Predicting Conductive Ear Conditions using Wideband Absorbance (WBA) and Wideband Tympanometry (WBT)

(Clients Robyn Choi and Chris Gonzalez; Team Ears)

AIM

1. Using test results obtained from *Wide Band Absorbance* (WBA) and *Wide Band Tympanometry* (WBT), apply machine learning techniques to determine testing frequencies that are most effective for diagnosing conductive hearing conditions in children.
2. Compare the predictive performance of machine learning models to typical ROC results obtained from WBA and WBT testing.

BACKGROUND (point out: why detecting hearing loss is important, why should we use a range of freq instead of a single freq => because more detail data=> more accuracy)

Analysing the absorbance of wideband stimuli has been available for decades. Wideband Absorbance is a measure of how effectively the middle ear can 'absorb' sound. This is usually done at ambient pressure ('WBA') or at tympanometry peak pressure ('WBT'). Recently, Wideband Absorbance has been measured over a range of frequencies instead of a single tone. This change produces effective results but creates much larger datasets, requiring heavier processing.

Receiver Operating Characteristics (ROC) analyses have been performed to achieve this, as done by Sliwa et al. [1] in their study, *Measurement of Wideband Absorbance as a Test for Otosclerosis*. The results showed effectiveness, but ROC analysis of potentially hundreds of data points may not be feasible in a busy audiology practice and may not be well understood.

In this project, we thus aim to further investigate and implement potential machine learning algorithms and statistical models to achieve a timely analysis of WBA/WBT results and a subsequent diagnosis.

--

VALUE PROPOSITION

Why use data science?

1. The data in this project (and in future) contains hundreds of frequencies from hundreds of patients. Creating an automated solution allows for more accurate and efficient predictions as the dataset provided is complex.

Understanding bottleneck/challenges in client's data interpretation

1. WBA machines produce much more complex and detailed data compared to conventional tympanometry.
2. Interpretation of data is impractical due to dependence on clinicians to interpret complex graphs produced by WBA machine.

Have a clear value proposition (efficiency, consistency/reproducibility, cost-saving etc.)

1. Successful results will enable audiologists and/or practitioners to read and understand the results of WBA tests more efficiently using a discrete set of frequencies instead of using hundreds of frequencies , ranging from 226Hz to 8000Hz.

Commented [DI(1]: Ear conditions? Or conductive conditions?

Commented [A(2R1]: According to the intro lecture its conductive ear

Commented [AJ(3]: @David not sure if this make sense, can you double check? I thought tpp is a measure as a result of the test?

Commented [DI(4R3]: Here's my understanding: WBA & WBT are both measures of wideband absorbance; WB*A* being at ambient pressure; hence A, and WBT being absorbance at tympanometry, hence the T

2. Tests will also be more accurate and consistent as results will provide a concise list of frequencies that are most important in detecting conductive hearing problems among children.
3. Due to the tests being conducted at less frequencies, it will cost less compared to WBA.

Commented [E(5]: imo this is hard to promise due to limited data

Commented [AJ(6R5]: @Edward we will find out! But yeah let's think about re-wording the deliverables and propositions so we don't over-promise

DELIVERABLES -D I. (Andy)

- Final report detailing the methodology, planned work and outcomes of the project. This report includes the following key sub-deliverables:
    1. Comparison of Machine Learning models to manually derived ROC
    2. Key visualization of findings
    3. Table of frequencies or frequency ranges that are most effective for diagnosis
    4. Code script used to process data and highlight frequency's that are most useful for conductive hearing diagnostics
    5. An interpretation of the model's degree of explainability
- Seminar presentation that is structured to communicate through the principles of data story telling (include reference to book). This guides the audience through problem definition, model building, solution delivery and actions from learnings.
- Provide a script file for the client that does the following:
    - Reproduces the results from the data file
    - Delivers a summary of the key outcomes with an accompanying visualization

METHODS

Part 1 - Planning

02 – Data exploration (Jess) => what is EDA and why you do it (rewrite)
The main task of this project is to identify most useful frequency regions in WBA and WBT for diagnosing conductive conditions in children's ears based on 239 observations, therefore we consider this to be a classification problem where the class label is the indicator ("OverallPoF") that shows whether the child passed various hearing tests.
After the Exploratory data analysis (EDA), we have the following main findings. Firstly, the "OverallPoF" indicator is imbalanced, with around 85% of observations passing hearing tests and 15% failing hearing tests. Secondly, descriptive statistics for each frequency band illustrate that most frequency bands are skewed in the absorbance, admittance, and phase data. Thirdly, a correlation matrix is used to display correlation coefficients for "OverallPoF" and different frequency bands. Absorbance data has the strongest correlation with "OverallPoF" at frequency band 1414.2136; YAdmittance data has the strongest correlation at frequency band 667.4199; Phase data has the strongest correlation at frequency band 1455.6532.

Commented [JX(7]: Can put in the appendix

Commented [AJ(8R7]: @Jessie hopefully in the email to EJ tomorrow she can give us an indication on whether this should be included or not 😊

Comments: Imbalanced binary classification: “logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes.”

Part 2 – Data Preparation

Data in industry rarely comes in an ideal format that is ready to be used for machine learning. The importance of preparing data comes as a result in theoretical statistics which states that a model’s accuracy and stability depends on the bias, variance, and irreducible error.

Inclusion of rogue data results in a misrepresentation of the true dataset, this will often promote an incorrect generalization of the model. Ultimately, inaccurate characterization of models leads to unexplainable results which reduces the interpretability of the models.

To prevent issues that stem from poor quality data, the cleaning process will seek to remove incomplete data, inconsistent data, and irrelevant features. Different feature selection techniques such as: variance thresholding, and applying appropriate models, can also be used to select relevant features.

Due to the variety of models applied in this research and their uniqueness in application, transformations of data will depend on the model applied. This could consist of refactoring labelled data, feature scaling or encoding data.

Part 3 – Model Building

To predict conductive ear conditions, appropriate classification models are chosen with consideration of the high-dimensional healthcare data, data with numerous variables. Unfortunately, simple Logistic Regression tends to overfit to high-dimensional data. A workaround is to introduce regularizations, resulting in lasso regression and ridge regression, or even a combination of both called the elastic net.

A different approach is to create a new low-dimensional set of features, derived from the original variables. This transformation can be done in 2 ways, unsupervised and supervised resulting in Principal Components Regression and Partial Least Squares Regression respectively.

Other suitable methods are tree-based models and Support Vector Machines. Random Forest, a tree-based classifier, can handle high-dimensional data as each tree is trained only on a random subset of variables. SVM works in high-dimensional spaces called hyperplanes and it performs worse in larger numbers of data, which perfectly matches our dataset.

Excluding PCR and PLSR, all the afore-mentioned models will value some frequency variables over the others, which helps in identifying useful frequency regions. This interpretation and the models’ accuracies can then be compared and presented through visualization.

Part 4 – Solution Delivery (Dennis)

The findings will be delivered via a report and presentation, which will cover all areas mentioned in the deliverables section. A repository of work (containing code and data) will also be provided to the client for further use or modifications.

Commented [DI(9]: Part -4 solution can probably just be one heading. Combine 06 and 07.

Commented [DI(10]: Already covered in deliverables ?

Commented [AJ(11R10]: @David maybe lets put deliverables under solution delivery? Thoughts?

Commented [DI(12R10]: I think we should keep deliverables section and put solution delivery under deliverables

Several data visualizations will be included along with the results, so that they could be better interpreted and understood. Since several models will be used for the modelling, a table will likely be included to highlight the accuracy results of said models. This would help the client understand why the team suggested a certain model over the others.

The team will also utilize line graphs to help picture the test results. For example, a line graph will be used to depict the WBA and WBT test results along the given frequency. This line graph will also help the team confirm the frequencies that might indicate conductive hearing, as there might be major differences in results on normal ear and an ear with conductive hearing on said frequencies.

Furthermore, a line graph would also be useful in depicting trends the team picked up during EDA or during the analysis itself. In addition, several other visualizations such as bar graphs or pie charts could also be used, for example if the team decides to depict data distribution.

PROJECT MANAGEMENT TIMELINE

- Gantt chart (in appendix). Cost. To clarify: 1 or 2 required.

APPENDIX

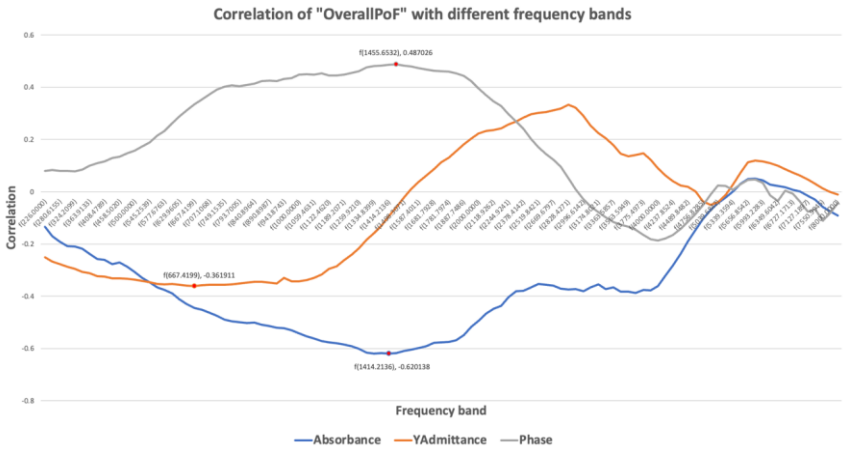
Individual plan and reflection on their contribution to the project for each member (max 400 words per member)

REFERENCES

Please refer to:

https://www.mybib.com/j/RobustBoilingGorilla

ARCHIVE



Exploratory Data Analysis

1. Performing data cleaning, making sure there are no empty columns or entries which do not match the format
2. Performing analysis on each variable, inspecting the distribution and/or trends of each variable.

Choice of different data exploration and analytical techniques

1. Performing factor analysis on the data using PCA and/or feature importance to rule out unnecessary variables.
2. Due to the imbalanced nature of the dataset, which consists of a total of 239 observations, with only 38 of them classified as failed. The team will work around this by utilizing different sampling techniques. This will be done carefully so they do not result in overfitting or losing valuable information which might help the model perform better. The sampling techniques used are:
    a. Sampling with/without replacement
    b. Oversampling/Under sampling
3. Creating and trialing multiple machine learning models to deduce the best model to predict conductive hearing in children. The following models will be used: Random forests, Logistic Regression, Decision Tree, K-Nearest Neighbours, Support Vector Machines.

How to make the process understandable and the output usable (e.g., parameter control, incorporating domain knowledge, communicating output uncertainty)

Visualizations to communicate the processes/results

1. Tables comparing the accuracy/performance of each model used.
2. Visualizations demonstrating the importance of different frequencies.
3. Graphing several entries of the WBA test results, comparing a failed result and a successful result to support the analysis result

Feature selection:

Feature selection is a process used to select relevant features as input for machine learning algorithms. Different feature selection techniques can be employed to reduce the number of features/variables by removing features that are irrelevant to the resulting model, while minimizing any negative impacts to performance. It has multiple benefits, such as:

- **Simpler models:** Resulting model is not complex and is easier to understand.
- **Shorter training times:** Less features used for training will reduce the time needed to train model.
- **Higher overall accuracy:** Model will only be trained on relevant features, which will increase its accuracy.

There are 3 main types of feature selection techniques, Wrapper, Filter, and Embedded feature selection. The dataset provided contains hundreds of features, hence wrapper methods are not recommended due to their large computation time.

Filter methods such as **Variance Thresholding** select features by ranking them using some sort of measure. Variance thresholding selects features that have variance above a certain threshold, as features with low variance contain less information.

Embedded methods perform feature selection in conjunction with the model construction phase. **Lasso and Ridge Regression** both work in a similar way, but ridge regression only weighs features based on importance, unlike lasso regression which also removes unimportant features.

Decision Trees and **Random Forests** automatically stores the importance of features after fitting the model inside a built-in variable. Relevant features can then be identified by ranking the scores within the variable, and a subset of data can be created using the relevant features. A new model is then constructed using the resulting subset.

--- JX Methodology attempt: