

Statistical Learning Theory Final Project Report

ML meets MD: Diagnosing Coughs and Wounds by Sound

Medical Symptoms Classification

Author(s): Sahar Ziv

Student ID(s): 312160369

Course Name: Statistical Learning Theory

Submission Date: June 2, 2024

Abstract

This section should provide a concise summary of the project's objectives, methods, key results, and conclusions.

1 Introduction

This project focuses on using machine learning techniques in order to classify medical symptoms based on self-reported audio recordings, with the goal of improving conversational agents in the medical field. The dataset comprises thousands of audio snippets, totaling over 8 hours of recording time, these recordings were created through a multi-step process where contributors first provided textual descriptions of symptoms, followed by recording corresponding audio. However, challenges such as incorrect labels and poor audio quality necessitate thorough data cleaning before model training. By addressing these issues, the project aims to create robust models that can accurately classify medical symptoms, thus enhancing the efficacy of conversational agents and improving healthcare diagnosis.

The research questions are:

- Can we classify the symptom of the patient to cough or infected wound only by the audio file?
- Which model classify better, audio file classification model or NLP model of the audio files transcriptions?

2 Data Description

The data found on Kaggle, the data was labeled using the Figure-Eight platform. The dataset contains 6661 audio recordings and a corresponding CSV file, where each recording has 3 indicators of audio quality measures (audio clipping, background noise audible and quiet speaker) with confidence score for those indications (between 0 and 1), an overall numeric score between 3 and 5 of the quality of the audio which is uncorrelated with the other audio quality measures. Furthermore, for each audio recording the CSV file contains the transcription of the audio, the writer ID, the speaker ID, and the label which is one of 25 medical symptoms. In order to develop the NLP model, I used the English language dictionary. In order to answer the research questions, a sub selection of the data is needed, where the label is either "Cough" or "Infected wound".

3 Exploratory Data Analysis (EDA)

In Figure 1, it's evident that the label frequency is quite balanced, indicating no need to account for unbalanced observations. The subset of data we're interested in comprises 599 observations.

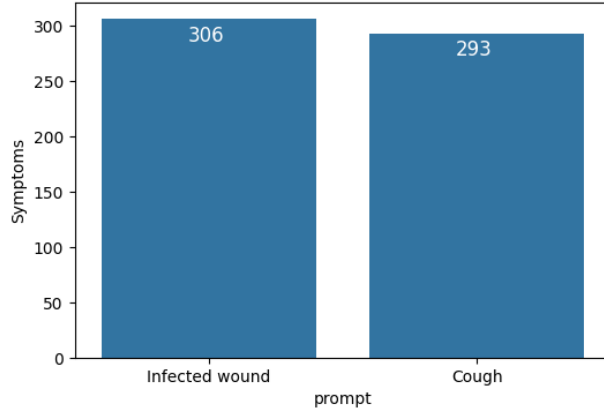


Figure 1: Label frequencies

In Figure 2, we can see the density of the overall quality score of the audio files by each label. We can see that there are more audio files with better overall quality for the "Infected wound" label than the "Cough" label.

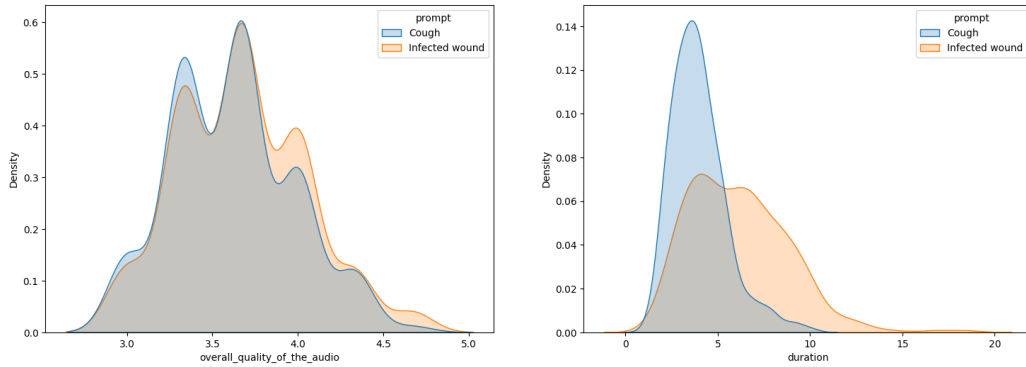


Figure 2: Left plot: Density of the overall quality score of the audio file for each label. Right plot: Density of the duration of the audio file for each label.

In Figure 3, we can see the absolute correlation between all the features I modified in the Model Engineering section. We can see really high correlation between some features (in bright red), which could result in bad predictions of our label. I decided to deal with it using LASSO and will be explained in the Model Engineering section.

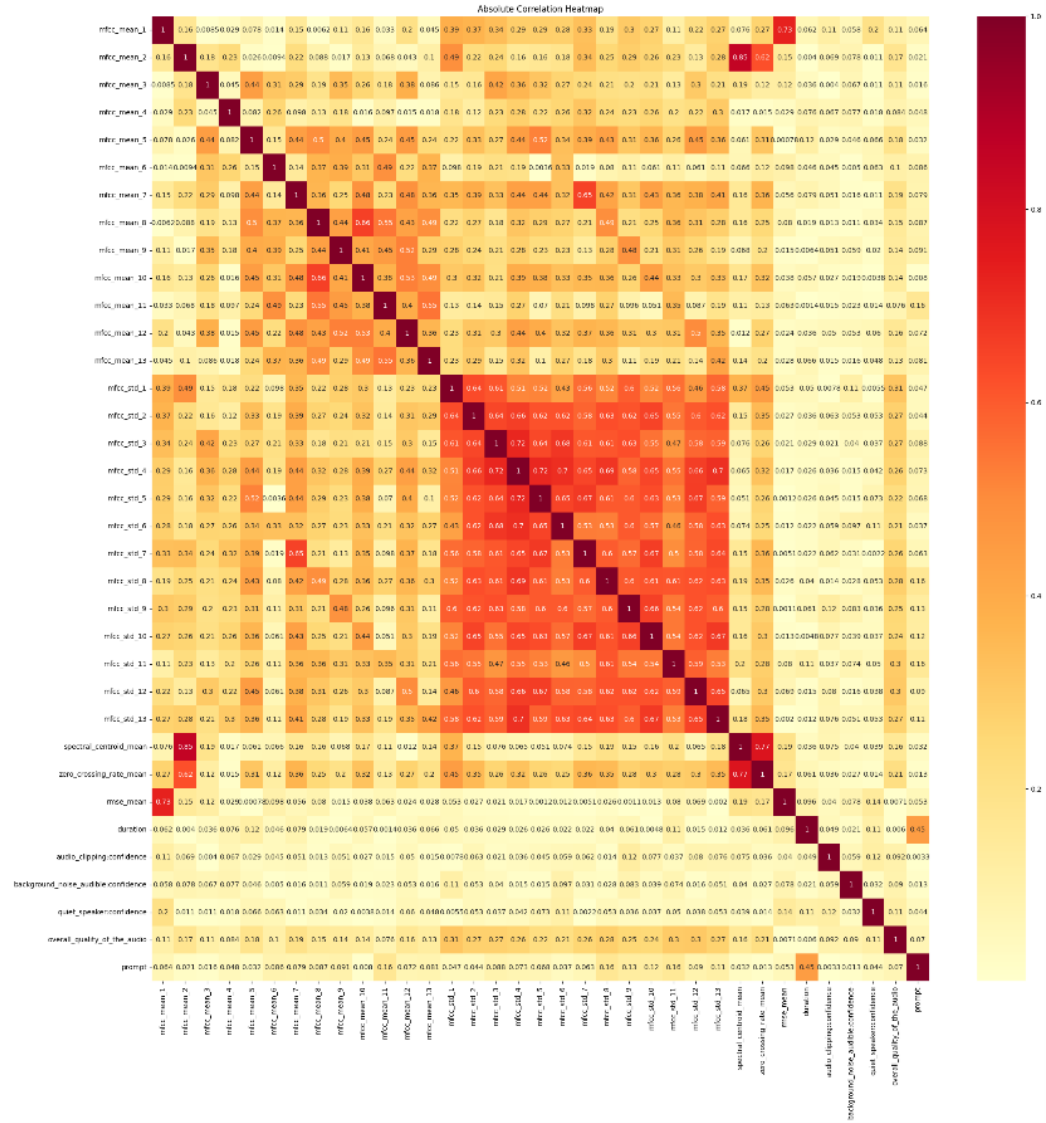


Figure 3: Label frequencies

4 Feature Engineering

In order to develop a model for classification from the audio files I converted each audio file to the following features:

- The duration of the audio file in seconds.
- The mean value of the MFCC coefficient values across all time series frames.
- The standard deviation value of the MFCC coefficient values across all time series frames.
- The mean and standard deviation spectral centroid across all time series frames, calculated by

$$\text{Centroid}(t) = \sum_k \frac{S[k, t] \cdot \text{freq}[k]}{\sum_j S[j, t]}$$

where, t is the time frame, S is a magnitude spectrogram and freq is the array of frequencies of the rows of S .

- The mean and standard deviation zero crossing rate across all time series frames, which means the mean rate at which the signal changes its sign (the number of times the audio waveform crosses the horizontal axis).
- The mean and standard deviation RMS (root mean square) across all time series frames.

To that I added the quality measures confidence score and the overall quality score, which resulted in 37 features.

In order to develop a model for classification from the transcriptions I used some conventional method in language process, I deconstructed words (I'm=I am, I've=I have...), lower cased and strip out punctuation. Then I removed the stop-words, lemmatized and removed the words 'cough', 'coughing', 'infected', 'infection', 'wound' and 'cut' and did one-hot encoding, to that again I added the quality measures confidence score and the overall quality score, which resulted in 174 features.

5 Modeling

For both classification methods I split the data to train and test 70%-30%, trained a scaler using only the train set and transforming both train and test set. I chose to fit 3 models Logistic Regression, KNN and SVM. For each model I have the following tuning parameters:

- Logistic Regression
 'penalty': LASSO or Ridge.
 'C': in $[1, e^3]$ – penalty inverse parameter, as it is smaller the stronger the regularization.
- KNN
 'K': number of nearest neighbors between $[3, 241]$ in odds numbers.
- SVM
 'kernel': linear, polynomial, or radial.
 'degree': between $[2, 6]$ for the polynomial kernel.
 'gamma': between $[0.00005, 1]$ for the radial kernel.

I will find the best tuning parameters by using 10-fold cross validation and measuring the accuracy of the validation sets for each combination of tuning parameters. The best combination model will be compared with the other models to determine the best model out of those three.

5.1 Mathematical Formulation

Denote $i = \{1, \dots, 599\}$ to be index for each medical self-reported phrase recorded. Denote a_{ij} to be each quality measure confidence score

where $j = \begin{cases} 1 & \text{Audio Clipping} \\ 2 & \text{Quiet Speaker} \\ 3 & \text{Background Noise} \end{cases}$, for each observation i . Denote b_i

to be a matrix of Mel-frequency cepstral coefficient size $\ell_i \times 13$, where $\ell_i = \left\lfloor \frac{\text{signal length}_i}{512} \right\rfloor + 1$ and $k = \{1, \dots, 13\}$ is the number of the coefficient, for each observation i . Denote c_{im} to be vectors of conversions of each audio file i ,

$$\text{where } m = \begin{cases} 1 & \text{Spectral Centroid} \\ 2 & \text{Zero Crossing Rate , each in length } \ell_i. \\ 3 & \text{RMSE} \end{cases}$$

The logistic regression equation is

$$\begin{aligned} \text{logit}[P(Y = \text{Cough}|X = x_i)] = & \beta_0 + \beta_j \cdot a_{ij} + \beta_4 \cdot \text{Overall quality score}_i \\ & + \beta_{4+k} \cdot \underset{\text{over } \ell_i}{\text{mean}}(b)_{ik} + \beta_{17+k} \cdot \underset{\text{over } \ell_i}{\text{std}}(b)_{ik} + \beta_{31} \cdot \text{duration}_i \\ & + \beta_{31+m} \cdot \underset{\text{over } \ell_i}{\text{mean}}(c)_{im} + \beta_{34+m} \cdot \underset{\text{over } \ell_i}{\text{std}}(c)_{im} \end{aligned} \quad (1)$$

Given a positive integer K and a test observation x_0 , the KNN classifier first identifies the K points in the training data that are closest to x_0 , represented by \mathcal{N}_0 . It then estimates the conditional probability for "Cough" as the fraction of points in \mathcal{N}_0 whose response values is "Cough":

$$P(Y = \text{Cough}|X = x_0) = \frac{1}{K} \cdot \sum_{i \in \mathcal{N}_0} \mathbb{I}(y_i \text{ is Cough}) \quad (2)$$

Finally, KNN classifies the test observation x_0 to the class with the largest probability.

The SVM equation is

$$\begin{aligned} & \text{Maximize } M \\ & \text{subject to } y_i \cdot (\beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \cdot K(x, x_i)) \geq M \cdot (1 - \varepsilon_i) \end{aligned} \quad (3)$$

where $K(x, x_i)$ is a kernel function:

- linear - $K(x_{ij}, x_{i'j}) = \sum_{j=1}^p x_{ij} \cdot x_{i'j}$
- polynomial - $K(x_{ij}, x_{i'j}) = (1 + \sum_{j=1}^p x_{ij} \cdot x_{i'j})^d$
- radial - $K(x_{ij}, x_{i'j}) = \exp(-\gamma \cdot \sum_{j=1}^p (x_{ij} - x_{i'j})^2)$

6 Results

This section should present the results in a clear and organized manner using tables, graphs, and charts. Include relevant evaluation metrics (accuracy, precision, recall, F1-score, etc.) to assess model performance comprehensively. Discuss any anomalies or unexpected results and provide possible explanations.

Header 1	Header 2	Header 3
Data 1	Data 2	Data 3
Data 4	Data 5	Data 6

Table 1: Caption for the table

Figure 5: Caption for the plot

7 Feature Importance and Interpretability

For models that allow for feature importance analysis, provide detailed interpretations of which features were most influential. Discuss any methods used to improve model interpretability.

8 Conclusions and Future Work

This section should summarize the key findings and their implications. Discuss the limitations of the current study and suggest areas for future research. Propose practical next steps based on the findings.

References

- [1] Author, “Title of the Paper,” *Journal Name*, vol. XX, no. XX, pp. XX-XX, Year.
- [2] Author, “Title of the Book,” Publisher, Year.

A Appendix A

This section should include supplementary material, additional figures/tables.