Laurene Thorson
Regis University
Denver, CO 80221
lthorson002@regis.edu

Heart Sound Classification through Machine Learning

ABSTRACT

Characterized datasets for "Heartbeat Sounds" are available on Kaggle. The sets consist of recorded audio files that were collected via an iPhone app and approximately 70% were labeled as "normal", "murmur", "extra heart sounds", and "artifact". The focus of this study was to focus on the "normal" and "murmur" file sets to create a machine learning classification model that would allow for a quick assessment of a patient's heart sounds, while in the doctor's office. Performing the code work in R, several functions were authored to loop through the directory containing the recordings, efficiently reading the ".wav" files, preprocessing the resulting signal, and ultimately extracting the acoustic signals and indices. Finally, training and test sets were built and evaluated using three machine learning techniques: K-nearest neighbors (KNN), Support Machine Vector (SVM), and Random Forest. Evaluation of the models resulted in the best classification results with SVM and caret optimization .

KEYWORDS

Heartbeat, Sounds, Murmur, Machine Learning, k-Nearest Neighbors, SVM, RandomForest, Classification

1. INTRODUCTION

When a normal heart is listened to with a stethoscope, heart sounds are heard in pairs and are a constant "lub-dub, lub-dub." The first "lub-dub" is the mitral and the tricuspid valves closing, while the second represents the aortic and pulmonary values closing immediately after. Additional sounds heard between these two sounds is typically referred to as a murmur. Murmurs can occur when blood is forced to flow through a narrowed valve (called stenosis), or when it leaks back through a defective valve (called regurgitation). These valve problems may be present at birth (congenital) or develop later in life because of rheumatic fever, coronary artery disease, infective endocarditis, or aging. (Texas Heart Institute, 2016). Murmurs can be "innocent" are require no treatment or lifestyle modifications, but they can also be congenital, such as a hole in the heart, or develop later in life due to rheumatic fever, coronary artery disease, infective endocarditis, or aging. Medicines are often prescribed by doctors to minimize the consequences of heart murmurs, so it is important that the condition be found and properly identified. The iPhone app technology allows for a quick assessment within the doctor office environment.

1.1. Dataset

The heartbeat dataset used in this evaluation was originally published on "peterjbentley.com" (Bentley, Nordehn, Coimbra, & Mannor, 2011) and was later hosted on the Kaggle website (King, 2017). Segregated into two datasets, the "A" set represents heart sound .wav audio files (31 normal, 34 murmur) that were collected from a general population using an iPhone microphone and associated app. Dataset B (202 normal, 66 murmur) was gathered during a clinical trial using a digital stethoscope and a similar app. The original datasets included other classifications, but this study only evaluated the "normal" and "murmur" categories. In both cases, the data was very noisy and required filtering due to issues like 60Hz electrical noise, stethoscope movement noise, and other external noise inputs.

1.2. Goal Expectation

The purpose of this effort is to leverage R to create code that read heartbeat .wav files from a directory, preprocessing and filtering them, and then leveraging machine learning packages and tools to train and test an efficient classifier. Performance assessments will be calculated using the "CrossTable" function from the R "gmodel" package (see sample output below). It is the opinion of this scientist, that the highest risk to the patient with a doctor-office-based heart beat assessment would be a mis-classifying a murmur heartbeat as a normal heartbeat. Therefore, the goal would be to "Minimize Murmur False Negatives" for the column highlighted below.

```
             | PredCaret
hbTestLabels |    normal |    murmur | Row Total |
-------------|-----------|-----------|-----------|
      normal |         4 |         0 |         4 |
-------------|-----------|-----------|-----------|
      murmur |         3 |        12 |        15 |
-------------|-----------|-----------|-----------|
Column Total |         7 |        12 |        19 |
-------------|-----------|-----------|-----------|
```

Figure 1: Sample Output of gmodel - CrossTable Function

Previous work on this challenge attempted to leverage Continuous Wavelet Transforms (CWT) to isolate and extract the S1 and S2 features. Then the region around the S1 peak was sampled, analyzed, and compared with the results shown in Figure 2.
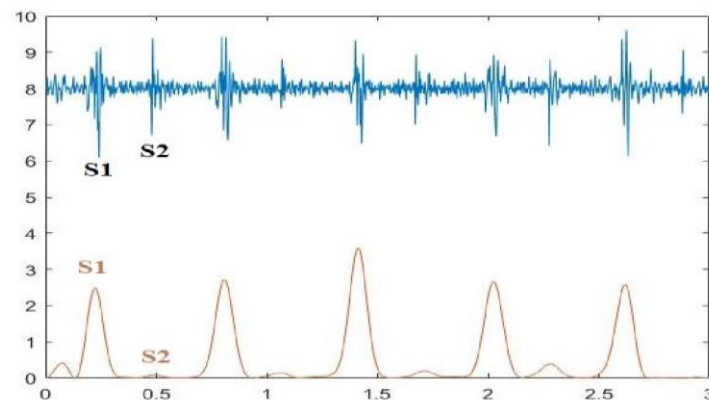

Figure 2: CWT Extraction of S1 and S2 Features

My analysis will, rather, focus on indices and envelopes related to the biological domain utilizing R packages that read in .wav files and provide simple indices ('acoustic_diversity' and 'bioacoustic_index') based on standard biodiversity indices (i.e. Shannon diversity index). These functions reflect roughly how 'diverse' the acoustic energy is distributed across the recorded frequency spectrum. While these functions do not identify patterns, it is the opinion of the author that they can be successfully leveraged to identify normal vs. murmur heart sound differences. (Oppel, 2017)

2. BACKGROUND
2.1. Human Heart Beat Characteristics

To properly process the data, several characteristics of human heart beats and heart sounds were required. Per the Mayo Clinic website, the normal heart rate for resting adults is 60 – 100 beats per minutes (approximately 1-2 Hz) and each beat contains a S1 (first "lub-dub") and S1 (second "lub-dub") heart sound. Any additional heart sounds between S1 and S2 typically represent a "heart murmur" as described in the table below. (Edward R. Laskowski, 2015).

Table 1: Heartbeat Sound Descriptions

Heartbeat Category	Description
Normal	Normal, healthy heart sounds" that have "a clear 'lub dub, lub dub' pattern"
Murmur	"Sounds as though there is a 'whooshing, roaring, rumbling or turbulent fluid' noise in one of two temporal locations: 1) between 'lub' and 'dub', or 2) between 'dub' and 'lub'

2.2. Domain Features and Measurement

Several different descriptive measurements have been evaluated to allow effective characterization and classification of the individual heart sound recordings. Many different websites related to the original dataset suggested determining the timing of the S1 and S2 signals and evaluating the signal between. Several of the assessment options from previous work are included in Table 2.

Table 2: Heart beat Assessment Variables

Domain	Techniques
Time	Energy Envelopes; Zero-Crossings; Min/Max Properties; Derivatives (Area Under the Curve)
Frequency	Fast Fourier Transforms (FFT); Power Densities
Time-Frequency	Short Time Fourier Transforms (STFT), Wavelet Transforms (WT)

2.3. Biological Domain Considerations

In addition to leveraging the techniques described in the previous section, there are several open source acoustic and/or biological sound packages available in R including "tuneR" to read in .wav files, and "Seewave", and "soundecology" to processes the information. These packages contain functions that generate indices for predefined sections of a signal or the signal as a whole. Originally, many of these R packages were created to categories bird songs, but can certainly be

modified to decipher heart sounds, as well. After review and analysis of the available indices the following were leveraged in this study:

    a. R "soundecology" package
        a) Acoustic Complexity Index (ACI) – Signal intensity variability
        b) Normalized Difference Soundscape (NDSI) – Segregates Biological sounds vs Anthropogenic (Man-made) noise
        c) Bioacoustic Index – Area under the curve for frequencies within a defined decibel (dB) range
        d) Acoustic Diversity Index (ADI) – Ratio of the signals beyond a defined threshold, falling to specified spectrum bins
        e) Acoustic Evenness Index (AEI) – Signal Evenness leveraging the "Gini Index" for bins in spectrogram
    2. R "Seewave" package
        a) Count of Zero Crossings in the waveform
        b) Acoustic Index based on the median of the amplitude envelope
        c) Total Entropy of a time wave
        d) Frequency Spectrum
        e) Shannon Spectral Entropy (noisy signal ~ 1; pure tone signal ~ 0)
        f) Spectrum Statistical Properties (meanfreq; sd; median; Q25; Q75; IQR; skew; kurtosis; sp.ent; sfm; mode; centroid)
        g) Frequency Amplitude peaks
        h) Fundamental Frequency parameters (meanfun; minfun; maxfun)
        i) Dominant Frequency parameters (meandom; mindom; maxdom; dfrange; duration)
        j) Modulation Index

3. ANALYSIS
3.1. File Processing and Data Analysis

The code for the .wav file processing is "Read and Process wav Directory" located at the GitHub url: https://github.com/lthorson002/MSDS692_DS-Practicum, which processes the heart beat recordings as follows:

    a. The .wav files in the defined working directory were translated using the "readwave" function from the "tuneR" package and truncated to 20 seconds in length.
    b. A bandpass filter between 75-1500 Hz was applied to the signal as most of the interesting bio-signals fall within this range. Then the data was normalized to allow for comparative assessments of signals gathered overtime to mitigate any intermittent noise factors. The figures below illustrate the preprocessing results.
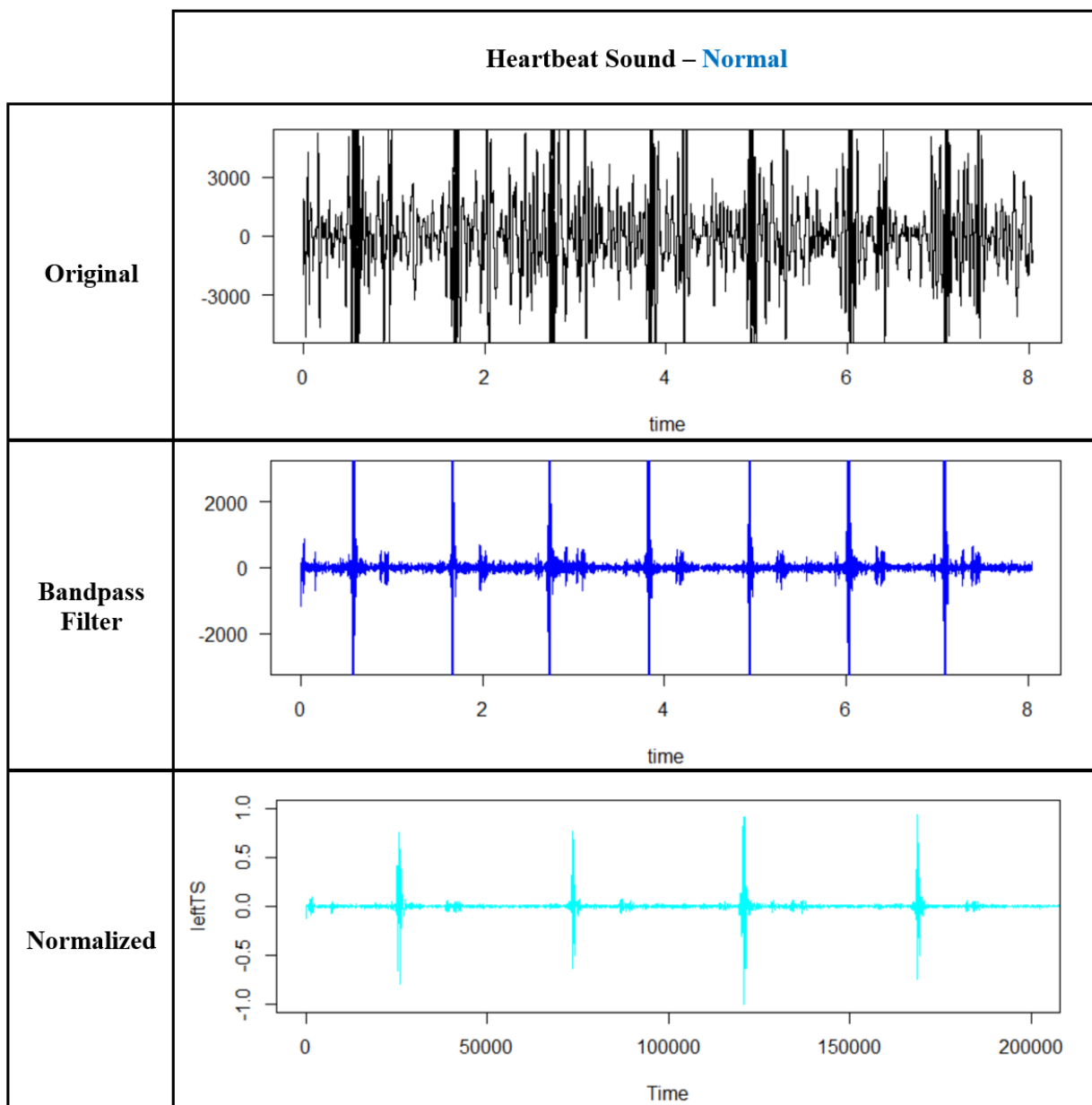
Figure 3: Preprocessing of Normal Heartbeat

	Heartbeat Sound – Murmur
Original	
Bandpass Filter	
Normalized	
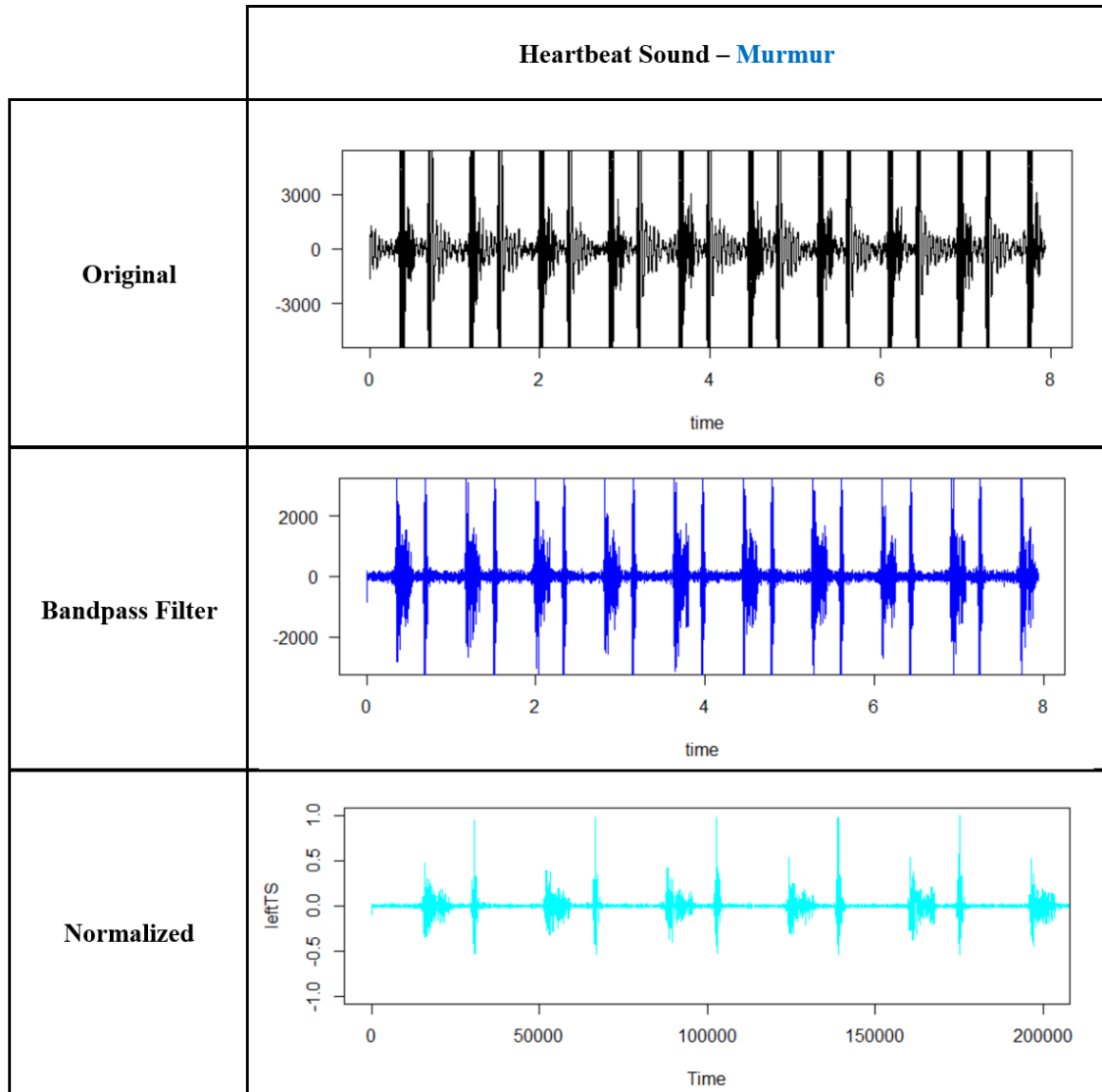
Figure 4: Preprocessing of Murmur Heartbeat

c.   Three user-defined R functions were created to loop through the directory containing the .wav files and ultimately create a dataframe with columns for the parameters described in the previous Section 2.3. These functions included: *SpectrumAnalyzehb:* reads, processes and filters .wav file; *beat:* applies the previous function to a specific directory; *processFolder:* creates and fills the dataframe as data is gathered during the looping.

d.   The "hbeat" dataframe which resulted from the "Read and Process wav Directory" R code was run against the normal/murmur labels using R's ggplot2 boxplot and facet_wrap function. Below are the resulting boxplots for each of the above parameters for the cleaner, less noisy Dataset A and the noisier Dataset B, segregated by heartbeat type. A review of the results highlights the ones with the least distribution overlap, which enhances machine learning opportunities: frequency

mean; standard deviation; median, skewness; kurtosis; spectral flatness centroid; acoustic complexity; zero crossings; acoustic diversity; acoustic evenness.
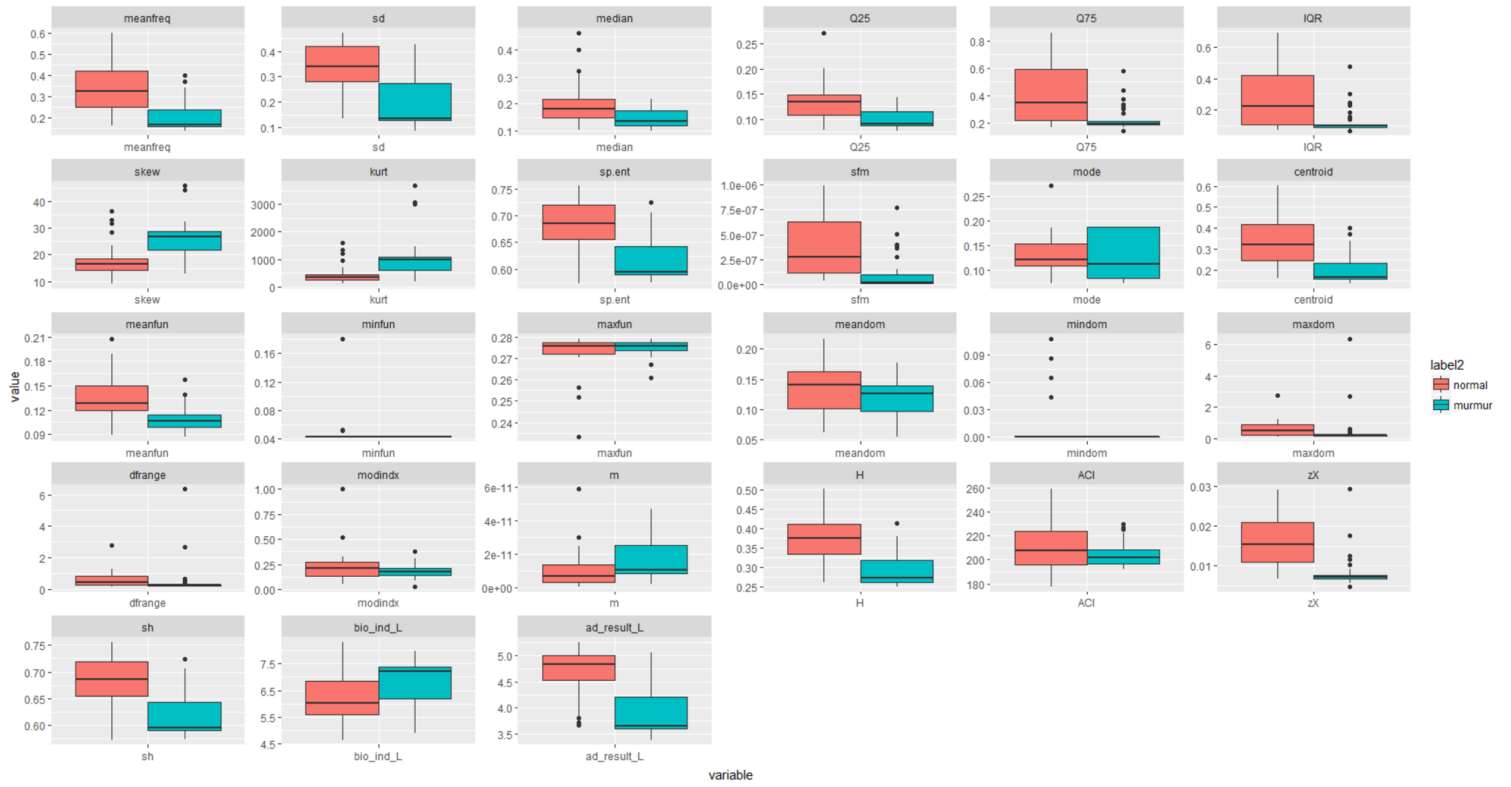
Figure 5: Boxplot Result for Data Set A (Less Noisy)
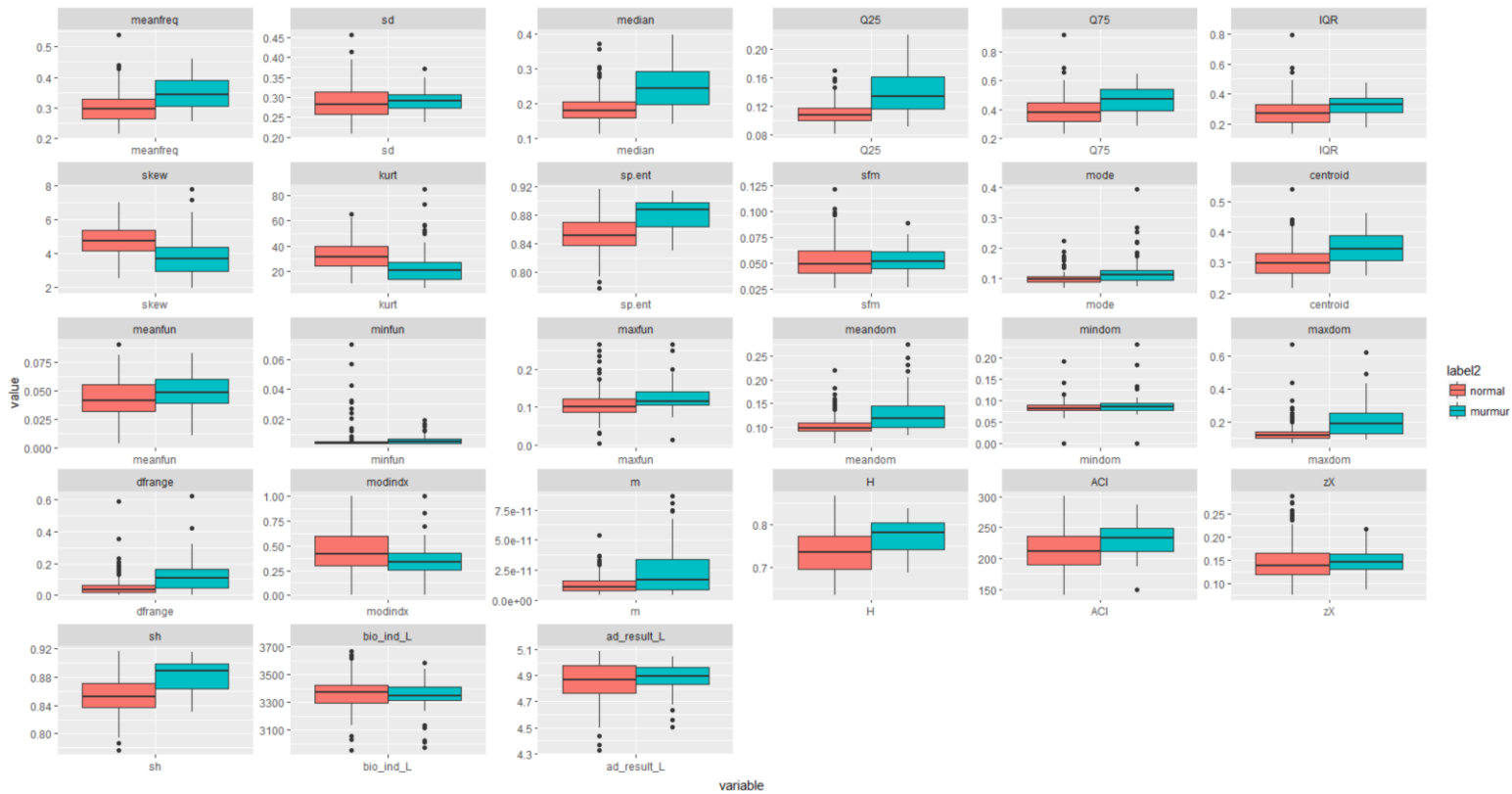
Figure 6: Boxplot Result for Data Set B (More Noisy)

3.2. Machine Learning Classification Techniques

The code for processing the R machine learning algorithms is the "Machine Learning" file located at the above GitHub url. The classification models leveraged during this evaluation included: *k-NN; Support Vector Machine; Random Forest.*

  a. k-NN is a simple algorithm that stores all known cases and classifies any new cases by how they match up with each of their k(count) of neighbors.
  b. In this evaluation, the "knn" function is from the R package "class" with k=3, based on manual evaluations. Training and test sets were created from the original "hbeat" data at 65% and 35%, respectively.
  c. Initial k-NN Classification results:
      i. Data Set A: k-NN results: *13.6% Chance of Murmur False Neg*

```
Total Observations in Table:  22

             | hbeat_pred
hbTestLabels |   normal |   murmur | Row Total |
-------------|----------|----------|-----------|
      normal |        5 |        2 |         7 |
-------------|----------|----------|-----------|
      murmur |        3 |       12 |        15 |
-------------|----------|----------|-----------|
Column Total |        8 |       14 |        22 |
-------------|----------|----------|-----------|
```

      ii. Data Set B: k-NN results: *15.7% Chance of Murmur False Neg*

```
Total Observations in Table:  89

             | hbeat_pred
hbTestLabels |   normal |   murmur | Row Total |
-------------|----------|----------|-----------|
      normal |       57 |       13 |        70 |
-------------|----------|----------|-----------|
      murmur |        9 |       10 |        19 |
-------------|----------|----------|-----------|
Column Total |       66 |       23 |        89 |
-------------|----------|----------|-----------|
```

d. R "caret" packaging, k-NN with training optimization (repeats=3, method="repeatedcv", tunelength=15):

    i.   Data Set A: k-NN results: *18.8% of Murmur False Negative*

```
Total Observations in Table:  22

             | PredCaret
hbTestLabels |   normal |   murmur | Row Total |
-------------|----------|----------|-----------|
      normal |        6 |        1 |         7 |
-------------|----------|----------|-----------|
      murmur |        4 |       11 |        15 |
-------------|----------|----------|-----------|
Column Total |       10 |       12 |        22 |
-------------|----------|----------|-----------|
```

    ii.   Data Set B: k-NN results: *15.7% of Murmur False Negative*

```
Total Observations in Table:  89

             | PredCaret
hbTestLabels |   normal |   murmur | Row Total |
-------------|----------|----------|-----------|
      normal |       59 |       11 |        70 |
-------------|----------|----------|-----------|
      murmur |       14 |        5 |        19 |
-------------|----------|----------|-----------|
Column Total |       73 |       16 |        89 |
-------------|----------|----------|-----------|
```

e. Initial SVM Classification results (guess: cost=1000 gamma=0.0001):

    iii. Data Set A: SVM results: *13.6% Chance of Murmur False Neg*

```
                | PredSVM
AllTest$label2  |   normal |   murmur | Row Total |
----------------|----------|----------|-----------|
         normal |        6 |        1 |         7 |
----------------|----------|----------|-----------|
         murmur |        3 |       12 |        15 |
----------------|----------|----------|-----------|
   Column Total |        9 |       13 |        22 |
----------------|----------|----------|-----------|
```

    iv. Data Set B: SVM results: *6.7% Chance of Murmur False Neg*

```
Total Observations in Table:  89

                | PredSVM
AllTest$label2  |   normal |   murmur | Row Total |
----------------|----------|----------|-----------|
         normal |       67 |        3 |        70 |
----------------|----------|----------|-----------|
         murmur |        6 |       13 |        19 |
----------------|----------|----------|-----------|
   Column Total |       73 |       16 |        89 |
----------------|----------|----------|-----------|
```

f. Round 2: After R "caret" package, SVM with training optimization (cost=10, gamma=0.1)

    iii.   Data Set A: SVM results: *9.0% of Murmur False Negative*

```
                | PredSVM2
AllTest$label2  |    normal  |   murmur  | Row Total |
----------------|------------|-----------|-----------|
       normal   |        5   |       2   |        7  |
----------------|------------|-----------|-----------|
       murmur   |        2   |      13   |       15  |
----------------|------------|-----------|-----------|
   Column Total |        7   |      15   |       22  |
----------------|------------|-----------|-----------|
```

iv. Data Set B: SVM results: *7.8% of Murmur False Negative*

```
Total Observations in Table:  89

                | PredSVM2
AllTest$label2  |    normal  |   murmur  | Row Total |
----------------|------------|-----------|-----------|
       normal   |       68   |       2   |       70  |
----------------|------------|-----------|-----------|
       murmur   |        7   |      12   |       19  |
----------------|------------|-----------|-----------|
   Column Total |       75   |      14   |       89  |
----------------|------------|-----------|-----------|
```

g. Round 3 After R "caret" package, SVM with training optimization and *some manual fine tuning*

v. Data Set A: SVM results: *9.1% of Murmur False Negative*
   a. (cost=100, gamma=0.001)

```
Total Observations in Table:  22

                | PredSVM3
AllTest$label2  |    normal  |   murmur  | Row Total |
----------------|------------|-----------|-----------|
       normal   |        6   |       1   |        7  |
----------------|------------|-----------|-----------|
       murmur   |        2   |      13   |       15  |
----------------|------------|-----------|-----------|
   Column Total |        8   |      14   |       22  |
----------------|------------|-----------|-----------|
```

vi. Data Set B: SVM results: *4.5% Chance of Murmur False Neg*
   a. (cost = .005, gamma = 1000)

```
Total Observations in Table:  89

                | PredSVM3
AllTest$label2  |    normal  |   murmur  | Row Total |
----------------|------------|-----------|-----------|
       normal   |       62   |       8   |       70  |
----------------|------------|-----------|-----------|
       murmur   |        4   |      15   |       19  |
----------------|------------|-----------|-----------|
   Column Total |       66   |      23   |       89  |
----------------|------------|-----------|-----------|
```

g. Initial randomForest Classification results;

vii. Data Set A: randomForest results: *9.1% Chance of Murmur False Neg*

```
                | RforPred
AllTest$label2  |    normal  |   murmur  | Row Total |
----------------|------------|-----------|-----------|
       normal   |        6   |       1   |        7  |
----------------|------------|-----------|-----------|
       murmur   |        2   |      13   |       15  |
----------------|------------|-----------|-----------|
   Column Total |        8   |      14   |       22  |
----------------|------------|-----------|-----------|
```

viii. Data Set B: randomForest results: *9.0% Chance of Murmur False Neg*

```
Total Observations in Table:  89

                | RforPred
AllTest$label2  |    normal |   murmur | Row Total |
----------------|-----------|----------|-----------|
        normal  |       68  |       2  |      70   |
----------------|-----------|----------|-----------|
        murmur  |        8  |      11  |      19   |
----------------|-----------|----------|-----------|
  Column Total  |       76  |      13  |      89   |
----------------|-----------|----------|-----------|
```

ix.   Data Set A: randomForest results: *13.6% Chance of Murmur False Neg*
         a.   Optimized with Repeat =2

```
Total Observations in Table:  22

              | PredCaret
AllTestY      |    normal |   murmur | Row Total |
--------------|-----------|----------|-----------|
      normal  |        6  |       1  |       7   |
--------------|-----------|----------|-----------|
      murmur  |        3  |      12  |      15   |
--------------|-----------|----------|-----------|
Column Total  |        9  |      13  |      22   |
--------------|-----------|----------|-----------|
```

x.   Data Set B: randomForest results: *7.9% Chance of Murmur False Neg*
         a.   Optimized with Repeat =3

```
Total Observations in Table:  89

              | PredCaret
AllTestY      |    normal |   murmur | Row Total |
--------------|-----------|----------|-----------|
      normal  |       68  |    .  2  |      70   |
--------------|-----------|----------|-----------|
      murmur  |        7  |      12  |      19   |
--------------|-----------|----------|-----------|
Column Total  |       75  |      14  |      89   |
--------------|-----------|----------|-----------|
```

Summary

References

Clean Data Set – Normal Heartbeat	
Noisy Data Set – Normal Heartbeat	