MISTY GROUP
Mohamed Abdelaziz k12137202
Lydia Mayer k11904969
Ivan Drinovac k12104744

Birds Audio Classification

Machine Learning & Pattern Classification [UE]

DATASET ANALYSIS

Table of Contents

Introduction

The provided Dataset consisted of:

Inputs:

- 1200 Separate Files with 548 columns each.
- 100 Rows / file each representing 0.2 seconds.
- ≈35 Feature sets of different lengths [1-60] / Row.

Labels:

- 1200 Separate Files describing 7 Classes.
- 100 Rows / file each representing the Annotations of an input row
- Different lengths of Annotation / Row in the form [5,5,0,1,0].

Preliminary Procedure:

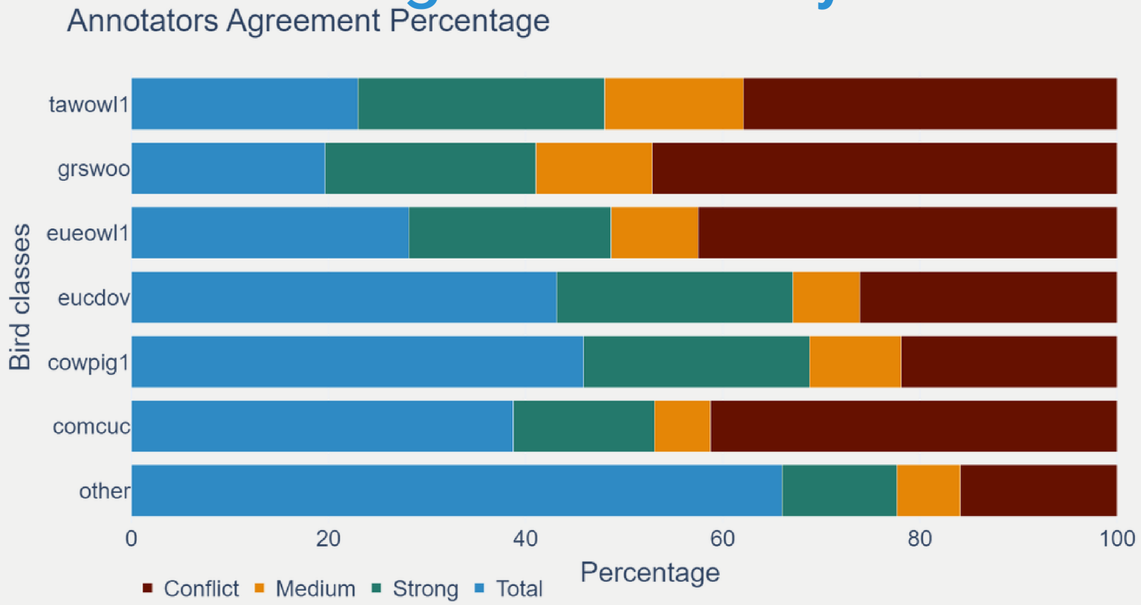
- Transforming both inputs and outputs "Separate files" into a **single data-frame** for easier analysis/training.
- **Labels:**

  - Calculating the Absolute Frequencies for each row with length 7 **[5,5,0] -> [1,0,0,0,2,0,0]**

  - Deriving the Relative Frequency for each Absolute Frequency, **[1,0,0,0,2,0,0] -> [0.33, 0, 0, 0, 0.67, 0, 0]** for a standard representation.
- **Features:**

  - Created a Dictionary **{Feature Set: [Feature Columns], ...}** for easier access in the future.

Annotators Agreement Analysis



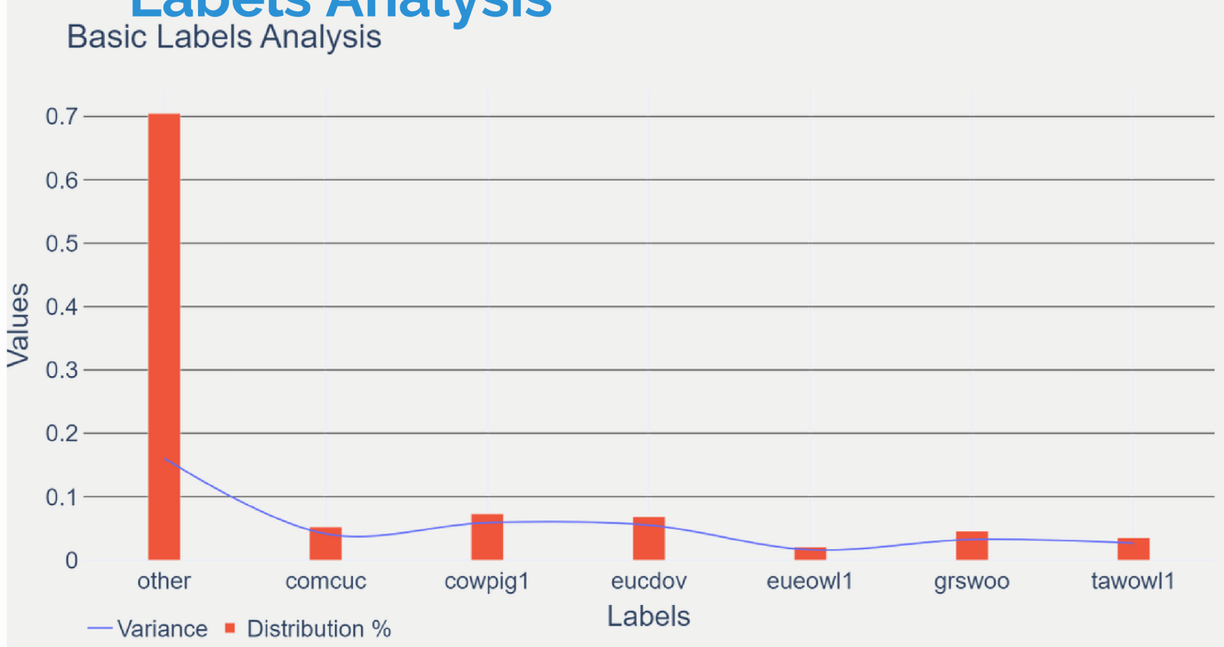
Procedure:

- From the "Standardized" labels data-frame, Selected all the rows which was labeled by any of the annotators. "values > 0"
- Created Bins of : [Total, Strong, Medium, Conflict], where **Total = 1** "All Annotators Agreed, **Strong [0.75<x<1]** "3 out of 4 agreed", **Medium [0.5<x<0.75]**, and **Conflict < 0.5** "Less than half of the annotators agreed"
- Summed over all the values for each bin and divided by the total number of rows *.

Conclusion:

- Category "other" had the highest agreement, while "grswoo" was the least agreed upon.
- **≈75%** of the data had at least **Medium Agreement** between the annotators.

Labels Analysis

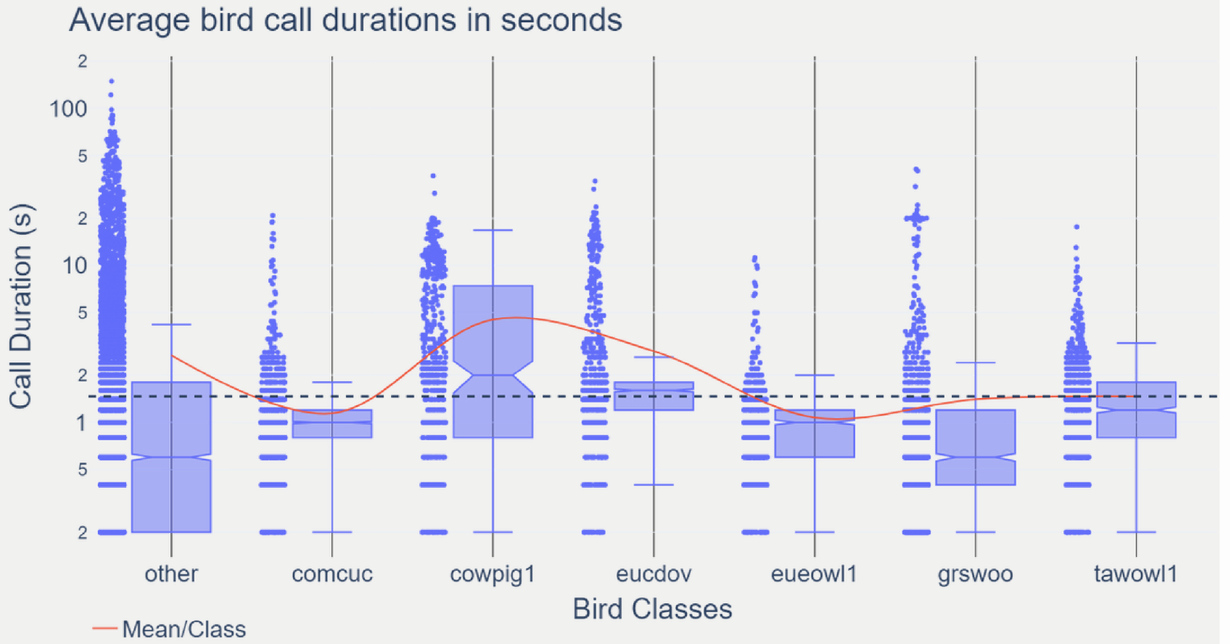


Procedure

- By **Summing over all the values** "including any Conflict in Annotations" the Dataset, and **dividing by** the number of the **Non-Zeros** for each label, to get a more precise insight.
- Calculating the Variance between Non-Zero labels / Class.

Conclusion

- The Dataset is highly Skewed towards the "other", and kind of balanced between the rest of the classes.
- There's Low Variance between the Classes "Since it was scaled from [0,1].



Procedure

- Selected All the Non-Zeros / label. "including Conflict in Annotations"
- Split each Selection into lists of consecutive rows. [rolling difference != 1]
- Calculated the lengths of each list to get how many rows were consecutive.
- Multiplied each value by 0.2 seconds.

Conclusion

- The "cowpig1" class had the longest avg call duration while "grswoo" had the lowest. [conflict correlation?]
- some of the files had more than 100 seconds of "other"/empty annotations -> Sparsity in the Data.
- The Median / Average rests around 1.5 seconds for all classes.

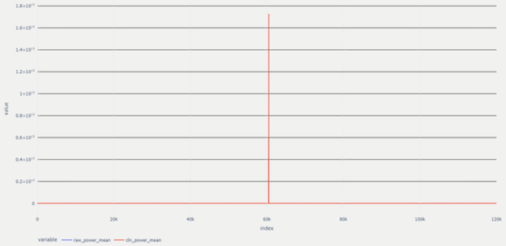
Features Analysis

Procedure

- Separated each **Feature-Set** [Legend on the Right] into its own matrix [120000xd] : d = # columns for feature.
- Calculated the Means / Variance for each feature set separately.

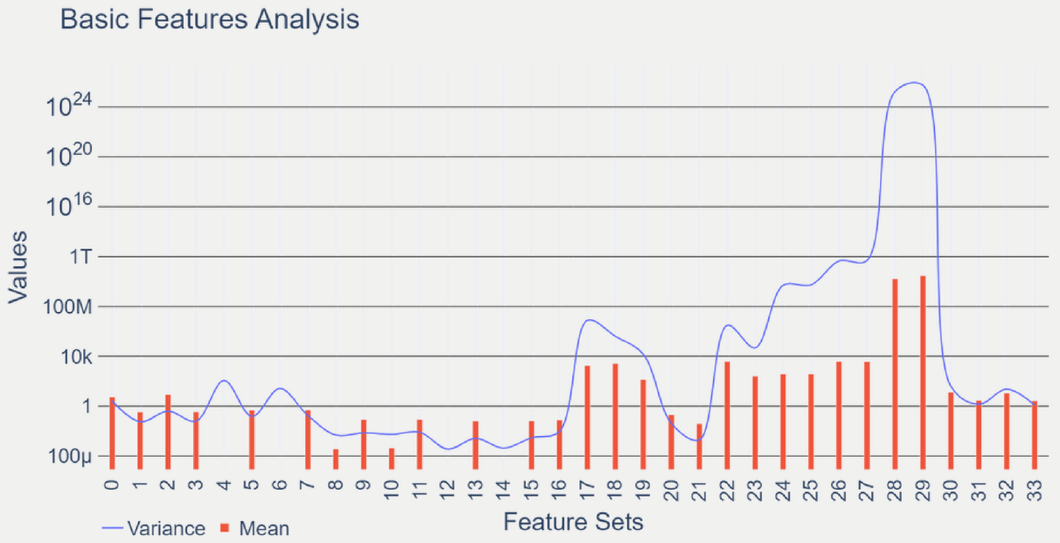
Conclusion

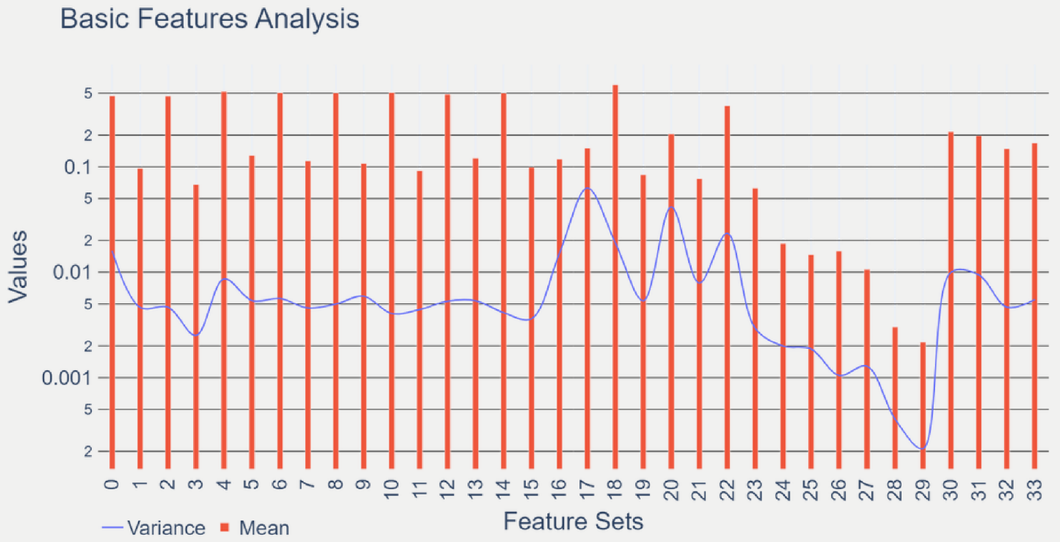
- Some of the features had very **High Variance** > 10000 [bandwidth_mean, badnwidth_stdm, centroid_mean, centroid std, power_mean, power_std], which suggests the existence of outliers within these features.
- Some of the features had very **Low Variance/Mean** ≈ 0 [raw_mfcc_mean, cln_mfcc_mean, raw_mfcc_d_mean, raw_mfcc_std, raw_mfcc_d2_mean, raw_mfcc_d2_std] which suggests existence of -ve/+ve values.



Solution

- Eliminate Outliers.
- Standardize Each Feature "regardless of the feature set" between [0,1] for better numerical stability.



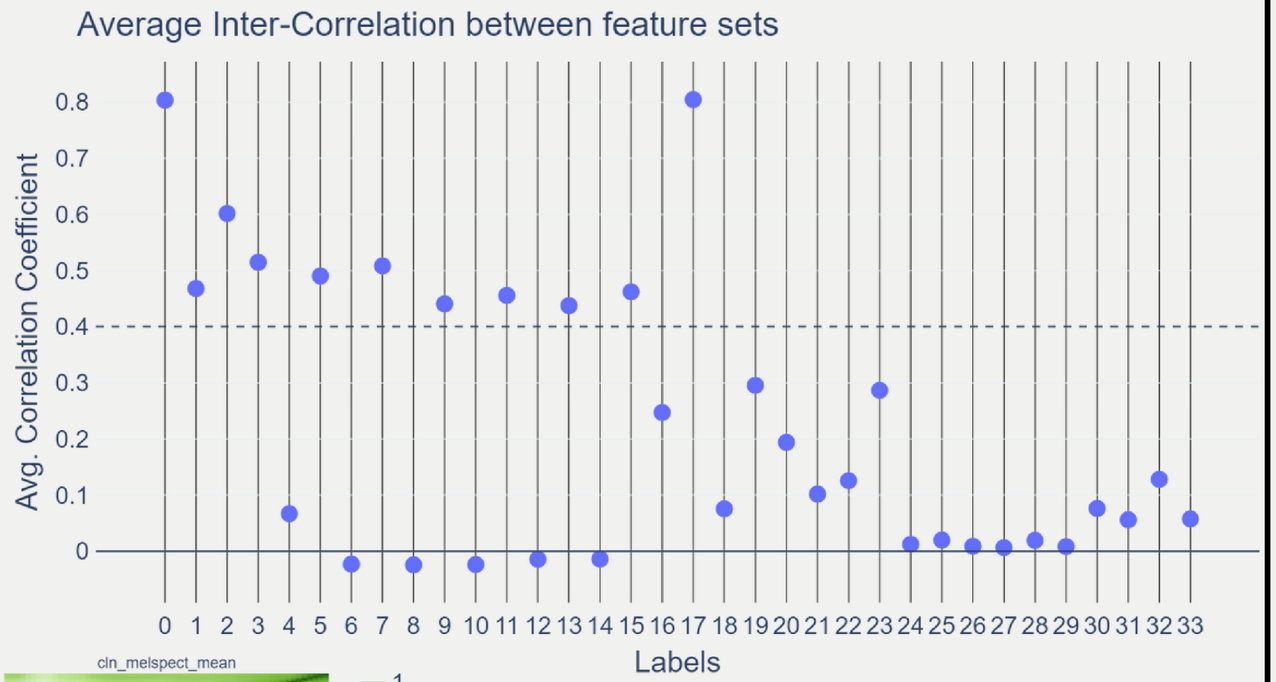


0	raw_melspect_mean
1	raw_melspect_std
2	cln_melspect_mean
3	cln_melspect_std
4	raw_mfcc_mean
5	raw_mfcc_std
6	cln_mfcc_mean
7	cln_mfcc_std
8	raw_mfcc_d_mean
9	raw_mfcc_d_std
10	cln_mfcc_d_mean
11	cln_mfcc_d_std
12	raw_mfcc_d2_mean
13	raw_mfcc_d2_std
14	cln_mfcc_d2_mean
15	cln_mfcc_d2_std
16	zcr
17	yin
18	bandwidth_mean
19	bandwidth_std
20	flatness_mean
21	flatness_std
22	centroid_mean
23	centroid_std
24	flux_mean
25	flux_std
26	energy_mean
27	energy_std
28	power_mean
29	power_std
30	raw_contrast_mean
31	raw_contrast_std
32	cln_contrast_mean
33	cln_contrast_std

Features Analysis [Inter-Correlation]

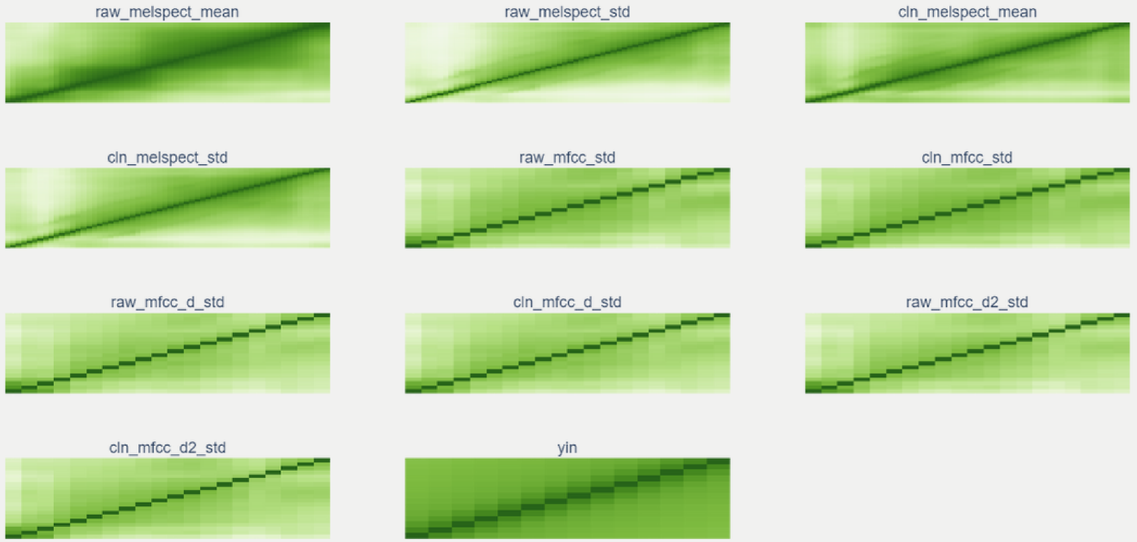
Procedure

- Calculated the Correlation Matrix For each **Feature Set,** [Legend on the right], separately to study the internal structure for each of them.
- Transformed the p-values to **Z-Fischer Correlation** "by applying arctanh() for the correlation matrix".
- Calculated the **Average** of all Coefficient of the Z-Fischer Matrix.



0	raw_melspect_mean
1	raw_melspect_std
2	cln_melspect_mean
3	cln_melspect_std
4	raw_mfcc_mean
5	raw_mfcc_std
6	cln_mfcc_mean
7	cln_mfcc_std
8	raw_mfcc_d_mean
9	raw_mfcc_d_std
10	cln_mfcc_d_mean
11	cln_mfcc_d_std
12	raw_mfcc_d2_mean
13	raw_mfcc_d2_std
14	cln_mfcc_d2_mean
15	cln_mfcc_d2_std
16	zcr
17	yin
18	bandwidth_mean
19	bandwidth_std
20	flatness_mean
21	flatness_std
22	centroid_mean
23	centroid_std
24	flux_mean
25	flux_std
26	energy_mean
27	energy_std
28	power_mean
29	power_std
30	raw_contrast_mean
31	raw_contrast_std
32	cln_contrast_mean
33	cln_contrast_std

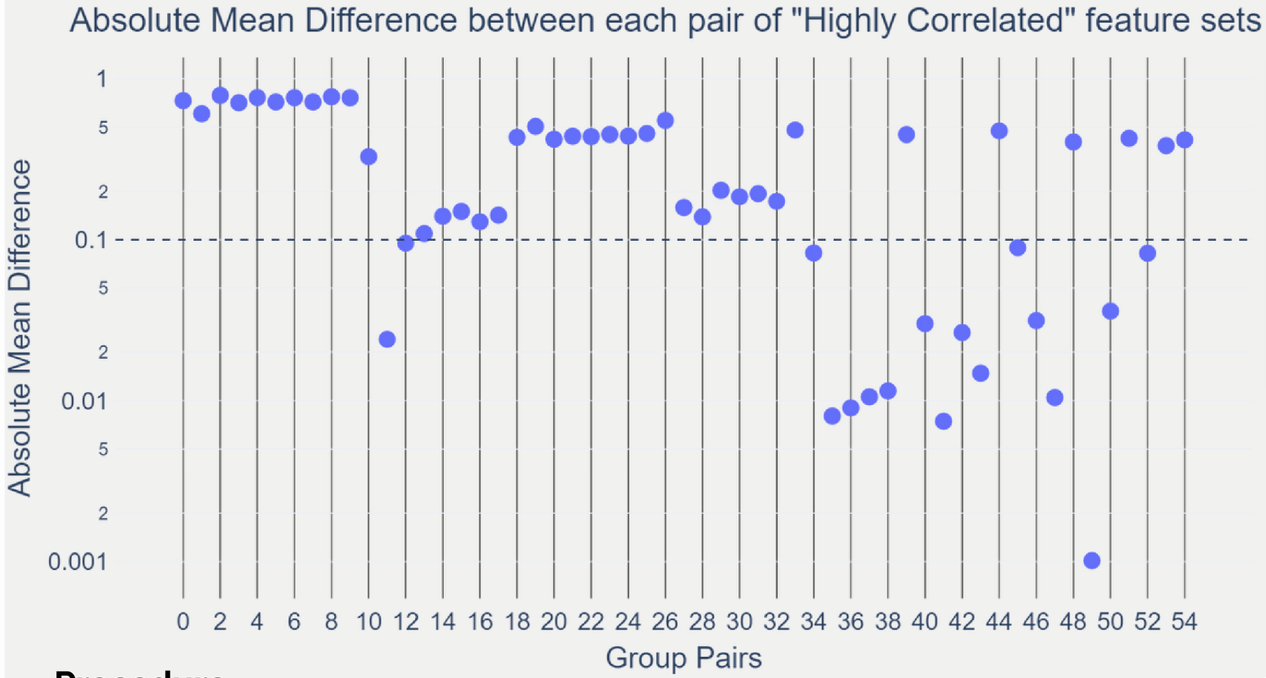
Top Inter Correlated Feature sets Matrices



Conclusion

- Some of the features had very **high average correlation** -> highly correlated **within themselves.**
- Confirmed by plotting some of these features [raw_melspect_mean/std, cln_melspect_mean/std raw_mfcc_std, cln_mfcc_std, yin, ..]

Features Analysis [Intra-Correlation]



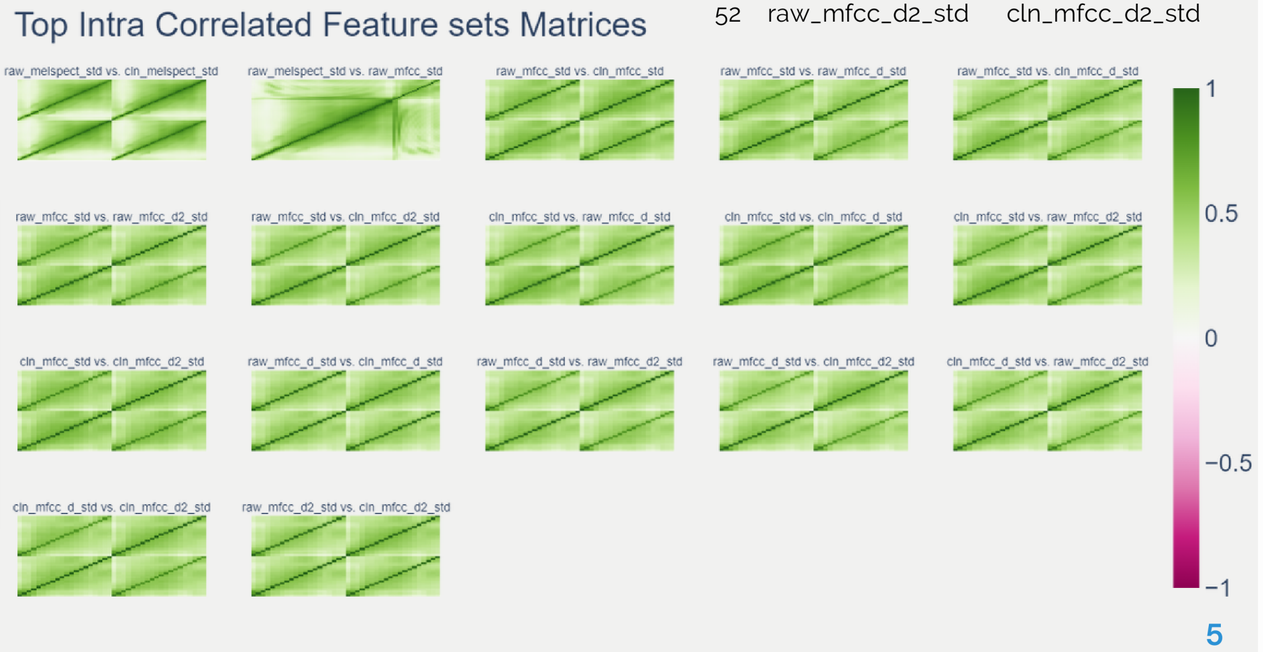
Procedure

- Calculated the Correlation Matrix for **each Pair of Highly Correlated Feature Set,** [Legend on the right], to study the relation ship across different features.
- Transformed the p-values to **Z-Fischer** Correlation by taking **arctanh()** for each correlation matrix.
- Calculated the Average of all Coefficient of the Z-Fischer Matrix.
- Calculated **Absolute Mean Difference** of the Average Correlation Coefficients **| |av1| - |av2| |** -> The lower the more correlated feature-sets pair.

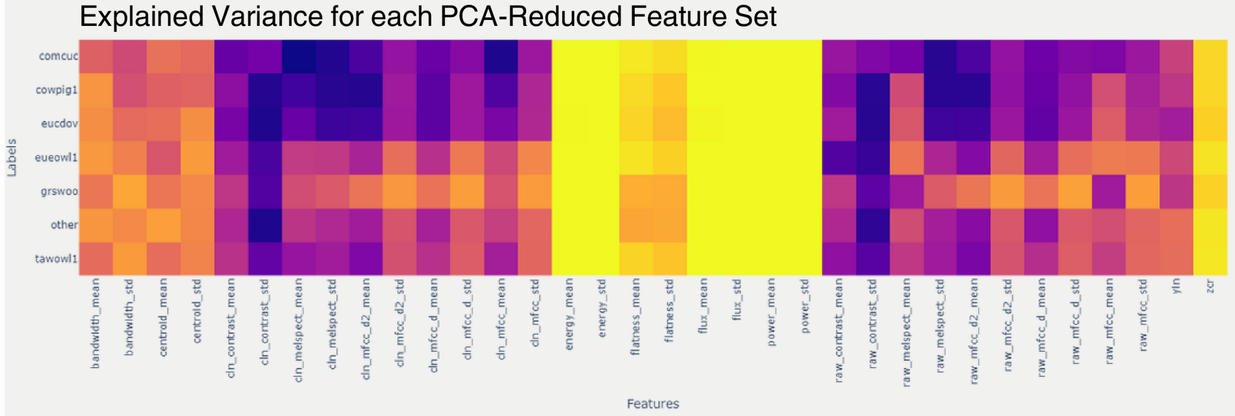
Conclusion

- Some of the feature sets have very low AMD -> Very High Correlation between them -> Redundant Features.
- Confirmed by plotting some of these feature pairs.
- Mainly the **MFCC** feature sets are **highly redundant.**
- There's some correlation between the **raw and cln** mel_spectrum to be confirmed by the explained_variance.

11	raw_melspect_std	cln_melspect_std
12	raw_melspect_std	raw_mfcc_std
34	raw_mfcc_std	cln_mfcc_std
35	raw_mfcc_std	raw_mfcc_d_std
36	raw_mfcc_std	cln_mfcc_d_std
37	raw_mfcc_std	raw_mfcc_d2_std
38	raw_mfcc_std	cln_mfcc_d2_std
40	cln_mfcc_std	raw_mfcc_d_std
41	cln_mfcc_std	cln_mfcc_d_std
42	cln_mfcc_std	raw_mfcc_d2_std
43	cln_mfcc_std	cln_mfcc_d2_std
45	raw_mfcc_d_std	cln_mfcc_d_std
46	raw_mfcc_d_std	raw_mfcc_d2_std
47	raw_mfcc_d_std	cln_mfcc_d2_std
49	cln_mfcc_d_std	raw_mfcc_d2_std
50	cln_mfcc_d_std	cln_mfcc_d2_std
52	raw_mfcc_d2_std	cln_mfcc_d2_std

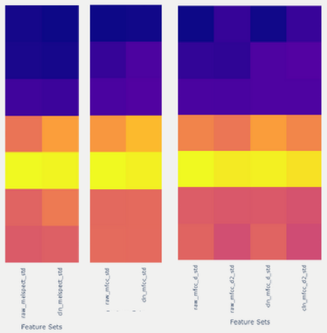


Labels/Features Analysis

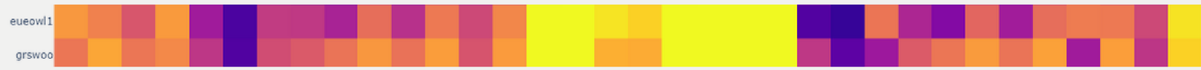


Conclusion

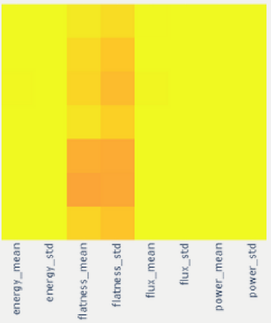
- Further Verified the Redundancy of some feature sets as their Explained Variance Matrix looks very similar



- The most explained classes by the dataset were [euewol1, grswo0]

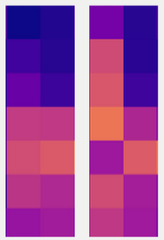


- The highest Explained Variance occurred in the low dimensional features. "even after eliminating the outliers and standardizing the inputs".



- The Most Similar classes in terms of explained variance ["Possibly" hard to distinguish between them], are [eueowl1, other]

- Refuted the previous hypothesis of the Redundancy between ["raw_melspect", "cln_melspect"]



Procedure

- We Used a **grid-search** for each label and each feature set independently.
- Using **PCA**, we **reduced the dimensions** of each feature-set into a single [dx1] : d = # of rows of the label.
- Calculated the **Explained Variance** of each label separately, in order to study how much the feature explains the label, "the higher explained variance the more the label is explained by this feature", assuming the features represent the labels.

Conclusion

Conclusion

- **Different** Annotators Agreement **Percentage** / Class with an **average of 75%**
- **Skewed Dataset** towards class "other" with **High % of total agreement**.
- Average Bird Call Duration is around **1.5 Seconds**.
- Different **High/Low Variance/Means** across the features
- Features with very **High-Inter Correlation**
- Same Features have **High Intra-Correlation** [MFCC, STD, ..]
- The most explained classes by the dataset were [euewol1, grswo0] while the least explained is comcuc1

Recommendation / Effect

- Can **Affect the final overall precision** of the trained model,
- Best dealt with **percentages for each row**.
- Can be partially remedied by **ignoring the "total" or "conflict"** when aggregating the label for each file.
- Can be used as a form of **Data Preprocessing/Augmentation** to **increase the sample** size of training/validation data
- **Rescale Features** between [0,1] and **Eliminate Outliers.**
- Utilize **Model Classes** that Make use of these Correlation **[CNN, LSTM, ..]**
- Eliminate Redundancy of features for overall better performance

Thank you..