ANC Project Report 1 - Linus

Motivation

To find clusters using dimension reduction techniques given the data collected and provided.

One of the most important aspects of dimensionality reduction is data visualization. Having to drop the dimensionality down to two or three, make it possible to visualize the data on a 2d or 3d plot, meaning important insights can be gained by analysing these patterns in terms of clusters and much more.

Data downloaded and overview

The downloaded files from One-Drive for this work:

- One Drive
    - FFT-A
    - FFT-Z
    - SPL-A
    - SPL-Z
    - 3rd-Octave-A
    - 3rd-Octave-Z
    - Spectrogram A
    - Spectrogram Z

All the above downloaded folders consist of 918 collected and annotated files. Total size of 35Gb.

The downloaded raw data are stored in the following path:

```
smb://10.97.31.159/SNTL_2/Linus/ANC-Project/ANC-Data/
```

Data preprocessing

The preprocessing of the provided data can be broken into 5 steps:

1. Reading
2. Converting
3. Standardizing
4. Labelling
5. Exporting

Reading

To read the provided .asc files and converting to numpy arrays, .npy files for data processing.

```
# convert asc files to npy files
path = '../ANC-Data/Spectrogram-A/Spectrogram (A)/'
```

```
npy_dest_path = '../ANC-Data-NPY/Spectrogram-A/'
files = glob.glob(path + '*')

for i in files:
    with codecs.open(i, encoding='utf-8-sig') as f:
        X = np.loadtxt(f, usecols=(0,1))
        np.save(npy_dest_path + i.split('/')[-1][:-3] + 'npy', X)
```

The converted and exported npy files are stored in the following path:

```
smb://10.97.31.159/SNTL_2/Linus/ANC-Project/ANC-Data-NPY/
```

Standardization

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

Standardize features by removing the mean and scaling to unit variance

The standard score of a sample `x` is calculated as:

```
z = (x - u) / s
```

where `u` is the mean of the training samples, and `s` is the standard deviation of the training samples.

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set.

Labelling

A new column, titled 'Class', is inserted into the dataframe. As the filenames provided were annotated with the sound class, the labelling is automated with a script that runs through the filenames and assigns the corresponding sound class as a datapoint under the 'Class' column.

Figure 1 shows the generated new 'Class' column, alongside the corresponding analysis entries for each sound class.

Although the 3 main classes are 'Plane', 'Helicopter' and 'Rumble', the provided annotated filenames included 2 more sub-classes ('Plane_Raining' and 'Plane_highAltitude') for 'Plane'.

The finer subclasses are labelled in the dataframe as such, therefore, making the total number of classes to be 5. This will provide us better insights for the variants of class 'Plane' when we visualize the data in a 2D space.

Exporting

```
df.to_csv('./testing_standardized_flattened.csv')
# convert to pickle to retain assigned data types the dataframe since csv
would all be converted to strings
df.to_pickle('./testing_standardized_flattened.pickle')
```

The exported pickle and csv files are stored in the following path:

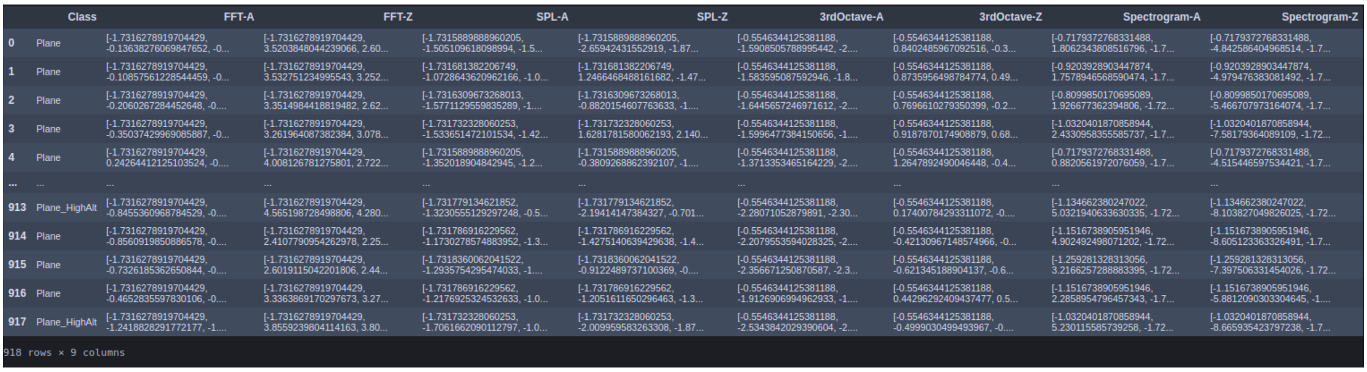
```
smb://10.97.31.159/SNTL_2/Linus/ANC_Project/Scripts/
```

Overview of dataframe

Dataframe:

- 918 entries (rows)
  - 700 Planes
  - Planes (Raining)
  - Planes (High altitude)
  - Helicopters
  - Rumble

- 9 features (columns)
  - Class
    - Planes
    - Planes_Raining
    - Planes_HighAlt
    - Helicopter
    - Rumble
  - FFT-A
  - FFT-Z
  - SPL-A
  - SPL-Z
  - 3rd-Octave-A
  - 3rd-Octave-Z
  - Spectrogram-A
  - Spectrogram-Z



	Class	FFT-A	FFT-Z	SPL-A	SPL-Z	3rdOctave-A	3rdOctave-Z	Spectrogram-A	Spectrogram-Z
0	Plane	[-1.7316278919704429, -0.13638276069847652, -0...	[-1.7316278919704429, 3.5203848044239066, 2.60...	[-1.7315889888960205, -1.505109618098994, -1.5...	[-1.7315889888960205, -2.65942431552919, -1.87...	[-0.5546344125381188, -1.5908505788995442, -2....	[-0.5546344125381188, 0.8402485967092516, -0.3...	[-0.7179372768331488, 1.8062343808516796, -1.7...	[-0.7179372768331488, -4.842586404968514, -1.7...
1	Plane	[-1.7316278919704429, -0.10857561228544459, -0...	[-1.7316278919704429, 3.532751234995543, 3.252...	[-1.731681382206749, -1.0728643620962166, -1.0...	[-1.731681382206749, 1.2466468488161682, -1.47...	[-0.5546344125381188, -1.583595087592946, -1.8...	[-0.5546344125381188, 0.8735956498784774, 0.49...	[-0.9203928903447874, 1.7578946568590474, -1.7...	[-0.9203928903447874, -4.979476383081492, -1.7...
2	Plane	[-1.7316278919704429, -0.2060267284452648, -0....	[-1.7316278919704429, 3.3514984418819482, 2.62...	[-1.7316309673268013, -1.5771129559835289, -1....	[-1.7316309673268013, -0.8820154607763633, -1....	[-0.5546344125381188, -1.6445657246971612, -2....	[-0.5546344125381188, 0.7696610279350399, -0.2...	[-0.8099850170695089, 1.926677362394806, -1.72...	[-0.8099850170695089, -5.466707973164074, -1.7...
3	Plane	[-1.7316278919704429, -0.35037429969085887, -0...	[-1.7316278919704429, 3.261964087382384, 3.078...	[-1.731732328060253, -1.533651472101534, -1.42...	[-1.731732328060253, 1.6281781580062193, 2.140...	[-0.5546344125381188, -1.5996477384150656, -1....	[-0.5546344125381188, 0.9187870174908879, 0.68...	[-1.0320401870858944, 2.4330958355585737, -1.7...	[-1.0320401870858944, -7.58179364089109, -1.72...
4	Plane	[-1.7316278919704429, 0.24264412125103524, -0....	[-1.7316278919704429, 4.008126781275801, 2.722...	[-1.7315889888960205, -1.352018904842945, -1.2...	[-1.7315889888960205, -0.3809268862392107, -1....	[-0.5546344125381188, -1.3713353465164229, -2....	[-0.5546344125381188, 1.2647892490046448, -0.4...	[-0.7179372768331488, 0.8820561972076059, -1.7...	[-0.7179372768331488, -4.515446597534421, -1.7...
...	...	...	...	...	...	...	...	...	...
913	Plane_HighAlt	[-1.7316278919704429, -0.8455360968784529, -0....	[-1.7316278919704429, 4.565198728498806, 4.280...	[-1.731779134621852, -1.3230555129297248, -0.5...	[-1.731779134621852, -2.19414147384327, -0.701...	[-0.5546344125381188, -2.28071052879891, -2.30...	[-0.5546344125381188, 0.17400784293311072, -0....	[-1.134662380247022, 5.0321940633630335, -1.72...	[-1.134662380247022, -8.103827049826025, -1.72...
914	Plane	[-1.7316278919704429, -0.8560919850886578, -0....	[-1.7316278919704429, 2.4107790954262978, 2.25...	[-1.731786916229562, -1.1730278574883952, -1.3...	[-1.731786916229562, -1.4275140639429638, -1.4...	[-0.5546344125381188, -2.2079553594028325, -2....	[-0.5546344125381188, -0.42130967148574966, -0...	[-1.1516738905951946, 4.902492498071202, -1.72...	[-1.1516738905951946, -8.605123363326491, -1.7...
915	Plane	[-1.7316278919704429, -0.7326185362650844, -0....	[-1.7316278919704429, 2.6019115042201806, 2.44...	[-1.7318360062041522, -1.2935754295474033, -1....	[-1.7318360062041522, -0.9122489737100369, -0....	[-0.5546344125381188, -2.356671250870587, -2.3...	[-0.5546344125381188, -0.621345188904137, -0.6...	[-1.259281328313056, 3.2166257288883395, -1.72...	[-1.259281328313056, -7.397506331454026, -1.72...
916	Plane	[-1.7316278919704429, -0.4652835597830106, -0....	[-1.7316278919704429, 3.3363869170297673, 3.27...	[-1.731786916229562, -1.2176925324532633, -1.0...	[-1.731786916229562, -1.2051611650296463, -1.3...	[-0.5546344125381188, -1.9126906994962933, -1....	[-0.5546344125381188, 0.44296292409437477, 0.5...	[-1.1516738905951946, 2.2858954796457343, -1.7...	[-1.1516738905951946, -5.8812090303304645, -1....
917	Plane_HighAlt	[-1.7316278919704429, -1.2418828291772177, -1....	[-1.7316278919704429, 3.8559239804114163, 3.80...	[-1.731732328060253, -1.7061662090112797, -1.0...	[-1.731732328060253, -2.009959583263308, -1.87...	[-0.5546344125381188, -2.5343842029390604, -2....	[-0.5546344125381188, -0.4999030499493967, -0....	[-1.0320401870858944, 5.230115585739258, -1.72...	[-1.0320401870858944, -8.665935423797238, -1.7...

```
918 rows × 9 columns
```

Figure 1: A snippet of the processed and exported dataframe. Each row is a data entry for the which consists

the sound class and its corresponding analysis entries.

Dimensionality Reduction

Several techniques have been experimented. The techniques include dimension reduction techniques like PCA and TSNE,

Principal Component Analysis (PCA)

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed [1].

t-distributed stochastic neighbour embedding (TSNE)

TSNE creates a probability distribution using the Gaussian distribution that defines the relationships between the points in high-dimensional space. TSNE uses the t-distribution to recreate the probability distribution in low-dimensional space [2].

PCA with TSNE

The idea is using both PCA and TSNE to visualize the high dimensional principal components in a 2D space.

The approach can be summarised in 2 steps. The first, applying PCA to reduce the high dimension data of say 12285 bins of standardized FFT values into a lower dimension. The seconnd, applying TSNE on the reduced principal components to visualize the possible clusters on a 2 dimensional space.

PCA here takes in the high dimensional features of the FFT-A with 918 entries of shape 12285 bins values and calculates the principal components. The number of components used in this work is set to 50. The extracted 50 principal components of the features were further reduced with TSNE into 2 dimensions for visualization purpose.

Since FFT and 3rd-Octave are in fixed bins, 12285 and 111 respectively, we do not need to segment the data points to provide a fixed shape when feeding into both reduction algorithms.

As for the Spectrogram and SPL features, since the raw audio files used to compute these features in unequal length in time initially, we would have to segment the data points in order to provide acceptable inputs for the reduction algorithms.

The following subsections shows the reduced dimensions in a 2D space for visualization. The features are plotted in both static and dynamic figures. The static plot applied 1000 iterations on the TSNE algorithm, whereas the dynamic plot shows a successions of 250-interval iterations in the range of 250 to 10000 iterations.

Single feature

FFT-A

12285 FFT-A bins > PCA > 50 principal components > TSNE > 2D Visualization

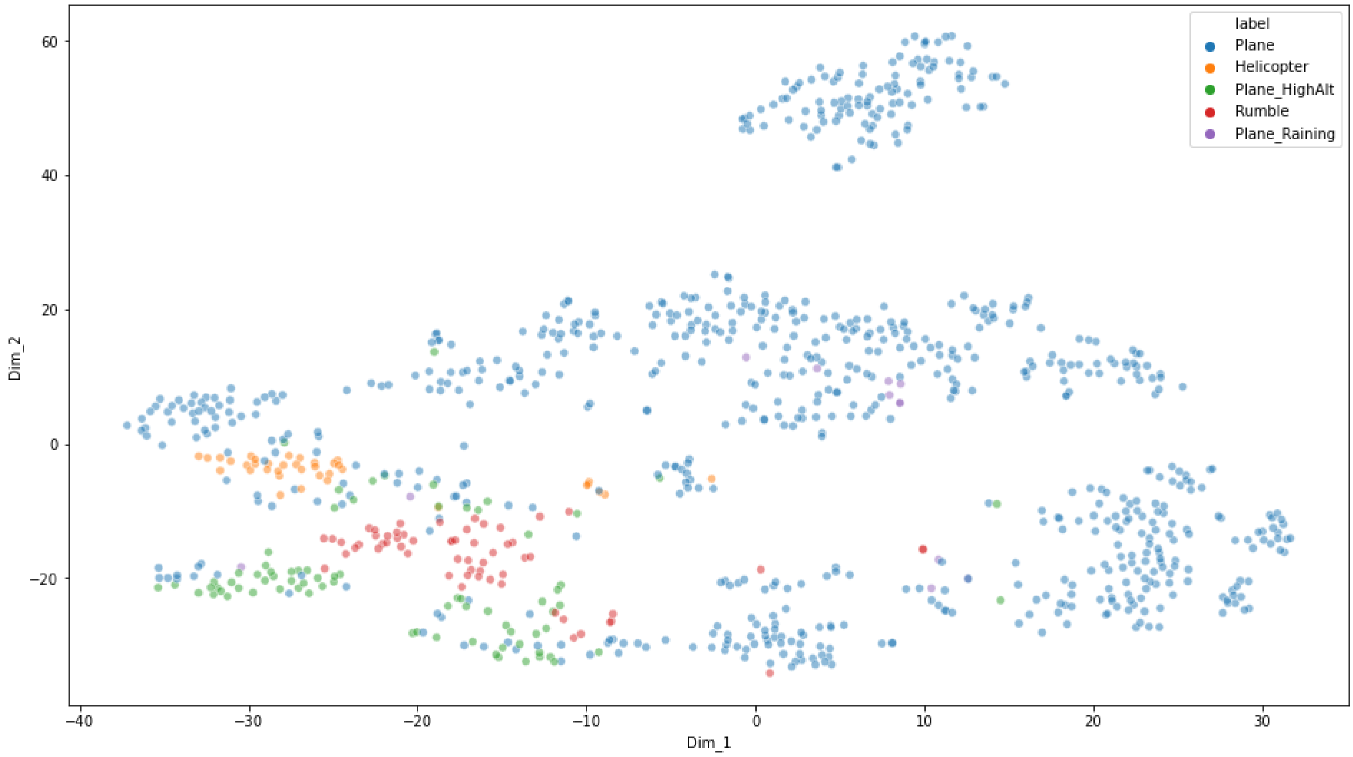


Figure 2: TSNE plot on the FFT-A feature, number of iterations = 1000

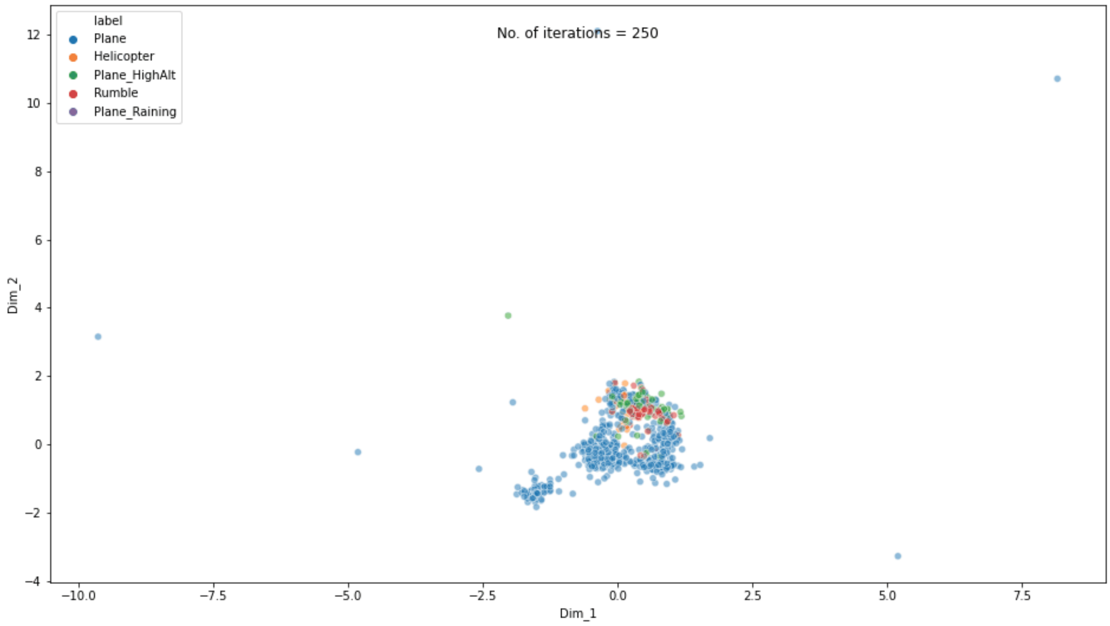


Figure 3: Dynamic plot of the FFT-A feature, iterations range from 250 to 10000

FFT-Z

12285 FFT-Z bins > PCA > 50 principal components > TSNE > 2D Visualization

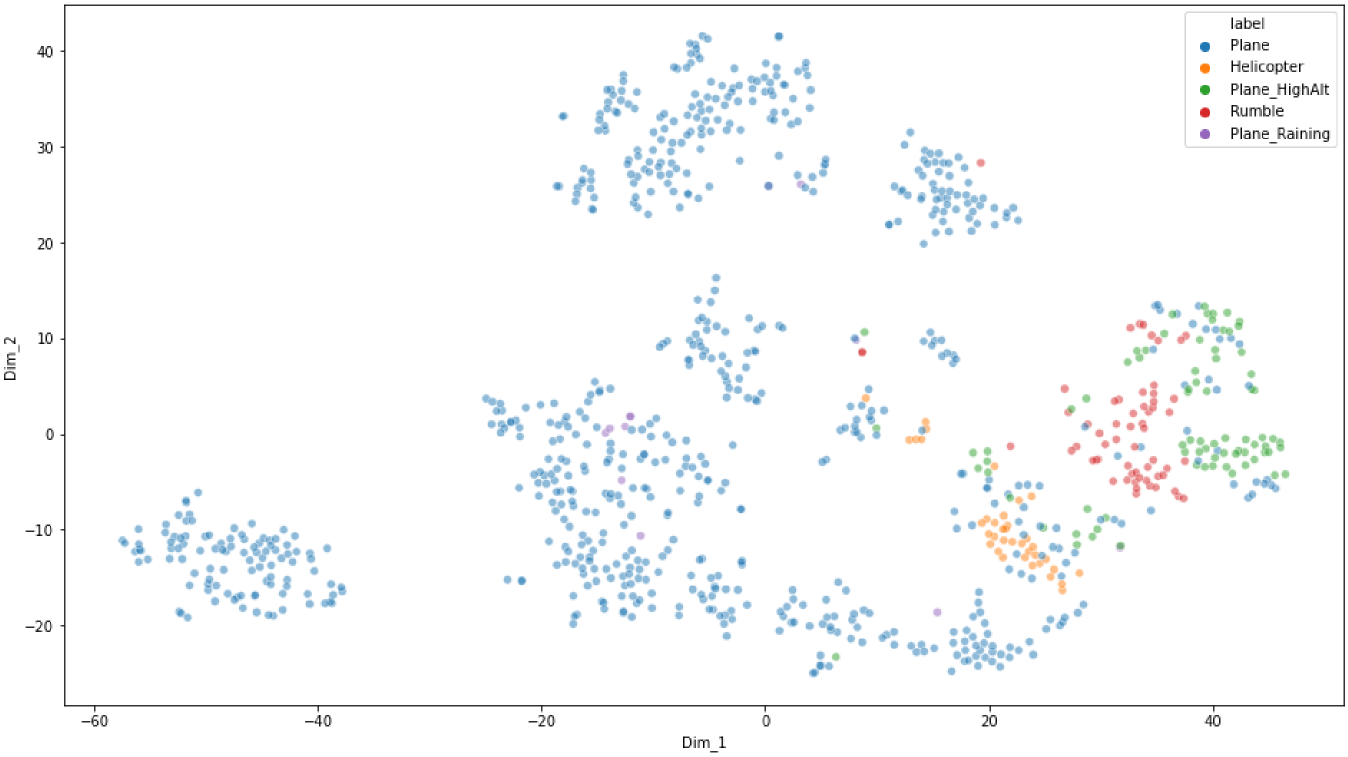


Figure 4: TSNE plot on the FFT-Z feature, number of iterations = 1000

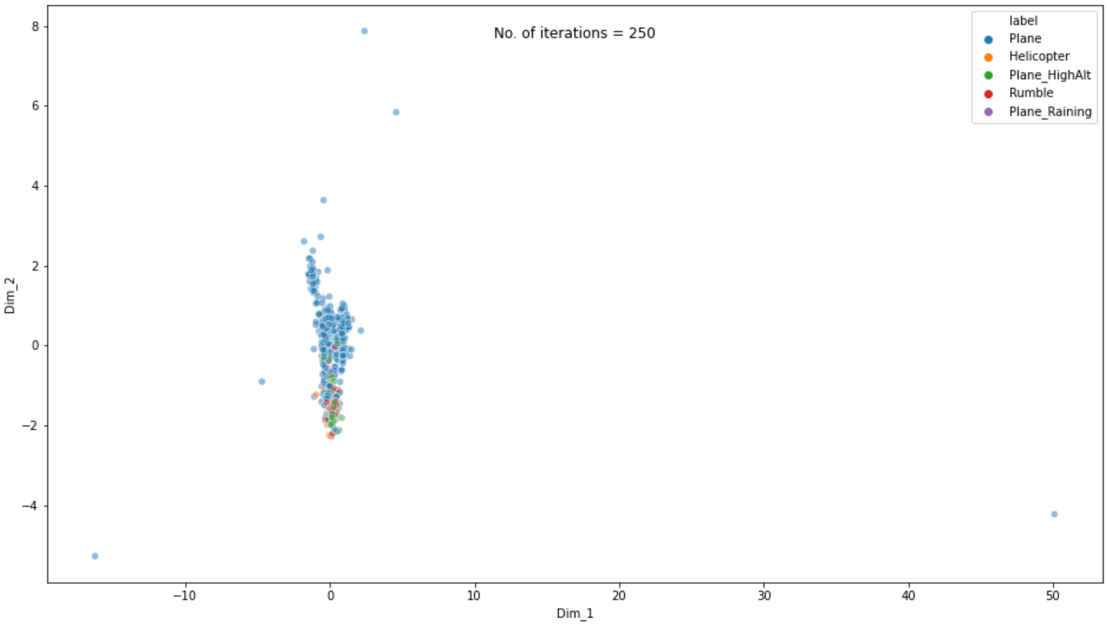


Figure 5: Dynamic plot of the FFT-Z feature, iterations range from 250 to 10000

3rd-Octave-A

111 3rd-Octave-A bins > PCA > 50 principal components > TSNE > 2D Visualization

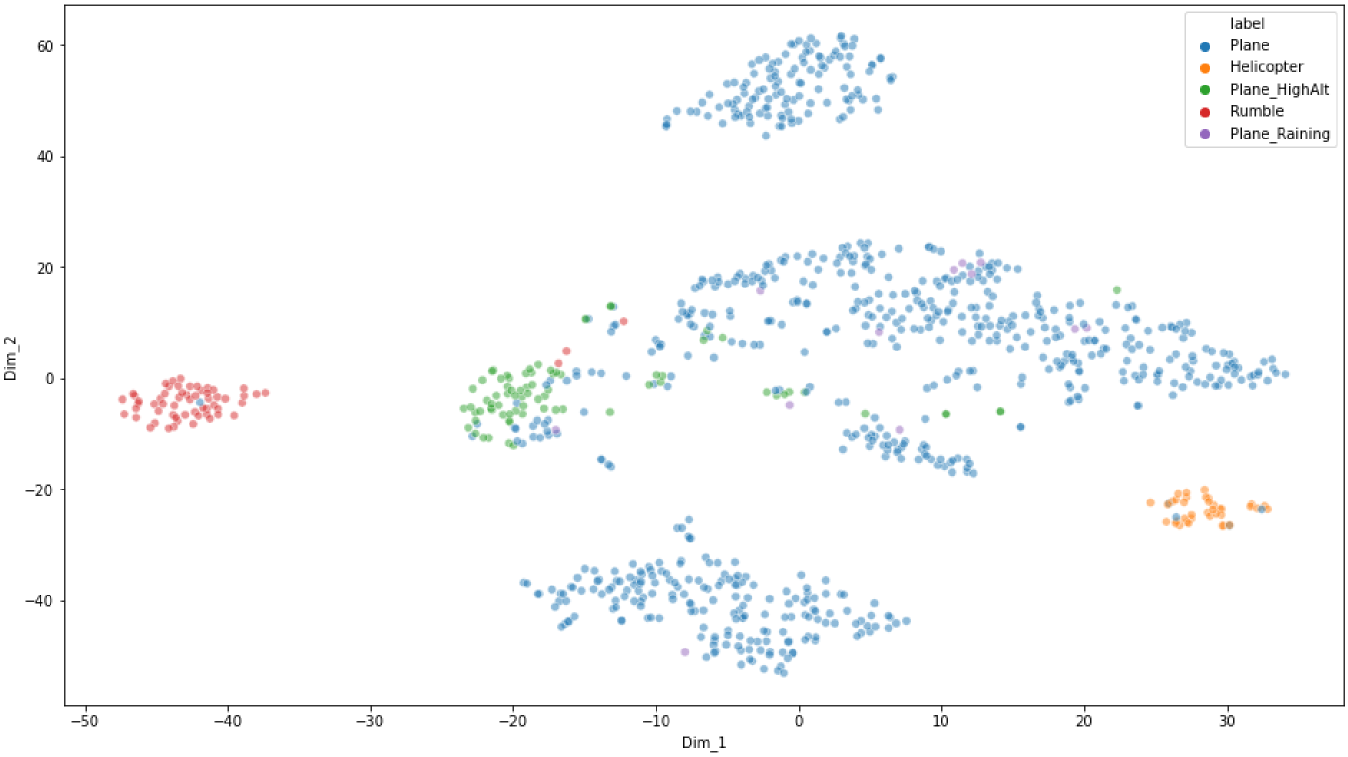


Figure 6: TSNE plot on the 3rd-Octave-A feature, number of iterations = 1000

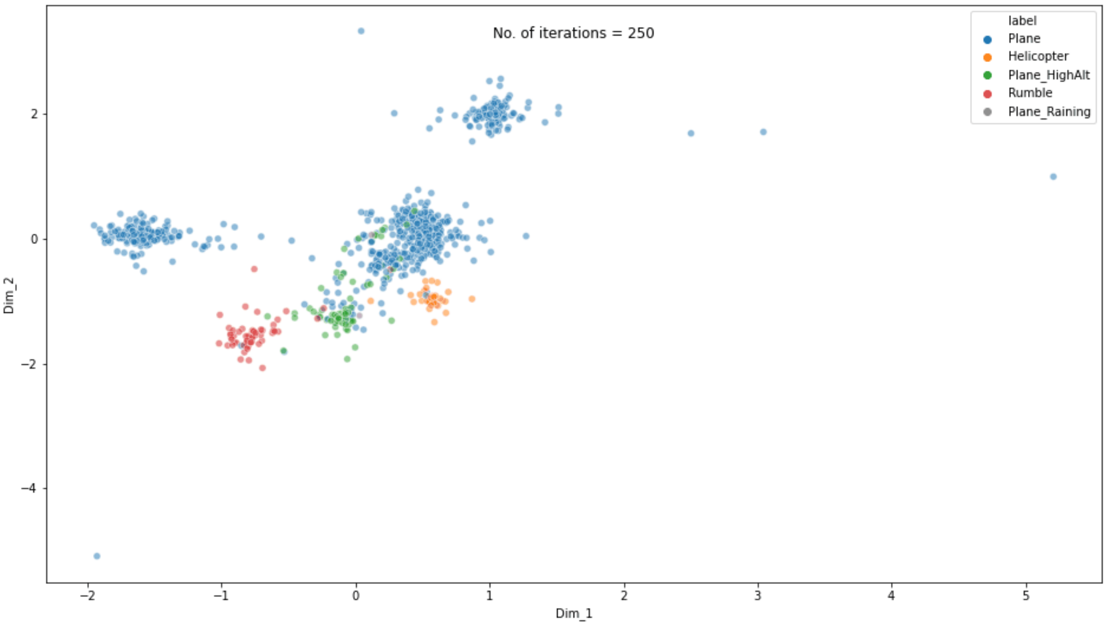


Figure 7: Dynamic plot of the 3rd-Octave-A feature, iterations range from 250 to 10000

3rd-Octave-Z

111 3rd-Octave-Z bins > PCA > 50 principal components > TSNE > 2D Visualization

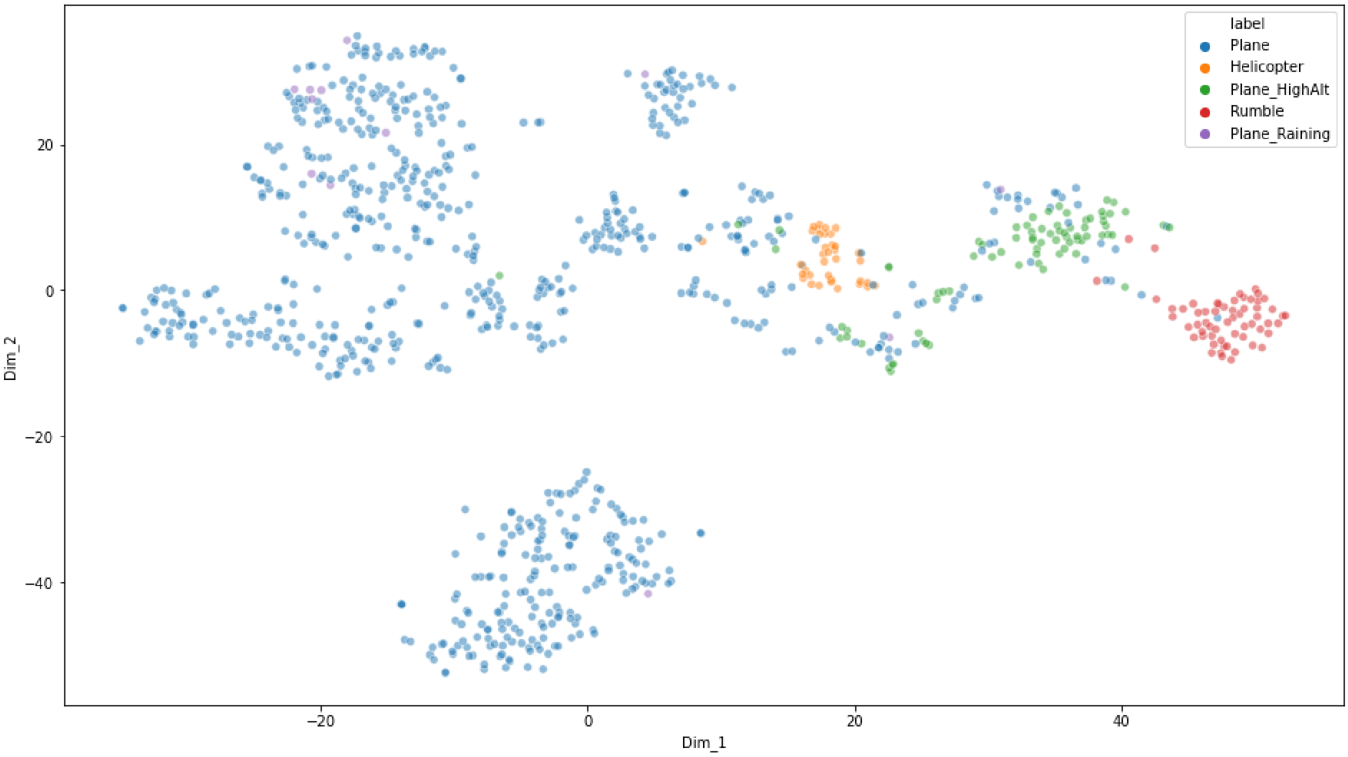


Figure 8: TSNE plot on the 3rd-Octave-Z feature, number of iterations = 1000

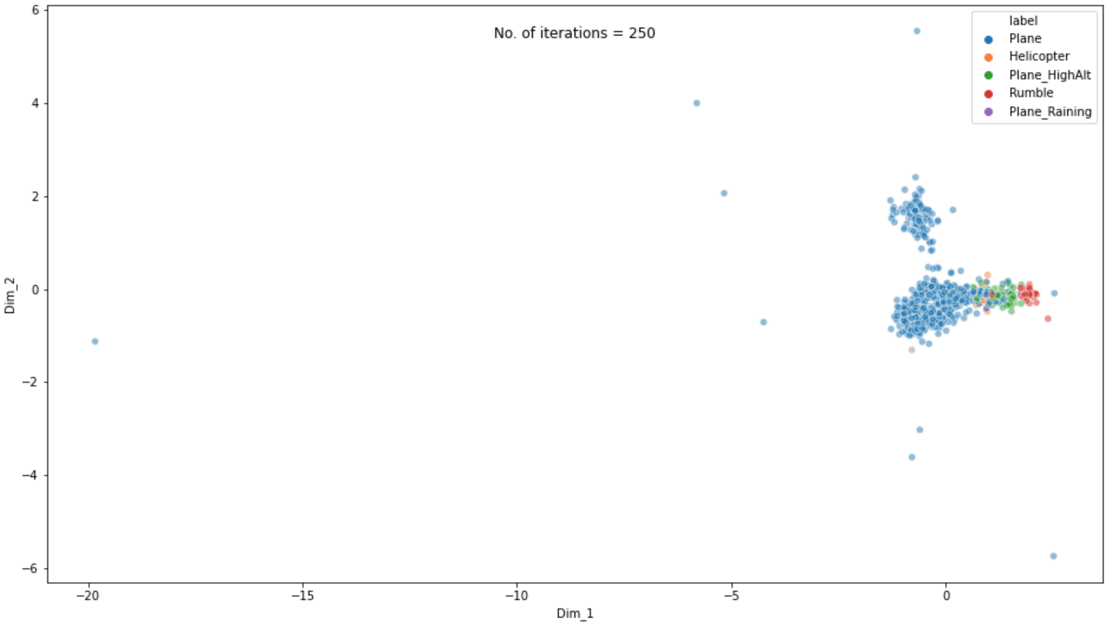


Figure 9: Dynamic plot of the 3rd-Octave-Z feature, iterations range from 250 to 10000

SPL-A

For the SPL feature, different audio length of the provided files yield different number of observable data points. Segmentation of data is required to fix the data points in order to fit the algorithm. SPL-A is segmented to fit the nearest 6000 data points, giving the shape of (918, 6000) representing the number of entries and SPL data points respectively.

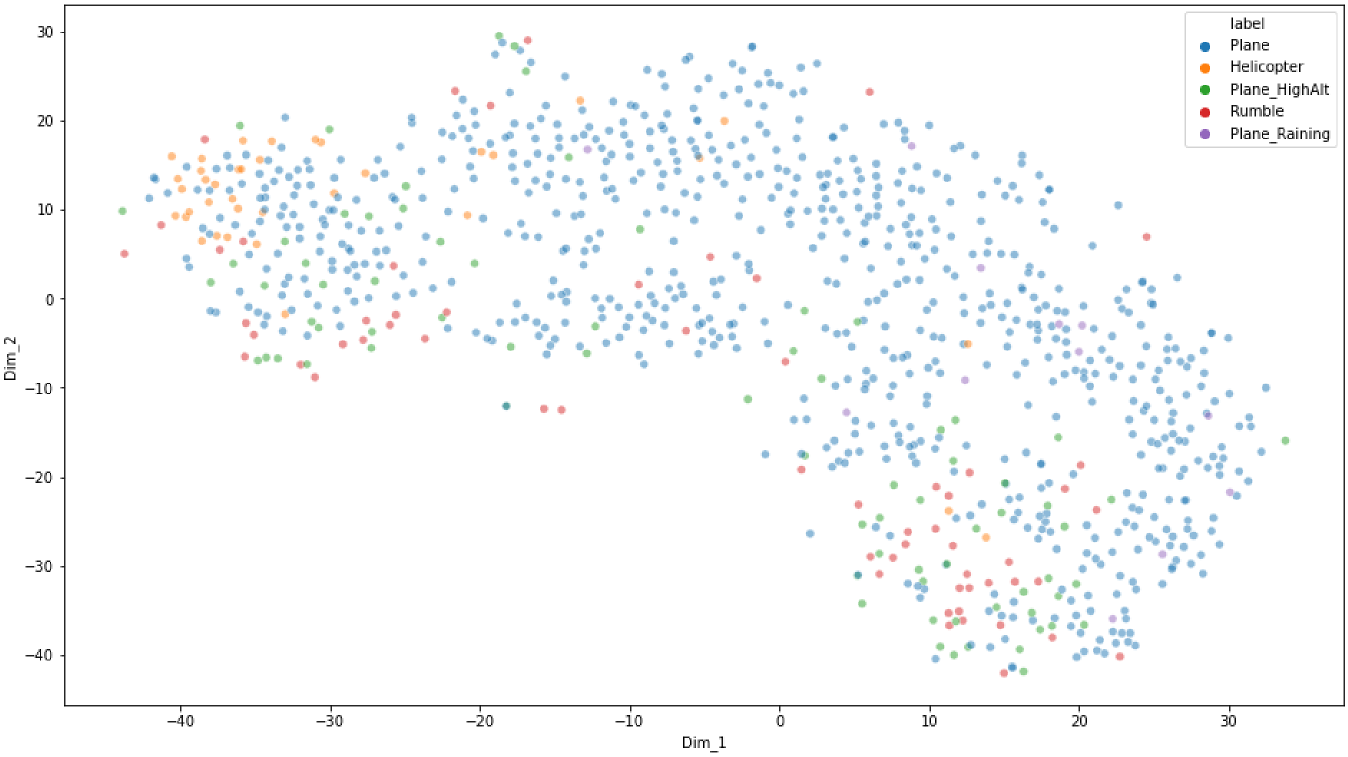


Figure 10: TSNE plot on the SPL-A feature, number of iterations = 1000

The figure above does not show a good correlation for the different classes based on the SPL feature. It is assumed that the very nature of SPL for say "Helicopter" class could very likely be of a similar level to that of a "Plane" or "Rumble" class data.

Combined features

3rd-Octave(A) and 3rd-Octvate(Z)

Concatenated 222 3rd-Octave-A and Z bins > PCA > 50 principal components > TSNE > 2D Visualization

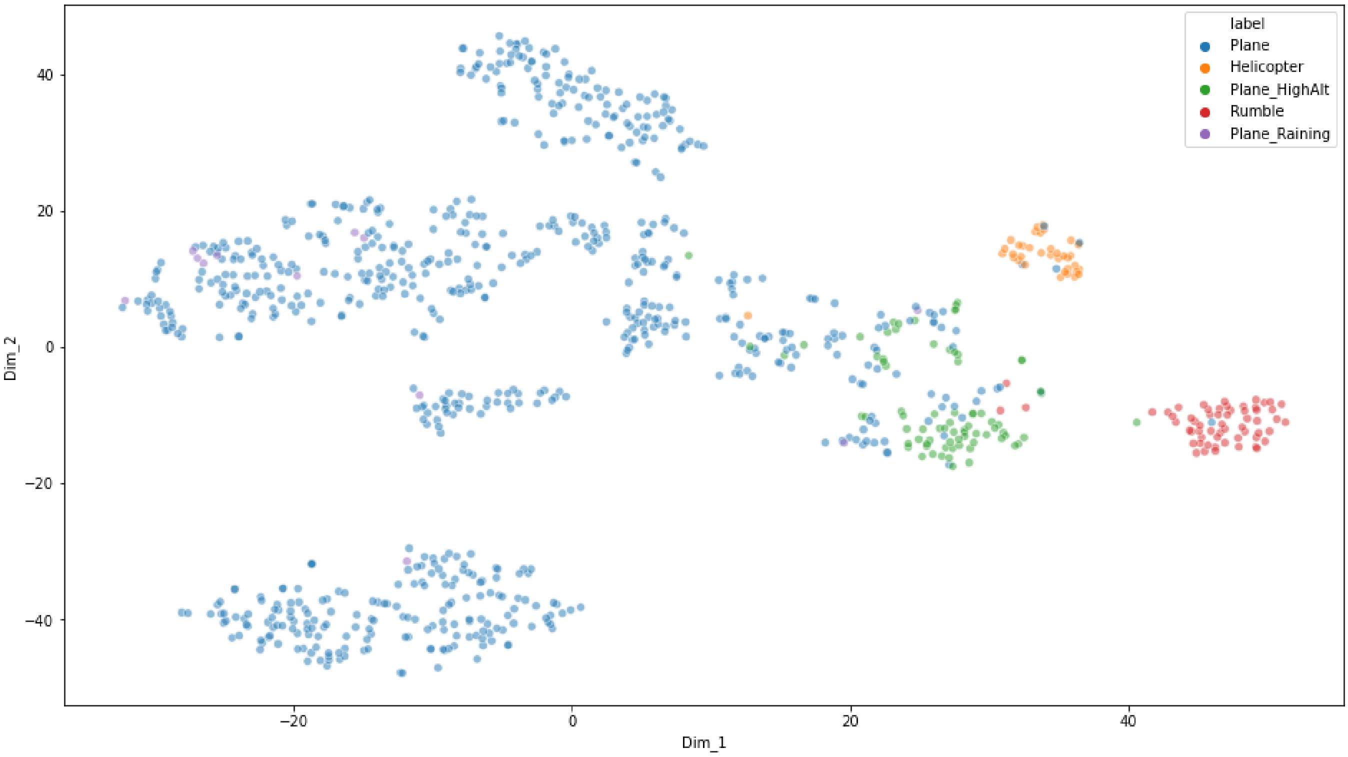


Figure 10: TSNE plot on the 3rd-Octave-A and Z features, number of iterations = 1000

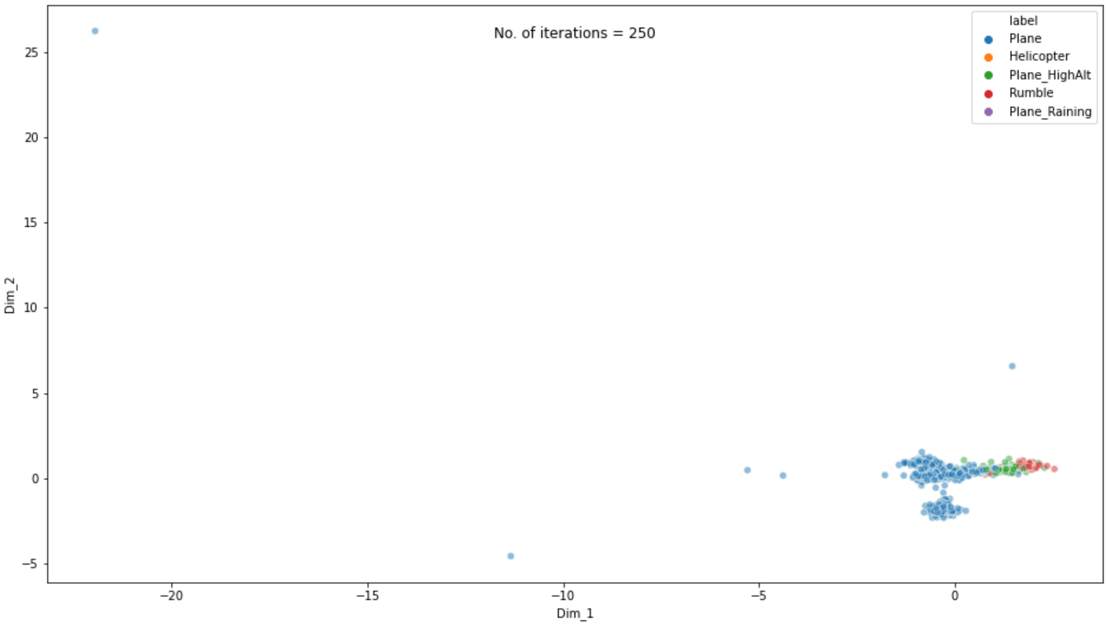


Figure 11: Dynamic plot of the 3rd-Octave-A and Z features, iterations range from 250 to 10000

Conclusion

For both FFT(A, Z) and 3rd-Octave(A, Z) features, we can observe clusters in the 2D space after applying both the dimensionality reduction algorithms.

The 3rd-Octave(A, Z) features are able to exhibit good distinct clusters as visualized from the above figures.

On top of the ability to visualize a segregation between the 3 coarse classes ("Plane", "Helicopter", "Rumble"), new found clusters for "Plane" class can be observed from the figures.

Both figures shown below describe the new clusters observed that could be of different signatures for a different range of filters to be applied on during operation.

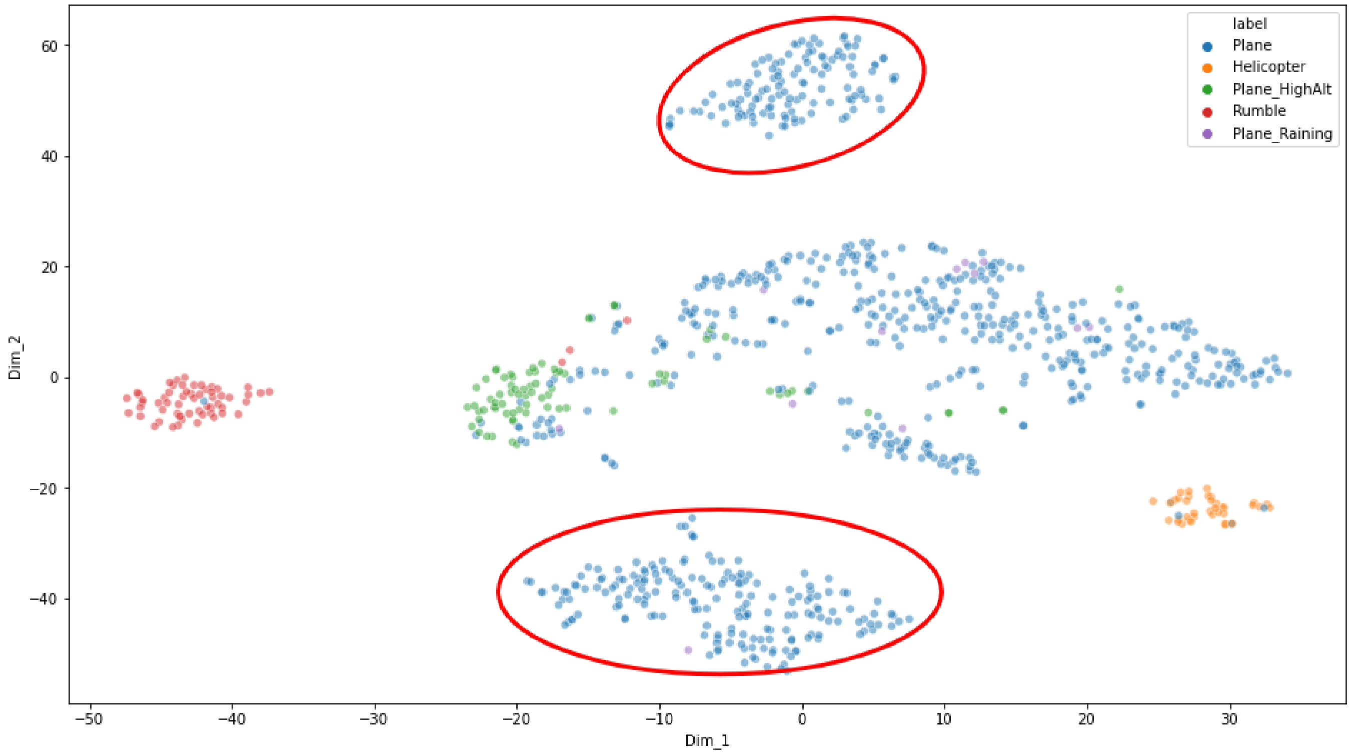


Figure 12: New clusters observed on TSNE plot for 3rd-Octave-A features

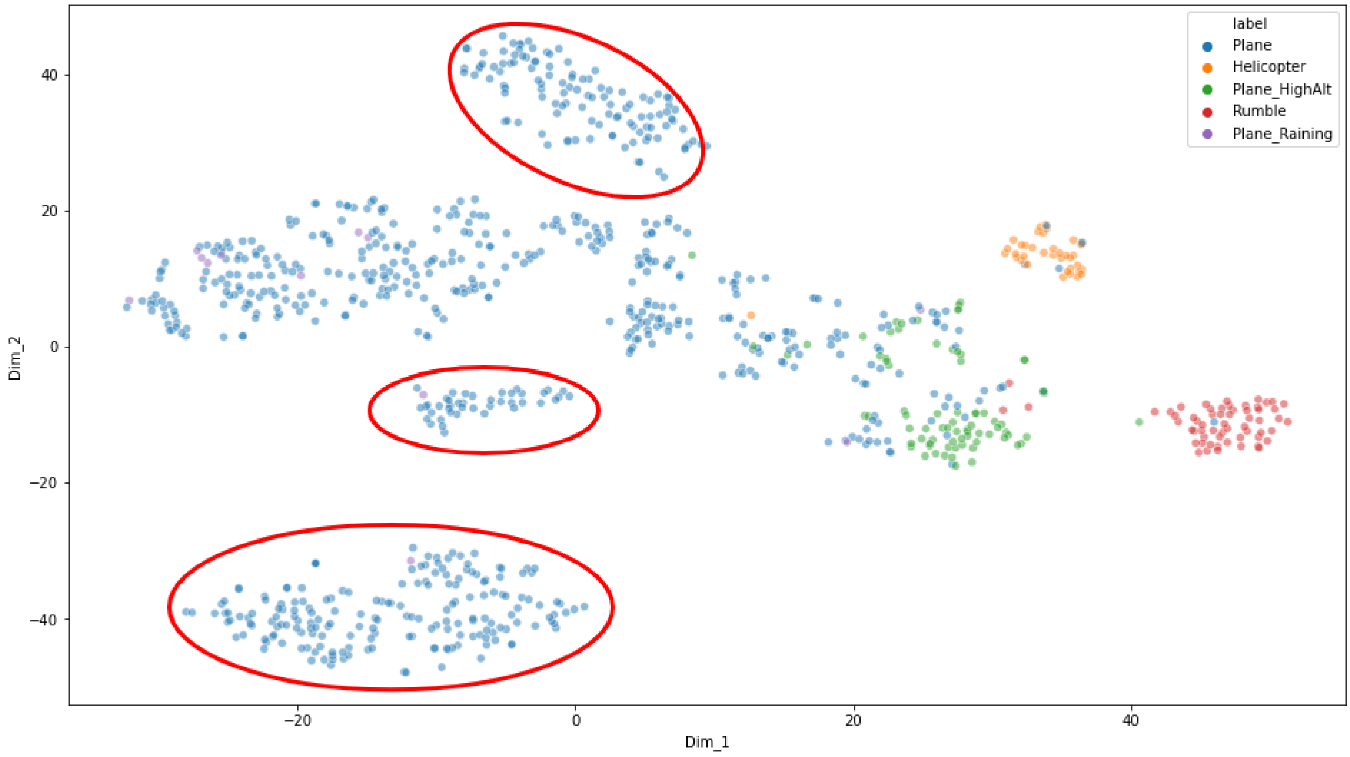


Figure 13: New clusters observed on TSNE plot for 3rd-Octave-A and Z features

Future work

Experiment on other clustering algorithms and techniques. Experiment on employing variational autoencoder (VAE) for deep model to learn and cluster with the latent space by employing reconstruction loss for optimization.

Reference

[1] https://blog.umetrics.com/what-is-principal-component-analysis-pca-and-how-it-is-used#:~:text=Principal%20component%20analysis%2C%20or%20PCA,more%20easily%20visualized%20and%20analyzed

[2] https://mlexplained.com/2018/09/14/paper-dissected-visualizing-data-using-t-sne-explained/#:~:text=t%2DSNE%20creates%20a%20probability,distribution%20in%20low%2Ddimensional%20space.&text=t%2DSNE%20uses%20multiple%20tricks%20to%20try%20to%20avoid%20this%20problem