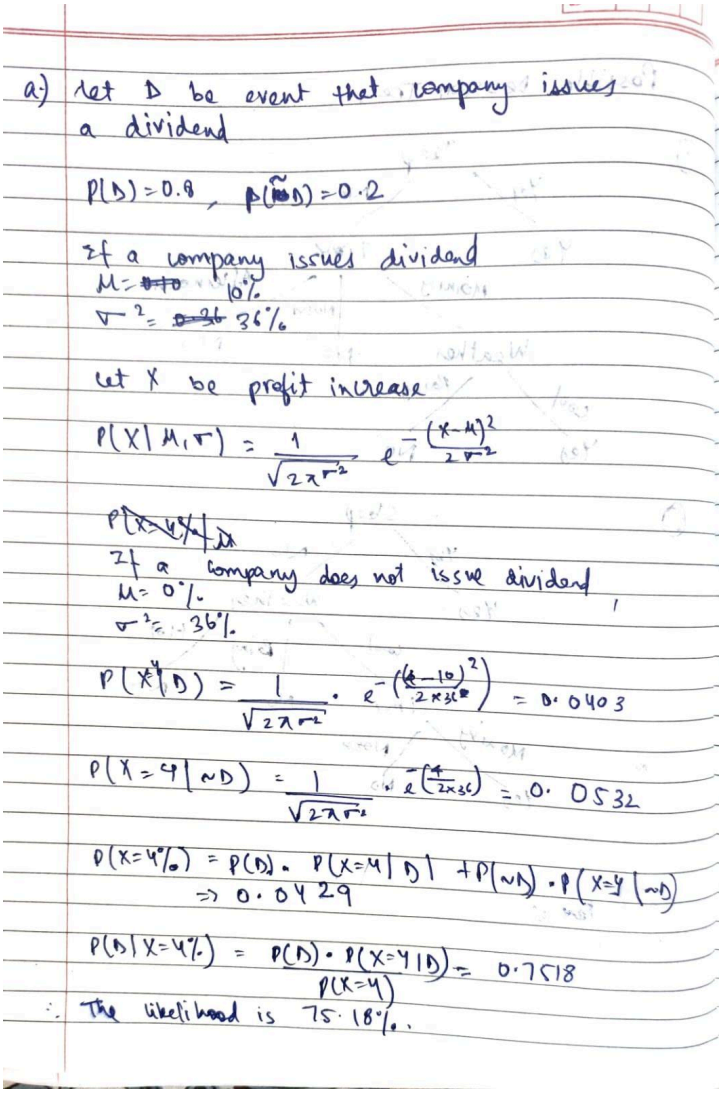
CSE343: Machine Learning
Assignment-2

REPORT

Anushka Srivastava (2022086)

SECTION A

a. The likelihood is 75.18%.



b. The decision trees are as follows:

b.) Step 1

$P(\text{Yes}) = 7/12$, $P(\text{No}) = 5/12$

$$H(Y) = -\frac{7}{12}\log_2\frac{7}{12} - \frac{5}{12}\log_2\frac{5}{12}$$

$$= 0.979$$

Class Timing

$$H(\text{Morning}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 0.811$$

$$H(\text{Noon}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$\Rightarrow 1$$

$$H(\text{Afternoon}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$\Rightarrow 1$$

$$IG = H(Y) - H(Y|X)$$

$$= 0.979 - \frac{4}{12} \times 0.811 - \frac{4}{12} - \frac{4}{12}$$

$$= 0.042$$

Sleep

$$H(\text{Sleep}=\text{Yes}) = \frac{6}{6}\log_2\frac{6}{6} = 0$$

$$H(\text{Sleep}=\text{No}) = \frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}$$

$$= 0.650$$

$$IG = 0.979 - \frac{6}{12}\times 0 - \frac{6}{12}\times 0.650$$

$$= 0.654$$

Weather

$$H(\text{Cool}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.721$$

$$H(\text{Rainy}) = 0 - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$$H(\text{Hot}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.970$$

$$IG = 0.979 - \frac{5}{12}(0.721) - 0 - \frac{5}{12}\times 0.970$$

$$\Rightarrow 0.274$$

Since sleep has maximum information gain, we use it as root node

$H(\text{sleep}=\text{yes}) = 0$; $H(\text{sleep}=\text{No}) = 0.650$

* Time | Sleep = No

H(Morning | Sleep = No)
$$\Rightarrow -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

H(Noon | Sleep = No)
$$= -0 - 0 = 0$$

H(Afternoon | Sleep = No)
$$\Rightarrow 0$$

$$IG = 0.654 - \frac{2}{6} = 0.321$$

Weather | Sleep = No

H(cool | Sleep = No) $= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

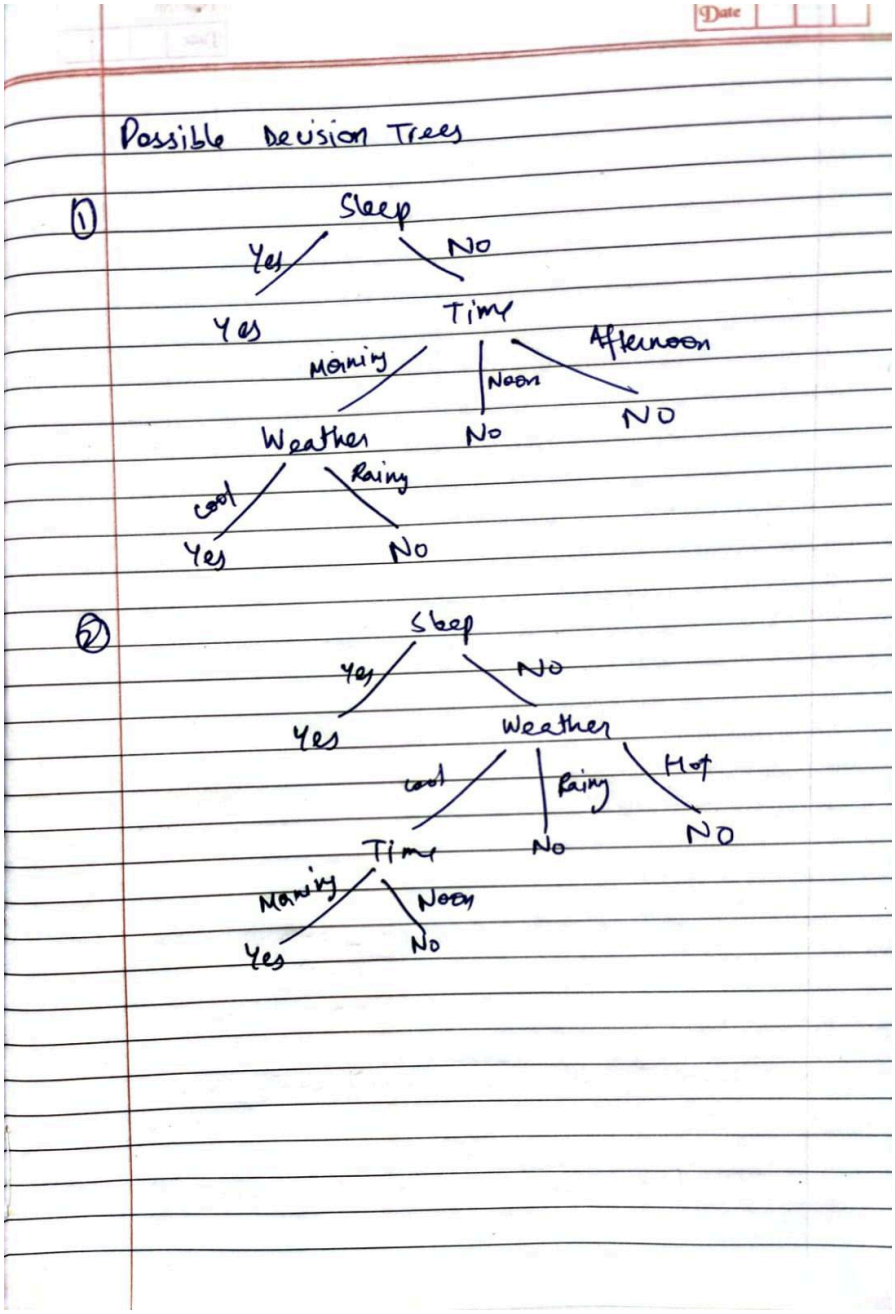
$\Rightarrow 1$

H(Rainy | Sleep = No) = 0

H(Hot | Sleep = No) = 0

$$IG = 0.654 - \frac{2}{6} = 0.321$$

The information gain for both Time and Weather is same and hence we can choose any parameter.



c. **Reference:** https://www.cs.cmu.edu/~avrim/ML10/lect0125.pdf

Page No.
Date

c. The number of mistakes M on S made by the Perceptron algorithm is at most $(1/\gamma)^2$ which is function of polynomial $(1/\gamma)$, where

$$\gamma = \min_{x \in S} \frac{|w^* \cdot x|}{\|x\|}$$

where w^* is a unit-length vector. If we scale examples to have Euclidean length 1, then γ is the minimum distance of any example to the plane $w^* \cdot x = 0$

Proof:

Claim 1: $w_{t+1} \cdot w^* \geq w_t \cdot w^* + \gamma$

Every time, a mistake is made, the length

Every time, a mistake is made, the dot product of our weight vector with the target increases by atleast γ.

If x was a positive example, then we get $w_{t+1} \cdot w^* = (w_t + x) \cdot w^* = w_t \cdot w^* + x \cdot w^* \geq w_t \cdot w^* + \gamma$.

Similarly, if x was a negative example, we get $(w_t - x) \cdot w^* = w_t \cdot w^* - x \cdot w^* \geq w_t \cdot w^* + \gamma$.

Claim 2: $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$.

Every time, a mistake is made, the length squared of our weight vector increases by atmost 1.

If x was a positive example, we get $\|w_t + x\|^2 = \|w_t\| + 2w_t \cdot x + \|x\|^2$. This is less than $\|w_t\|^2 + 1$ because $w_t \cdot x$ is negative.

Page No.
Date

Claim 1 implies that after M mistakes, $w_{M+1} \cdot w^* \geq \gamma M$.
On other hand, Claim 2 implies that after M mistakes, $\|w_{M+1}\| \leq \sqrt{M}$. Now, all we need to do is use the fact that $w_t \cdot w^* \leq \|w_t\|$, since w^* is a unit vector. So, this means we must have $\gamma M \leq \sqrt{M}$, and thus $M \leq 1/\gamma^2$

The number of mistakes (including margin mistakes) made by Margin Perceptron (γ) on S is at most $8/\gamma^2$.
As before, each update increases $w_t \cdot w^*$ by atleast γ. For the original algorithm, we had
$\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$, which implies
$\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2\|w_t\|}$

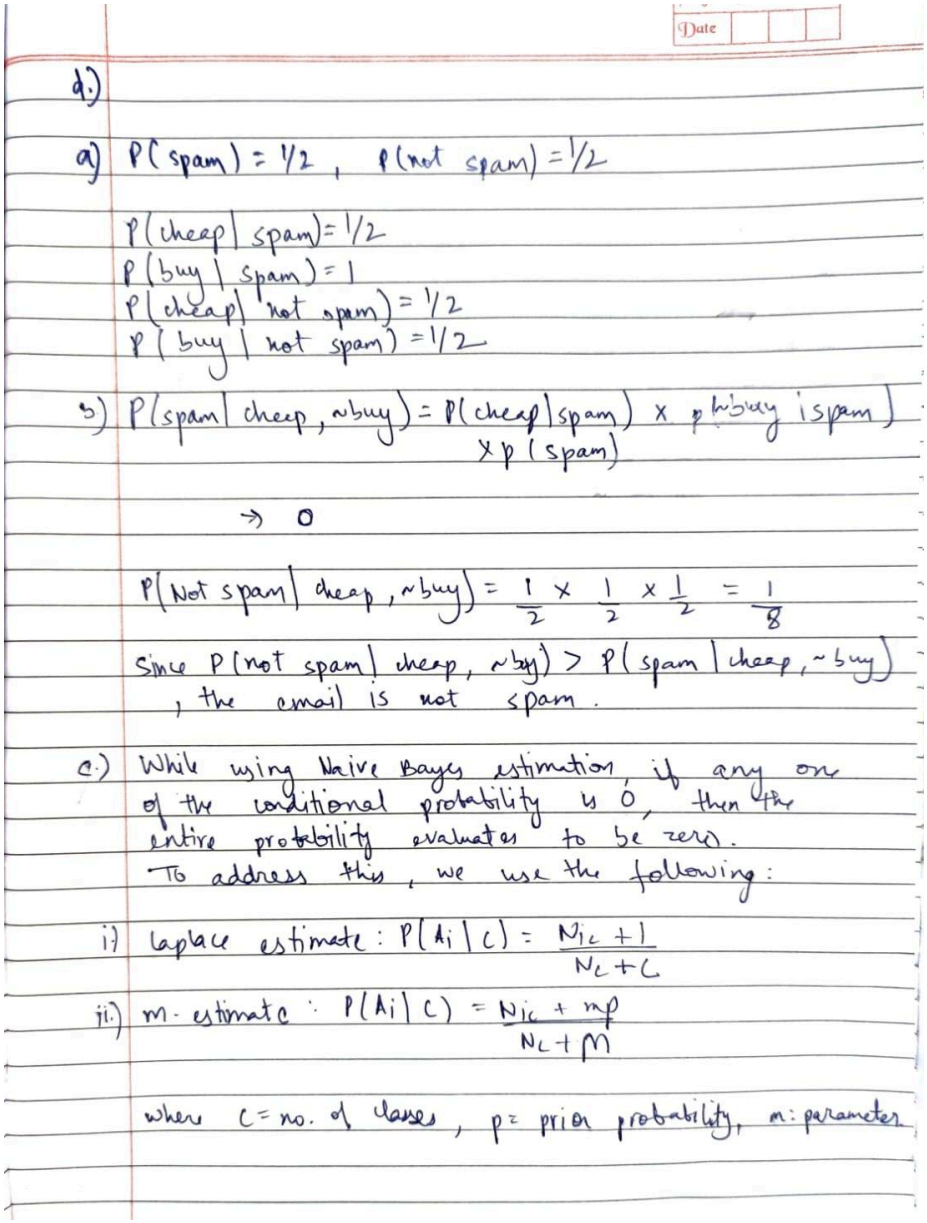
For the new algorithm, we get
$$\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2\|w_t\|} + \frac{\gamma}{2}$$

which we can see by breaking each x into its orthogonal part and its parallel part. We can now solve this directly, but just to get a simple upper bound, just notice that if $\|w_t\| \geq 2/\gamma$, then $\|w_{t+1}\| \leq \|w_t\| + 3\gamma/4$
So, after M updates, we have:
$$\|w_{M+1}\| \leq \frac{2}{\gamma} + \frac{3M\gamma}{4}$$

Solving $M\gamma \leq 2/\gamma + 3M\gamma/4$, we get

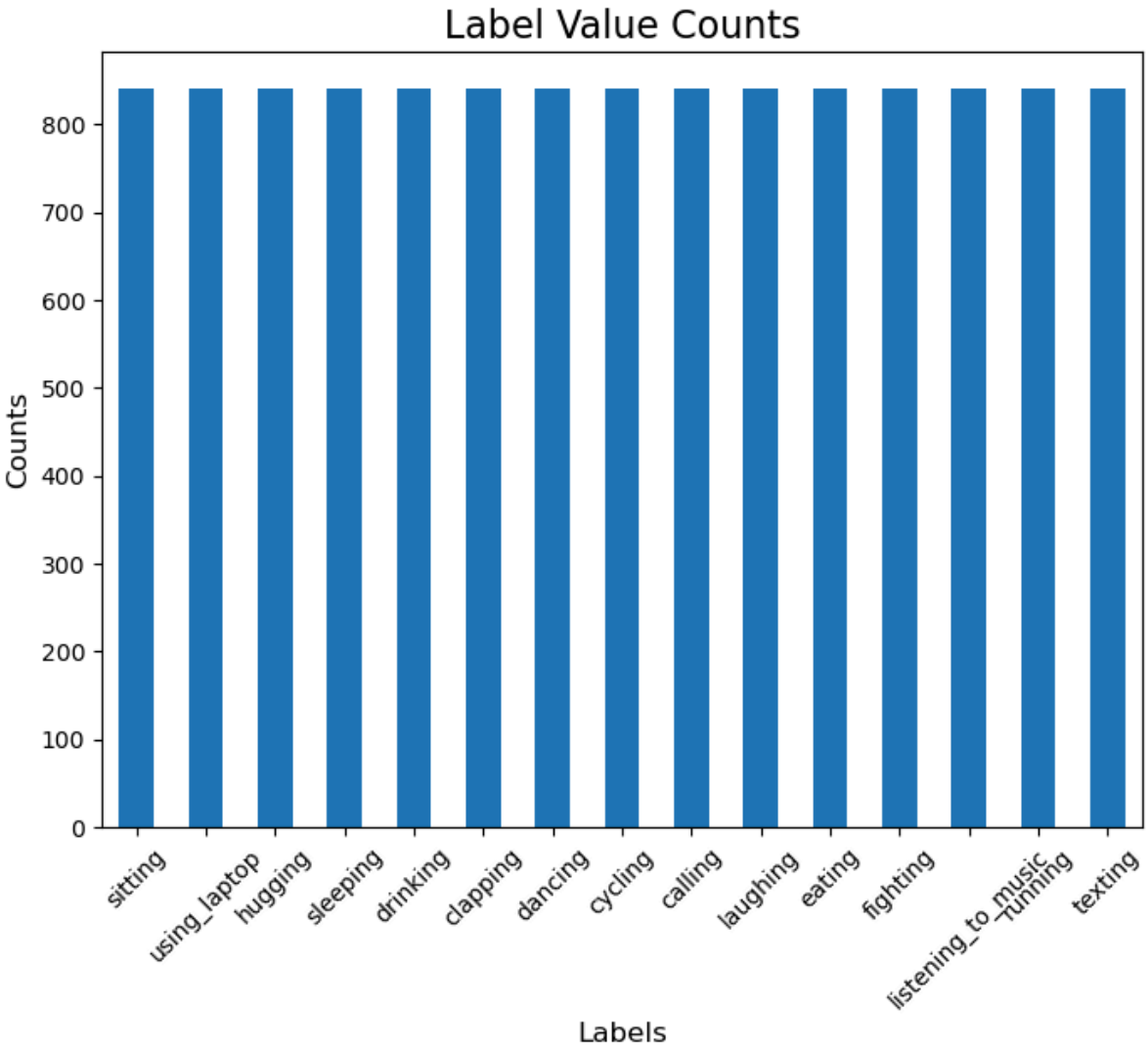
$M \leq 8/\gamma^2$, as desired.

d. Answer

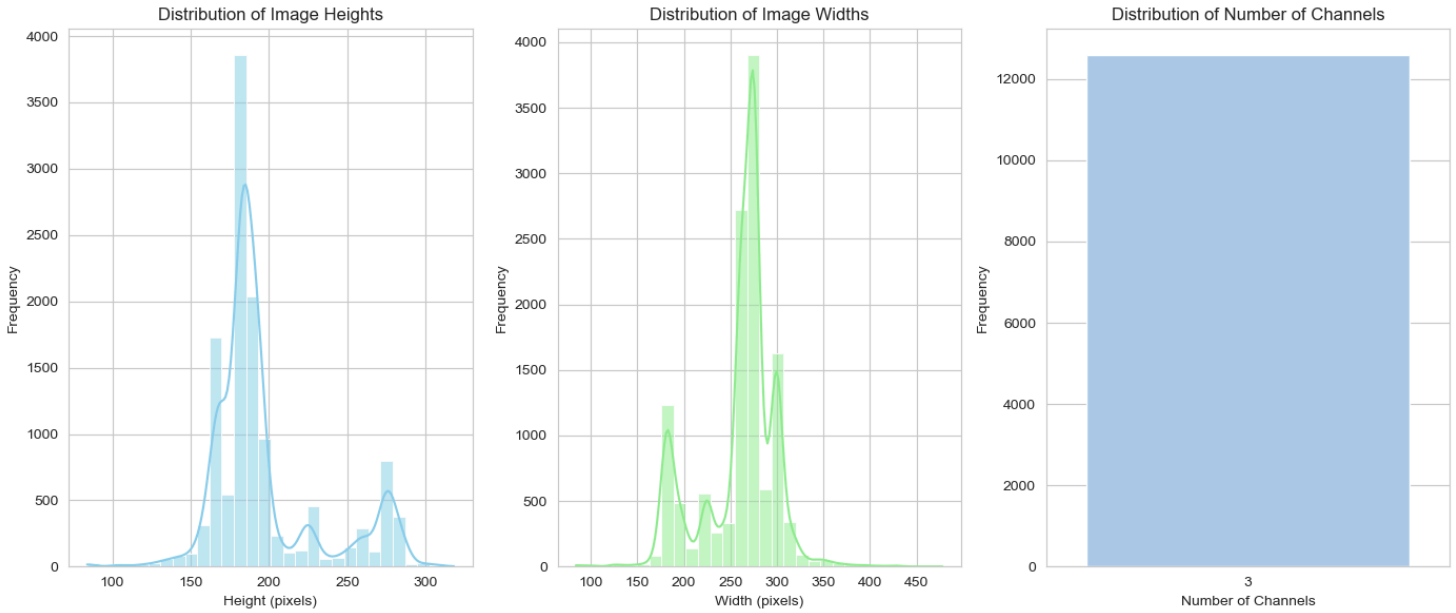


SECTION C

1. EDA
    a. We create a bar plot of Labels vs Counts.

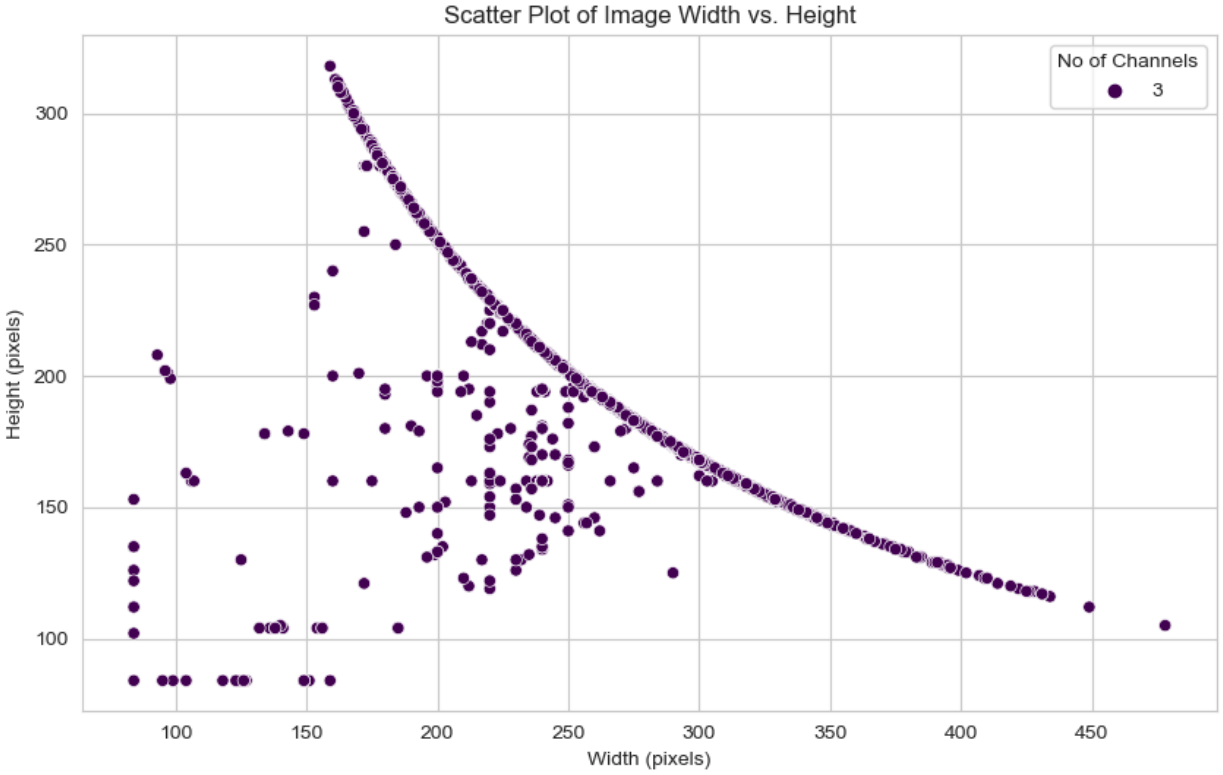


We notice that every class has an equal number of images, that is 840. Hence, we do not notice any class imbalance in our dataset.
We plot graphs for the distribution of image sizes and the number of channels.



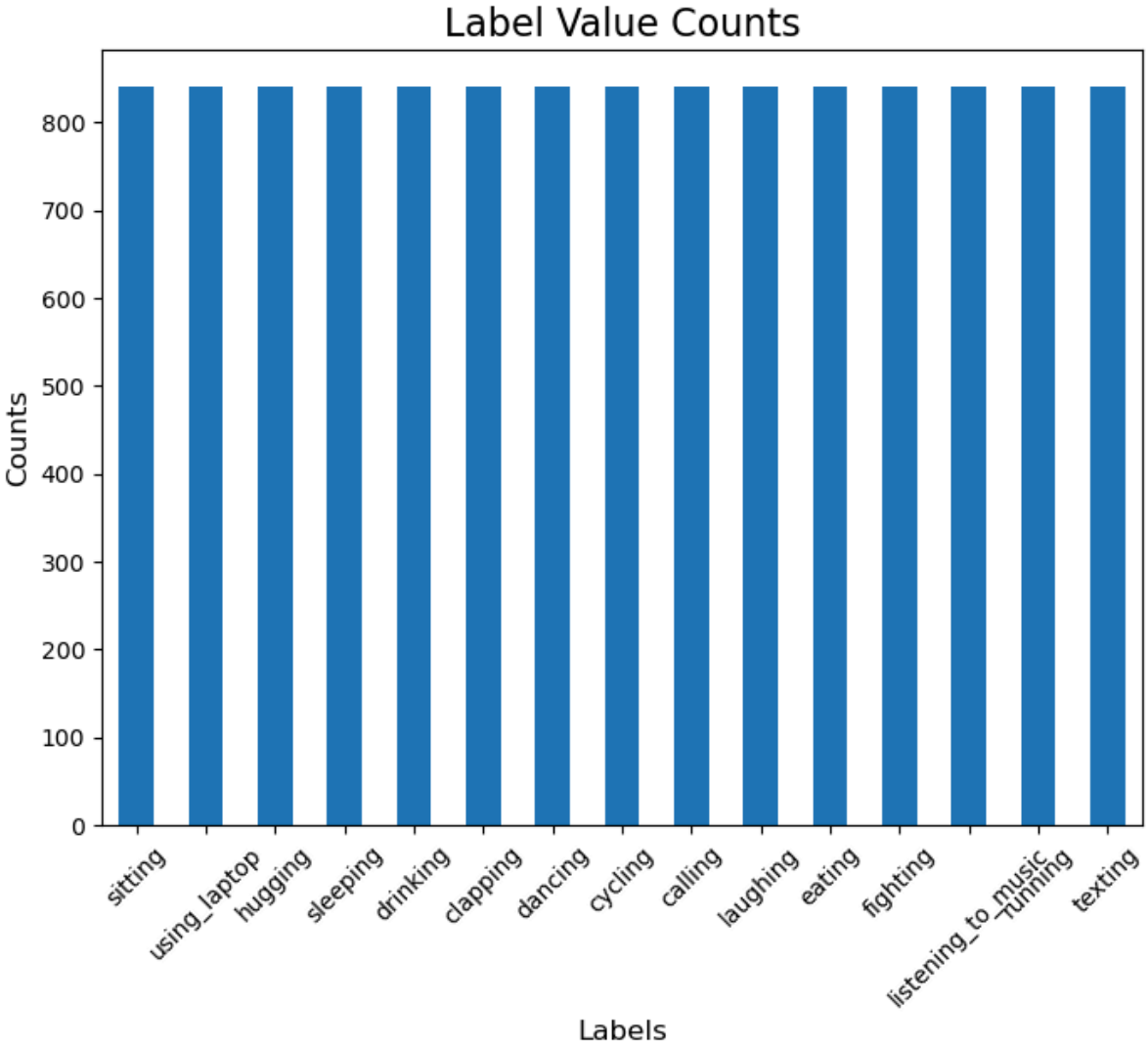
We notice that the height and weight follows a nearly normal distribution. Most images have a height between 150-200 pixels and a width between 250-300 pixels. All images have 3 channels.

We also plot a relationship between height and width.



We notice that the height and width have an inverse relationship, that is, when the width increases, the height decreases, and vice versa.

b. Visualization to display the distribution of classes



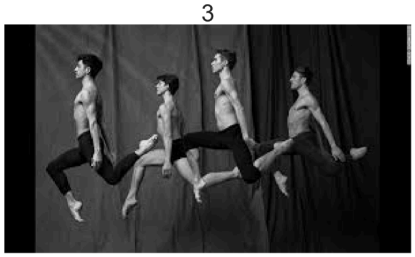
Each class has an equal number of objects.
We display 1 image from each class.

0

1

2

3



4

5

6

7

8

9

10

11

12

13

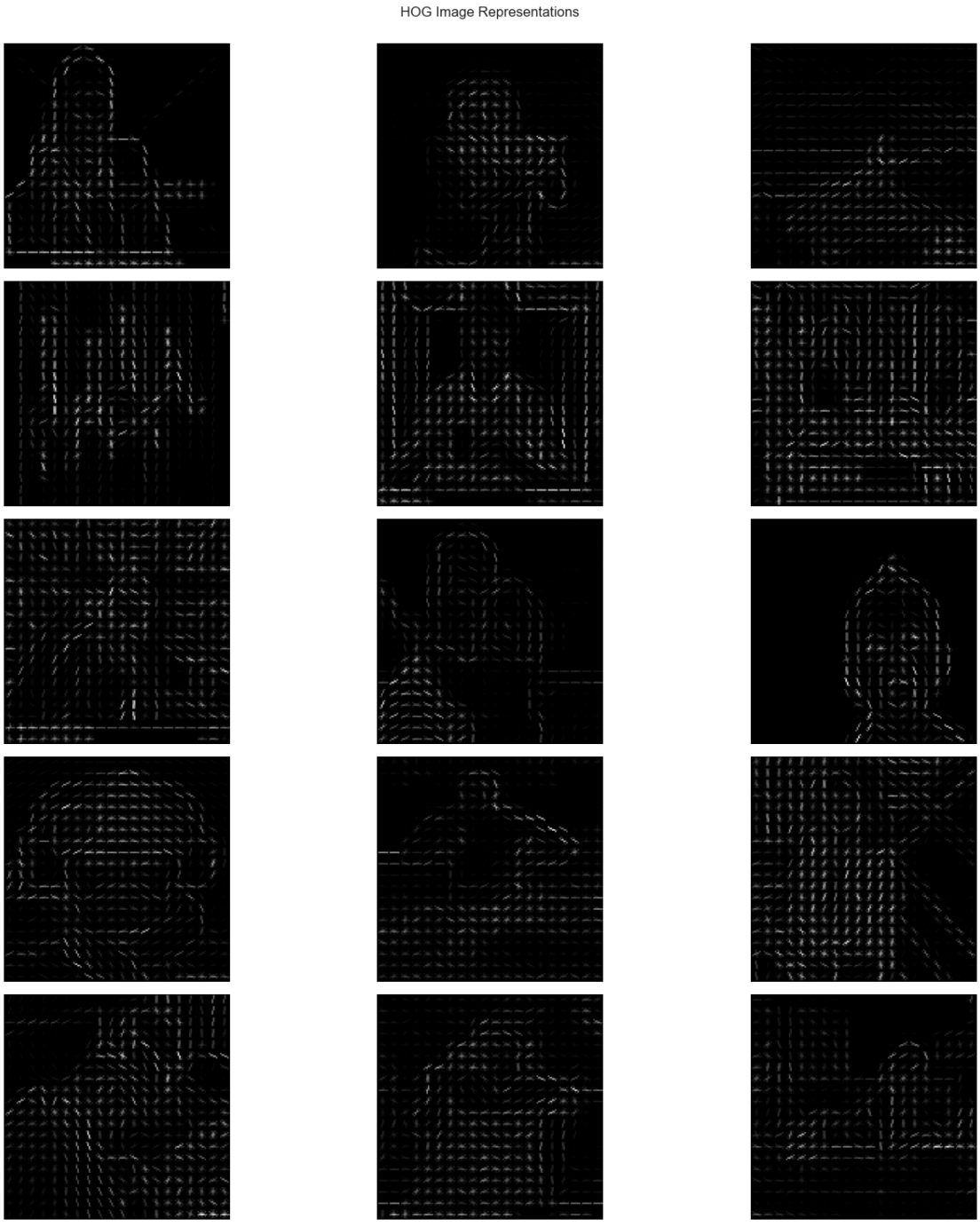
14

Our findings of an inverse relationship between height and width are verified as we notice from these images that the width decreases as height increases and vice versa.

c. From our bar plot, we notice that there are no class imbalances in our dataset, and hence, there is no need for any resampling or augmentation techniques.

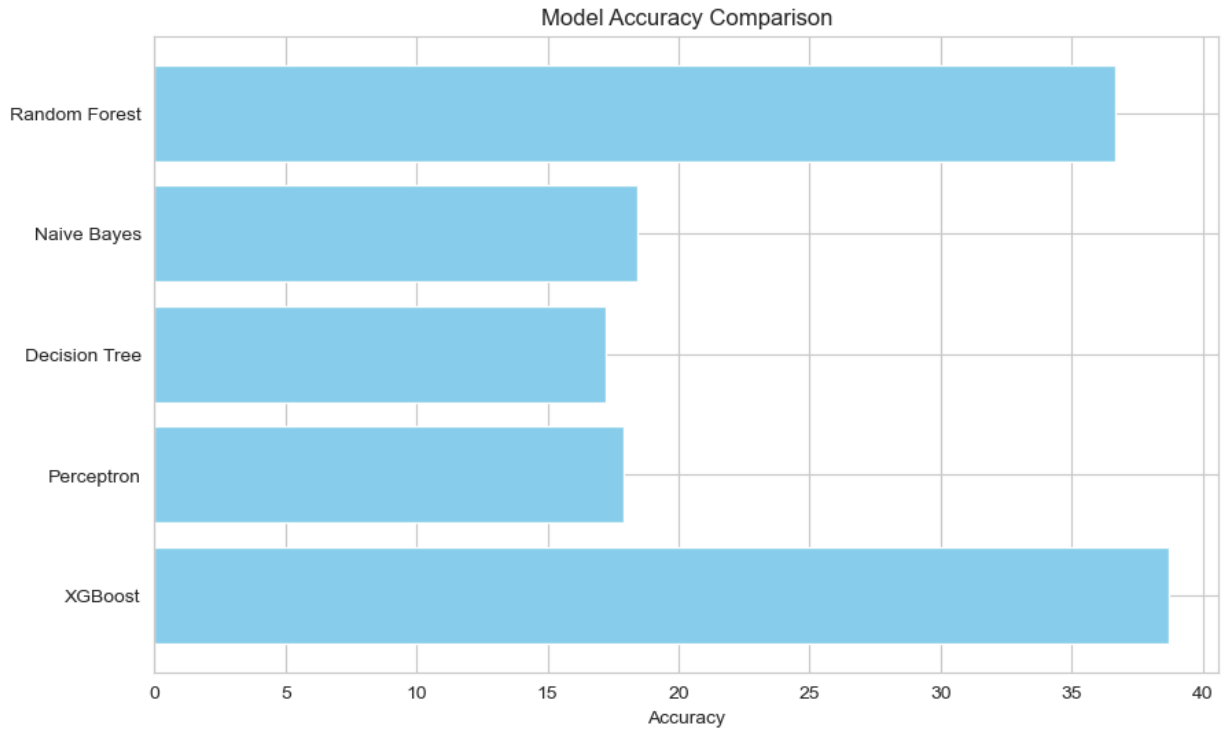
2. Feature Extraction

We perform Local Binary Pattern, Color Histogram, and SIFT Feature Extraction on our dataset. We also extract the Histogram of Oriented Gradients (HOG) from selected images for visualization purposes, but we do not use the extracted features from HOG to train our dataset.



We noticed that the features extracted using LBP, Color Histogram, and SIFT are of constant length. LBP returns a Numpy array of length 26. Color Histogram returns a Numpy array of length 512, and SIFT returns a Numpy array of length 128.

3. Model Training
   We train our dataset on Random Forest, Naive Bayes, Decision Tree, Perceptron, and XGBoost. The following accuracies are observed:
   - Random Forest: 36.67%
   - Naive Bayes: 18.45%
   - Decision Tree: 17.22%
   - Perceptron: 17.90%
   - XGBoost: 38.69%



XGBoost gives the best accuracy, followed by Random Forest.

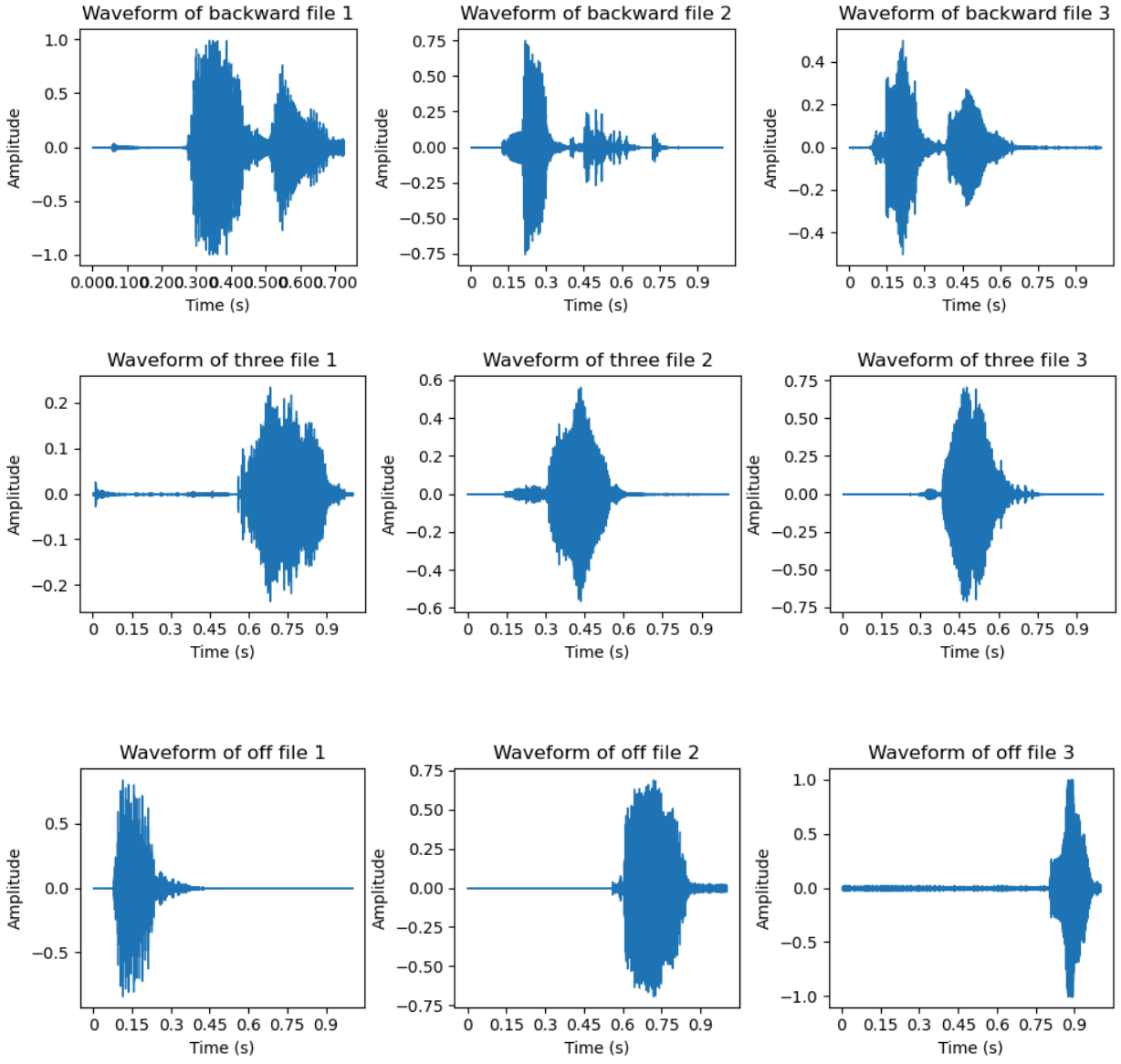
SECTION B (BONUS)

1. EDA
    a. Statistical Summary

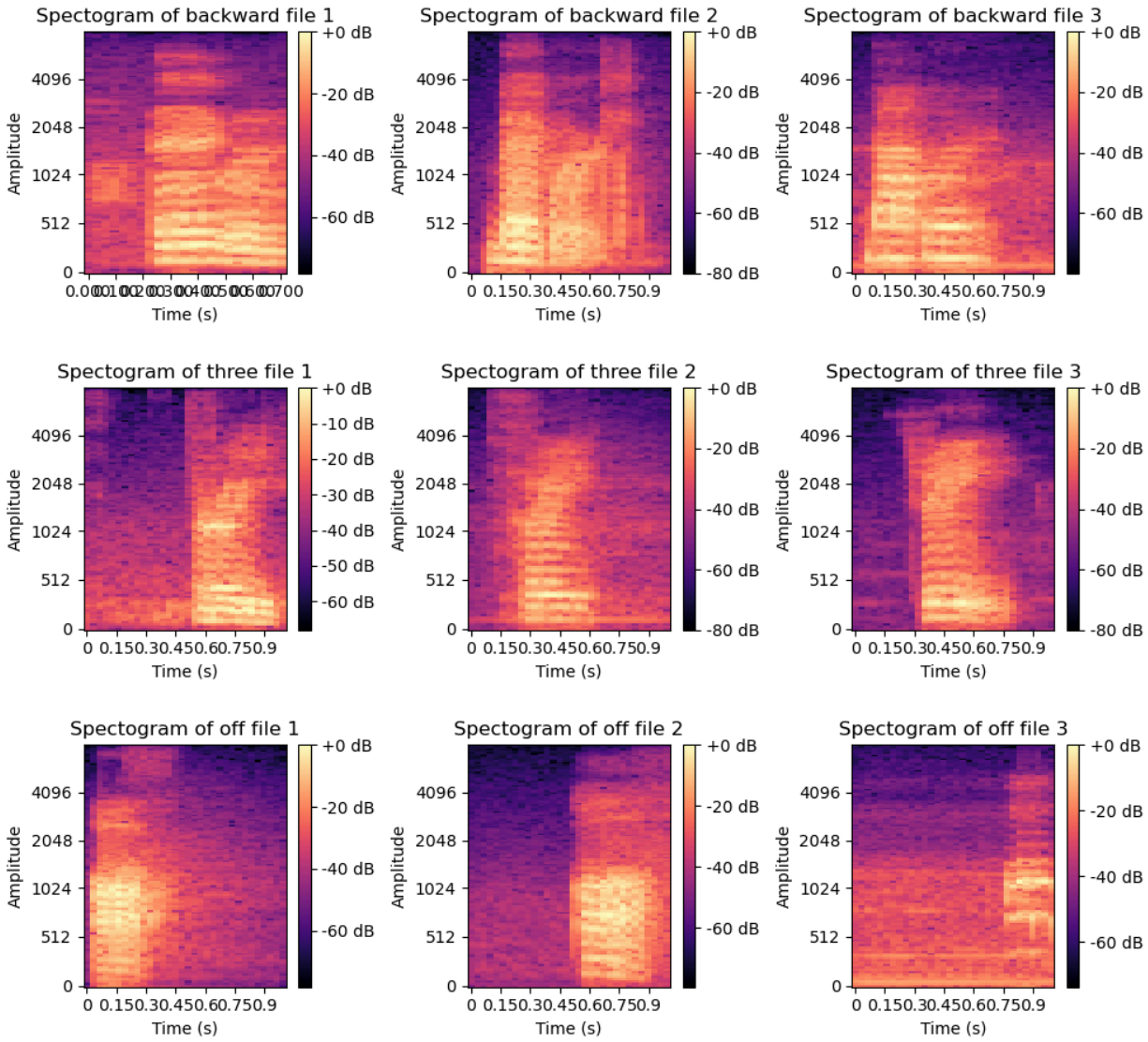
	mean_amplitude	min_amplitude	max_amplitude	std_amplitude	mean_duration	min_duration	max_duration	std_duration
backward	8.732750e-05	-1.0	0.999969	0.088580	0.986390	0.448000	1.0	0.062522
bed	6.745070e-05	-1.0	0.999969	0.085238	0.970498	0.213312	1.0	0.087145
bird	-1.830991e-04	-1.0	0.999969	0.095966	0.969454	0.325062	1.0	0.092588
cat	-2.848711e-04	-1.0	0.999969	0.075283	0.972065	0.384000	1.0	0.082372
dog	-4.608297e-04	-1.0	0.999969	0.093553	0.972226	0.426625	1.0	0.084865
down	4.458849e-05	-1.0	0.999969	0.087863	0.983559	0.325062	1.0	0.067175
eight	-2.228617e-05	-1.0	0.999969	0.076257	0.980846	0.256000	1.0	0.072684
five	1.695451e-05	-1.0	0.999969	0.092054	0.984320	0.384000	1.0	0.065863
follow	-1.721267e-04	-1.0	0.999969	0.103674	0.981981	0.341313	1.0	0.074247
forward	8.766092e-05	-1.0	0.999969	0.107183	0.984746	0.384000	1.0	0.066149
four	-1.378545e-04	-1.0	0.999969	0.095869	0.982999	0.278625	1.0	0.069495
go	-1.678797e-04	-1.0	0.999969	0.096900	0.978847	0.341313	1.0	0.076190
happy	-1.958642e-04	-1.0	0.999969	0.073446	0.974747	0.384000	1.0	0.079458
house	1.365303e-04	-1.0	0.999969	0.082099	0.974842	0.371500	1.0	0.082296
learn	-1.415776e-05	-1.0	0.999969	0.099211	0.974443	0.371500	1.0	0.091885
left	2.245471e-05	-1.0	0.999969	0.080381	0.984954	0.298625	1.0	0.065198
marvin	-3.048353e-04	-1.0	0.999969	0.095987	0.977728	0.384000	1.0	0.075275
nine	-7.837252e-07	-1.0	0.999969	0.085355	0.984835	0.341313	1.0	0.066458
no	2.396618e-05	-1.0	0.999969	0.088122	0.980128	0.384000	1.0	0.073198
off	3.893072e-05	-1.0	0.999969	0.088514	0.984939	0.394688	1.0	0.060932
on	3.273442e-05	-1.0	0.999969	0.089415	0.981885	0.384000	1.0	0.069206
one	-6.647508e-06	-1.0	0.999969	0.085383	0.978958	0.325062	1.0	0.076265
right	3.956014e-05	-1.0	0.999969	0.083249	0.982247	0.384000	1.0	0.069393
seven	-1.123295e-04	-1.0	0.999969	0.082817	0.984920	0.371563	1.0	0.062049
sheila	-1.687832e-04	-1.0	0.999969	0.083635	0.977648	0.426625	1.0	0.075049
six	-3.445910e-05	-1.0	0.999969	0.069262	0.987542	0.384000	1.0	0.056673
stop	-1.697145e-05	-1.0	0.999969	0.080679	0.984549	0.362687	1.0	0.064572
three	-9.996239e-05	-1.0	0.999969	0.081729	0.983931	0.341313	1.0	0.064122
tree	7.703565e-06	-1.0	0.999969	0.075752	0.970302	0.298625	1.0	0.087430
two	-1.324027e-04	-1.0	0.999969	0.086349	0.981600	0.278625	1.0	0.071753
up	1.073004e-05	-1.0	0.999969	0.074206	0.976156	0.298625	1.0	0.076972
visual	1.372075e-05	-1.0	0.999969	0.091905	0.982301	0.341313	1.0	0.070831
wow	3.483538e-05	-1.0	0.999969	0.096585	0.970225	0.325062	1.0	0.089879
yes	-6.172316e-05	-1.0	0.999969	0.079941	0.983431	0.384000	1.0	0.067037
zero	-7.398344e-05	-1.0	0.999969	0.096609	0.986978	0.298625	1.0	0.057204

We notice that the minimum and maximum amplitude are the same for all audio classes. Mean and Standard Deviation are very low, indicating that the data seems normalized or already pre-processed.
The maximum duration is 1 s for all classes. The mean durations are consistent.

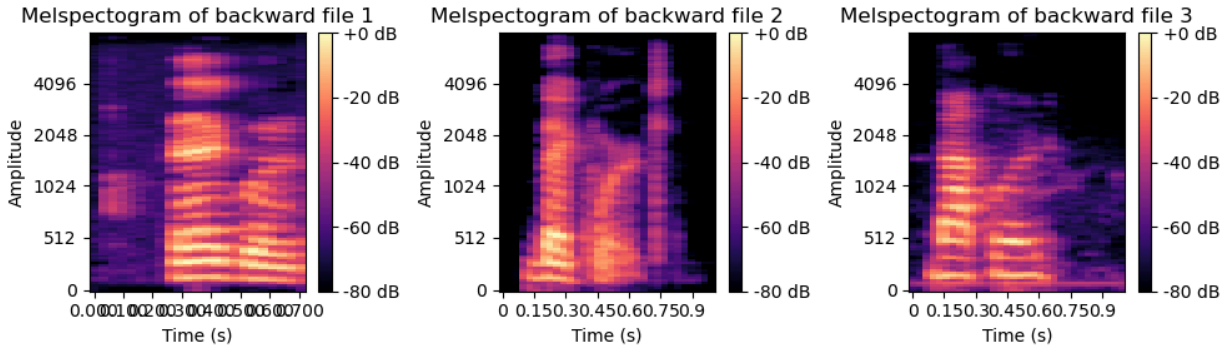
b. Waveforms

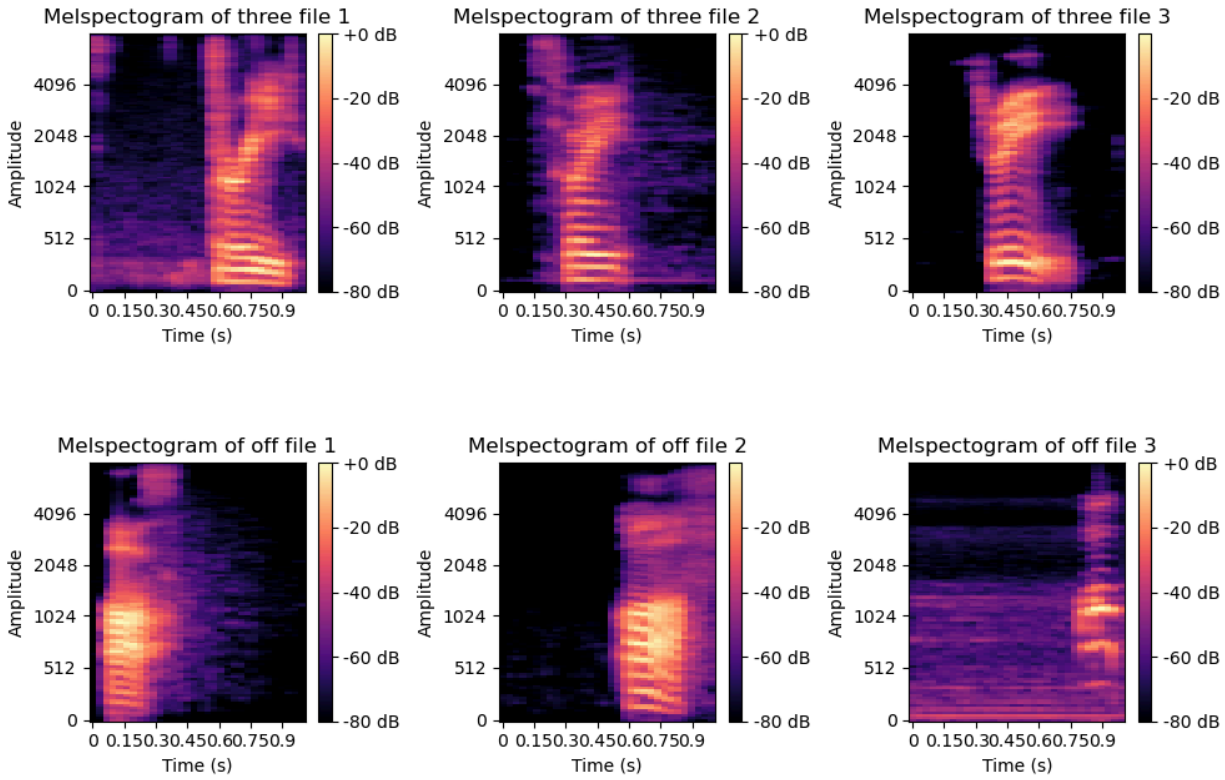


Spectogram



Melspectogram

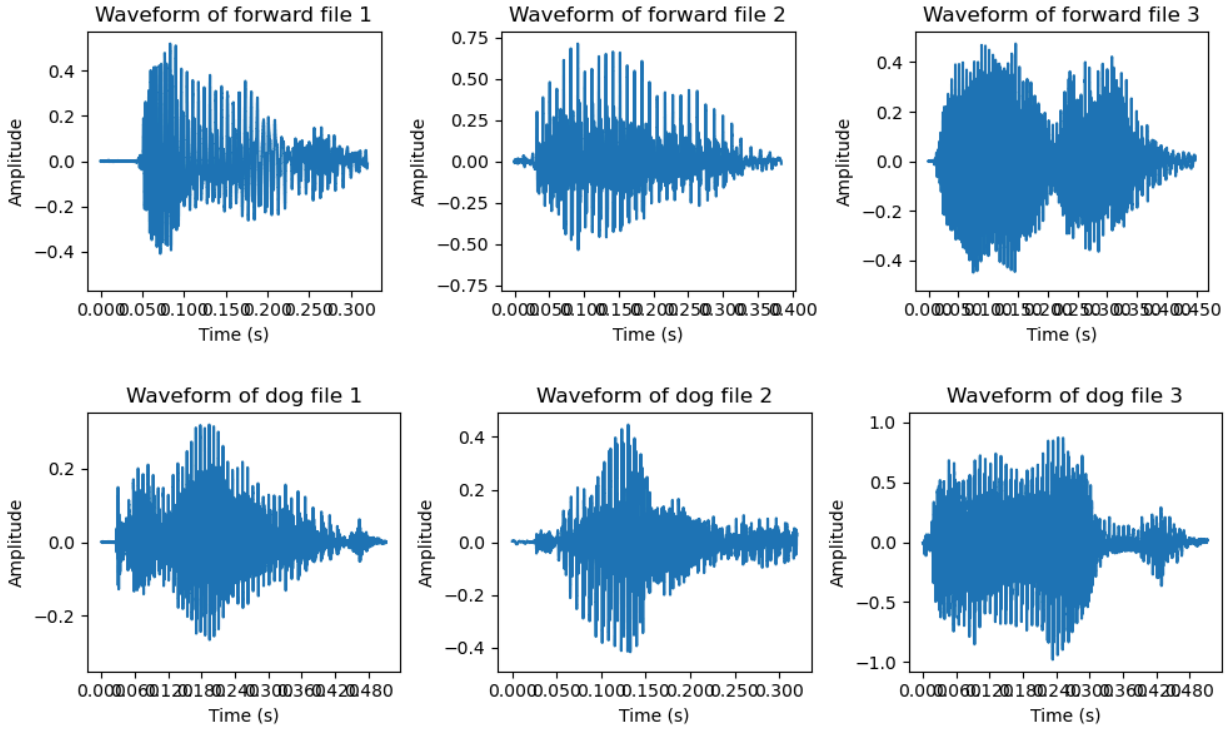


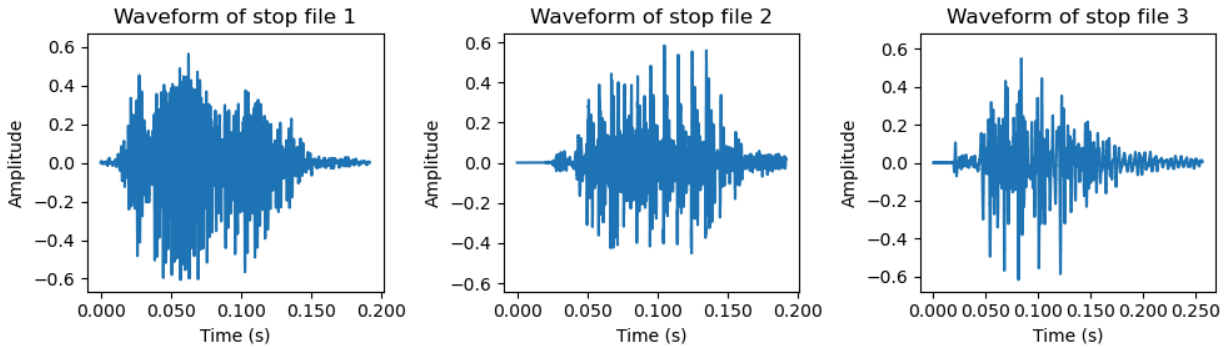


We observe silent segments from the waveforms at the beginning and end of our audio.

c. Since we notice class imbalance, we select a minimum threshold to filter out from the classes. In this case, we select 1550 as the minimum number and hence, we downsample the dataset to 1550 images per classes. We ignore the _background_noise_ for this purpose.

d. We remove silent segments from our audio and notice the waveforms again.





2. Feature Extraction

We extract MFCC, Chroma, Spectral Centroid, Spectral Bandwidth, Zero Crossing Rate, Spectral Rolloff, and RMS value.

We notice that MFCC Mean returns an array of length 13, and Chroma Mean returns an array of length 12.

This brings our dataset to have 31 columns.

We then apply label encoding to our labels.

3. We split our dataset into 80:20 ratios and apply Min Max Scaling to the training and testing dataset. We get the number of records as 43400 in the training dataset and 10850 in the testing dataset.