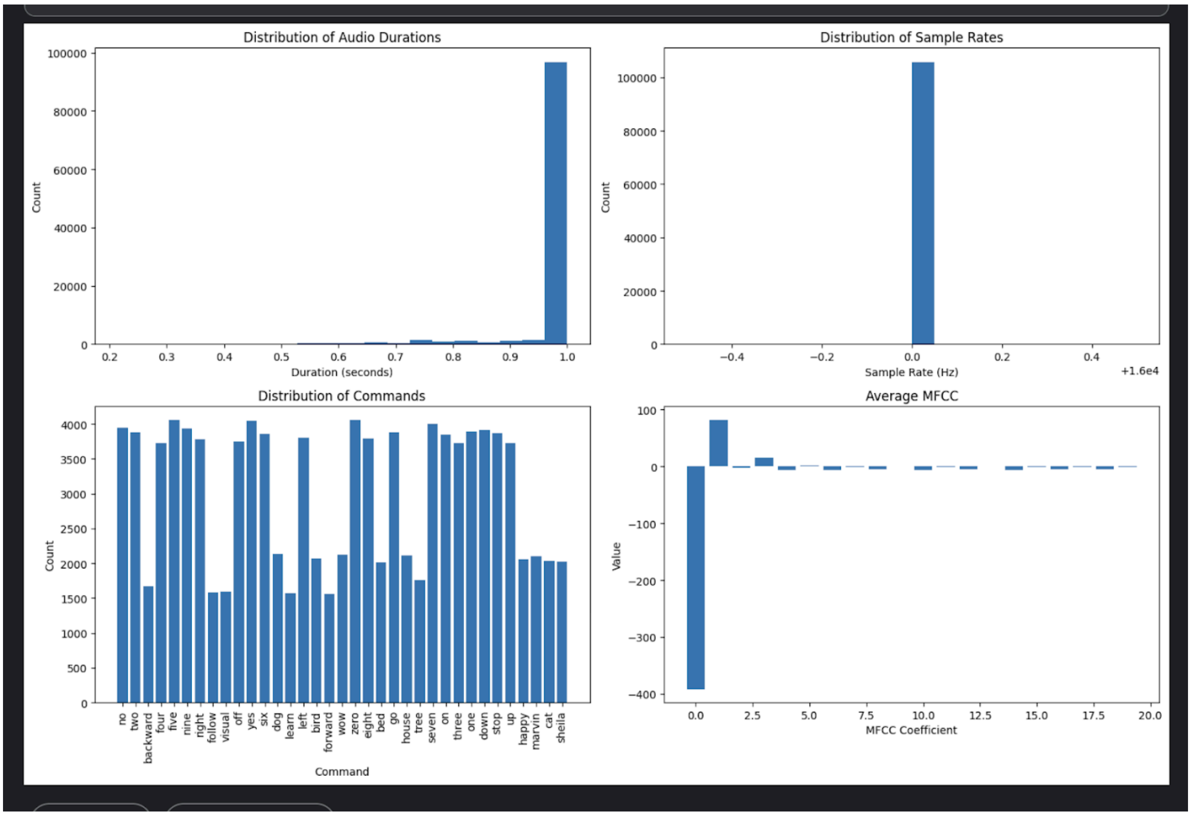
Report on Speech Command Recognition Project

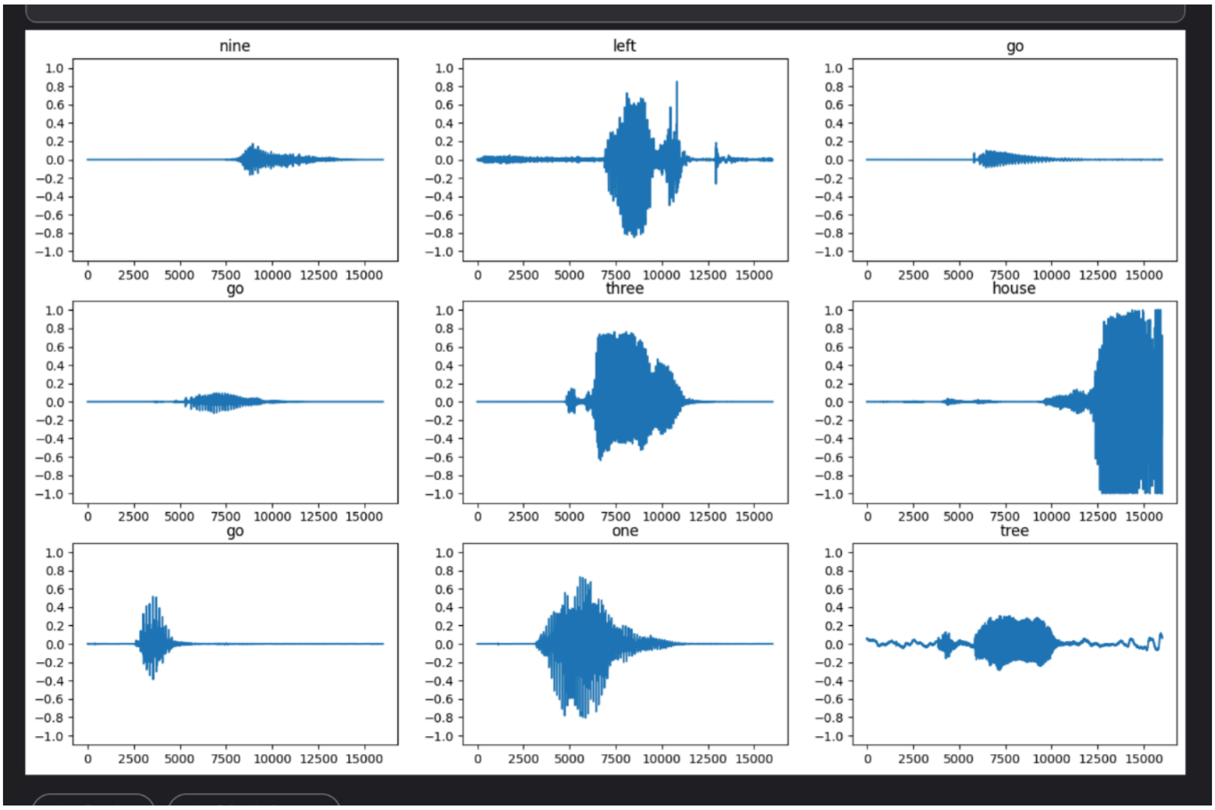
1. Logistics
- Start Time: Wed Sep 11, 10:00 AM
- End Time:Wed Sep 11, 04:30 PM

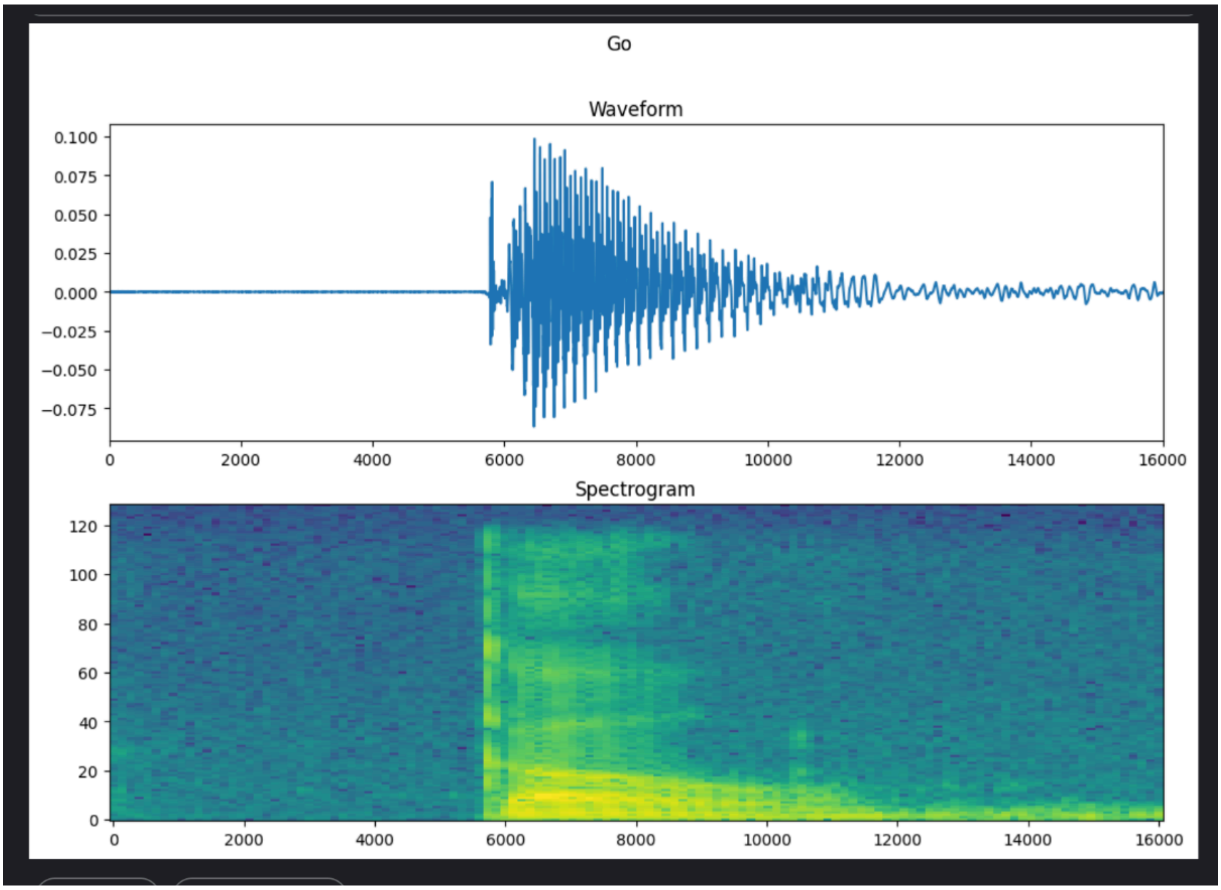
2. Task

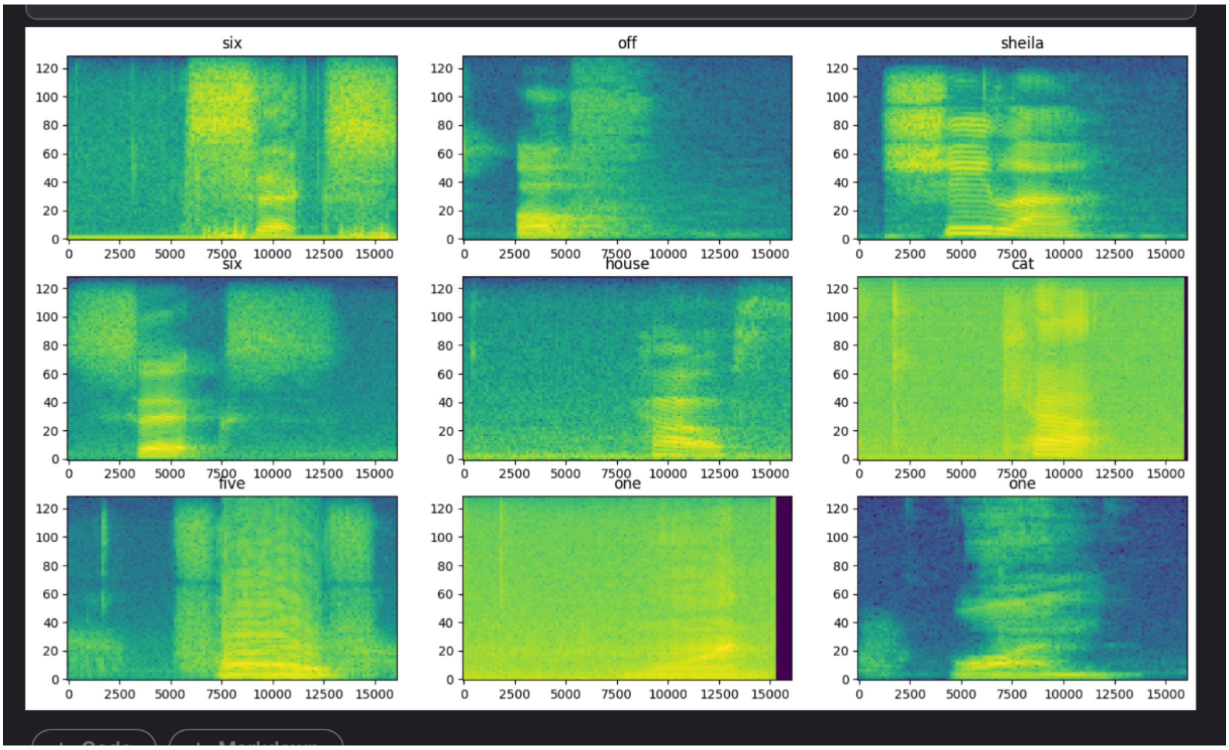
1. Paper Summary:
   The paper titled [“Deep Speech: Scaling up end-to-end speech recognition”](https://arxiv.org/abs/1804.03209) explores deep learning models for speech recognition, particularly focusing on the scalability and end-to-end nature of the models. The authors propose architectures and techniques that improve accuracy and efficiency in recognizing spoken language.

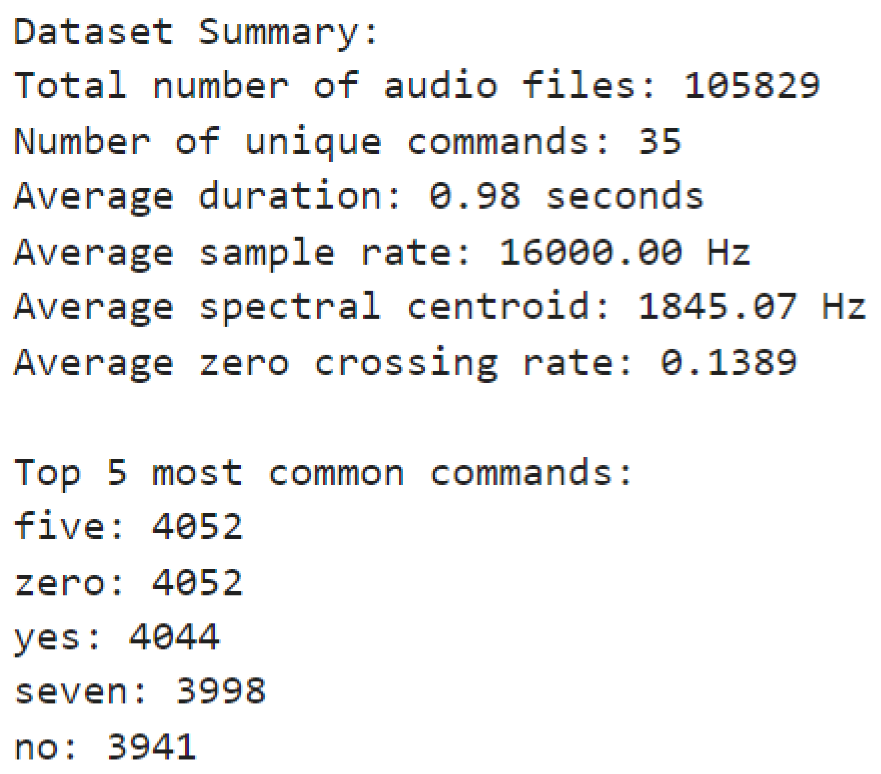
2. Dataset Analysis:
   - Dataset: The dataset from the paper was downloaded and analyzed.
   - Statistical Analysis: The dataset comprises audio recordings of various spoken commands. Statistical analysis was performed to describe the distribution of command types, sample length, and audio features. This involved code snippets that demonstrate data loading, feature extraction, and visualization.




nine
left
go
go
three
house
go
one
tree
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0
2500
5000
7500
10000
12500
15000


Go
Waveform
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0
2000
4000
6000
8000
10000
12000
14000
16000
Spectrogram
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000
14000
16000

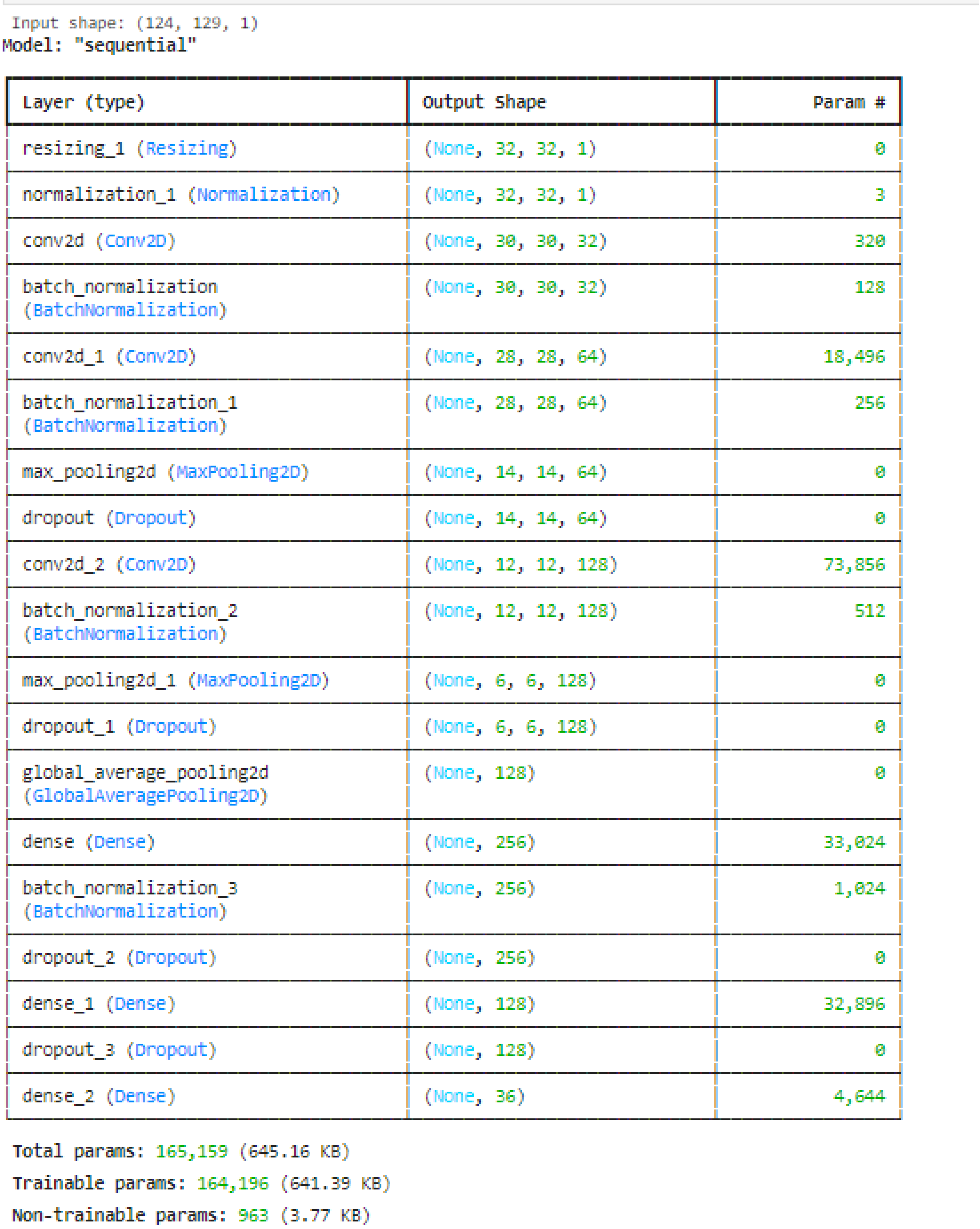




```
Dataset Summary:
Total number of audio files: 105829
Number of unique commands: 35
Average duration: 0.98 seconds
Average sample rate: 16000.00 Hz
Average spectral centroid: 1845.07 Hz
Average zero crossing rate: 0.1389

Top 5 most common commands:
five: 4052
zero: 4052
yes: 4044
seven: 3998
no: 3941
```

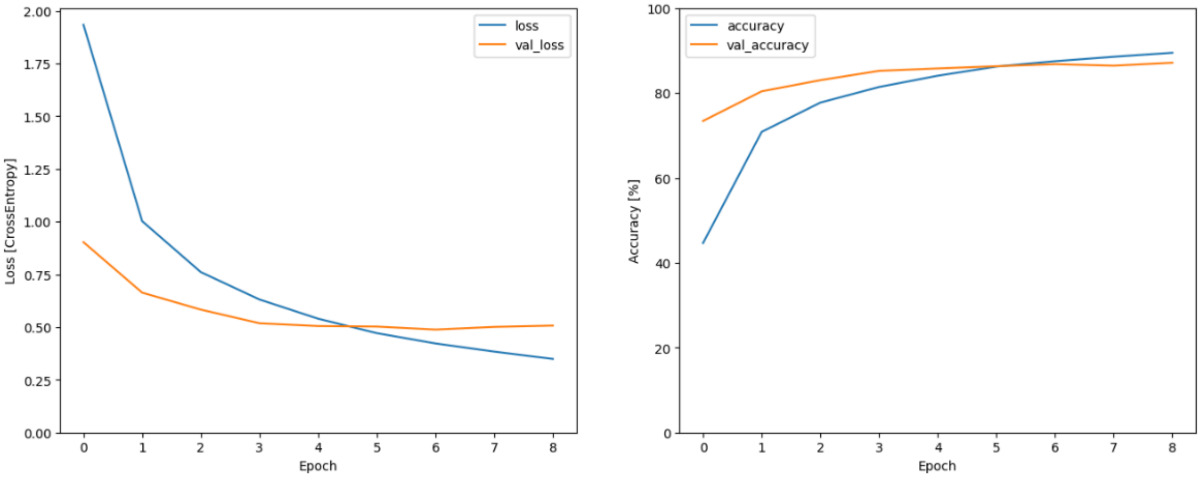
3. Classifier Training:
   - Model: A classifier was trained to recognize commands in the dataset.

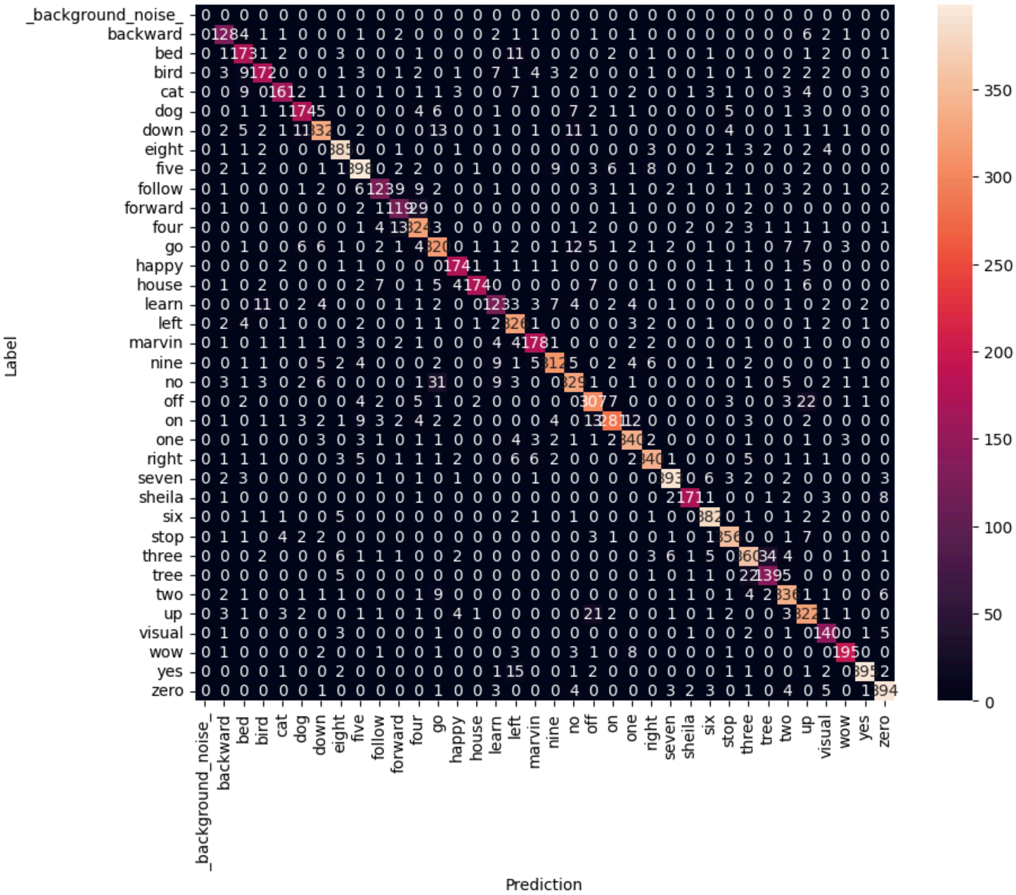


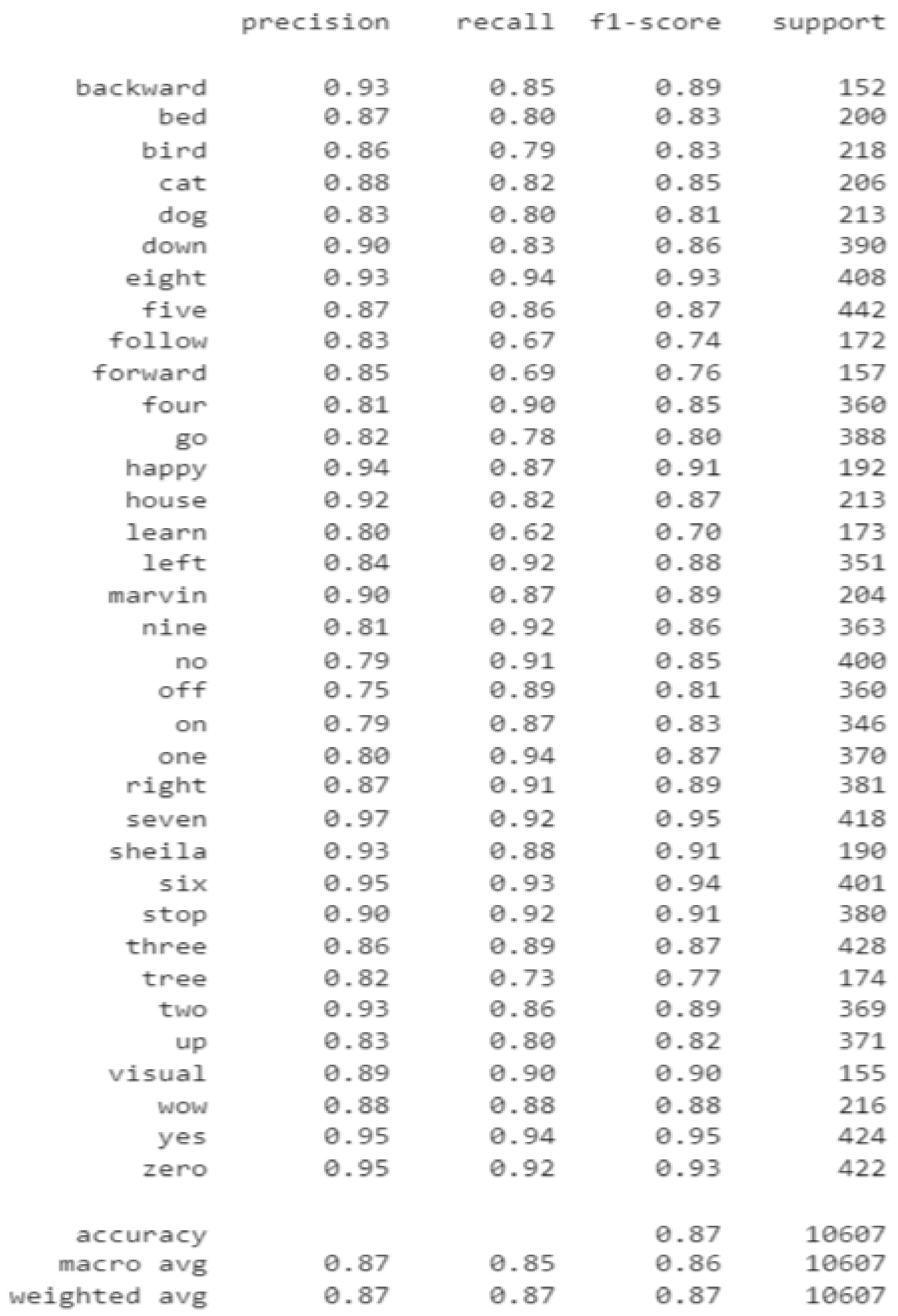
```
Input shape: (124, 129, 1)
Model: "sequential"
┌──────────────────────────────────────┬──────────────────────────────┬──────────────────┐
│ Layer (type)                         │ Output Shape                 │          Param # │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ resizing_1 (Resizing)                │ (None, 32, 32, 1)            │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ normalization_1 (Normalization)      │ (None, 32, 32, 1)            │                3 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ conv2d (Conv2D)                      │ (None, 30, 30, 32)           │              320 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ batch_normalization                  │ (None, 30, 30, 32)           │              128 │
│ (BatchNormalization)                 │                              │                  │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ conv2d_1 (Conv2D)                    │ (None, 28, 28, 64)           │           18,496 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ batch_normalization_1                │ (None, 28, 28, 64)           │              256 │
│ (BatchNormalization)                 │                              │                  │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ max_pooling2d (MaxPooling2D)         │ (None, 14, 14, 64)           │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dropout (Dropout)                    │ (None, 14, 14, 64)           │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ conv2d_2 (Conv2D)                    │ (None, 12, 12, 128)          │           73,856 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ batch_normalization_2                │ (None, 12, 12, 128)          │              512 │
│ (BatchNormalization)                 │                              │                  │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ max_pooling2d_1 (MaxPooling2D)       │ (None, 6, 6, 128)            │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dropout_1 (Dropout)                  │ (None, 6, 6, 128)            │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ global_average_pooling2d             │ (None, 128)                  │                0 │
│ (GlobalAveragePooling2D)             │                              │                  │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dense (Dense)                        │ (None, 256)                  │           33,024 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ batch_normalization_3                │ (None, 256)                  │            1,024 │
│ (BatchNormalization)                 │                              │                  │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dropout_2 (Dropout)                  │ (None, 256)                  │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dense_1 (Dense)                      │ (None, 128)                  │           32,896 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dropout_3 (Dropout)                  │ (None, 128)                  │                0 │
├──────────────────────────────────────┼──────────────────────────────┼──────────────────┤
│ dense_2 (Dense)                      │ (None, 36)                   │            4,644 │
└──────────────────────────────────────┴──────────────────────────────┴──────────────────┘
 Total params: 165,159 (645.16 KB)
 Trainable params: 164,196 (641.39 KB)
 Non-trainable params: 963 (3.77 KB)
```

4. Performance Report:

```
Text(0, 0.5, 'Accuracy [%]')
```



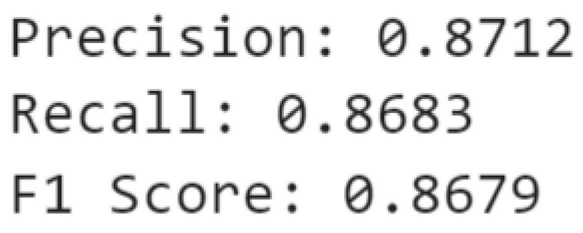




```
              precision    recall  f1-score   support

    backward       0.93      0.85      0.89       152
         bed       0.87      0.80      0.83       200
        bird       0.86      0.79      0.83       218
         cat       0.88      0.82      0.85       206
         dog       0.83      0.80      0.81       213
        down       0.90      0.83      0.86       390
       eight       0.93      0.94      0.93       408
        five       0.87      0.86      0.87       442
      follow       0.83      0.67      0.74       172
     forward       0.85      0.69      0.76       157
        four       0.81      0.90      0.85       360
          go       0.82      0.78      0.80       388
       happy       0.94      0.87      0.91       192
       house       0.92      0.82      0.87       213
       learn       0.80      0.62      0.70       173
        left       0.84      0.92      0.88       351
      marvin       0.90      0.87      0.89       204
        nine       0.81      0.92      0.86       363
          no       0.79      0.91      0.85       400
         off       0.75      0.89      0.81       360
          on       0.79      0.87      0.83       346
         one       0.80      0.94      0.87       370
       right       0.87      0.91      0.89       381
       seven       0.97      0.92      0.95       418
      sheila       0.93      0.88      0.91       190
         six       0.95      0.93      0.94       401
        stop       0.90      0.92      0.91       380
       three       0.86      0.89      0.87       428
        tree       0.82      0.73      0.77       174
         two       0.93      0.86      0.89       369
          up       0.83      0.80      0.82       371
      visual       0.89      0.90      0.90       155
         wow       0.88      0.88      0.88       216
         yes       0.95      0.94      0.95       424
        zero       0.95      0.92      0.93       422

    accuracy                           0.87     10607
   macro avg       0.87      0.85      0.86     10607
weighted avg       0.87      0.87      0.87     10607
```



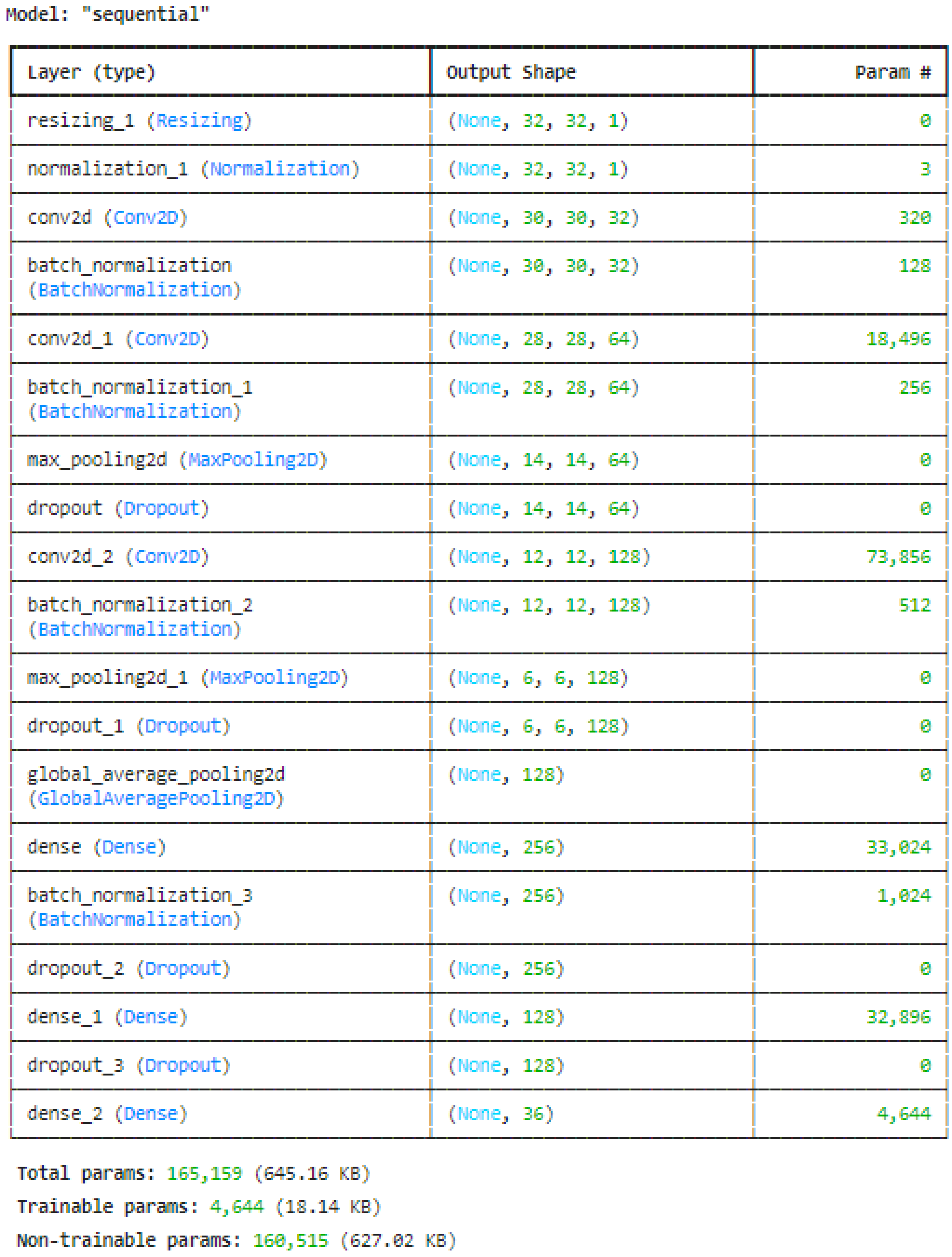
```
Precision: 0.8712
Recall: 0.8683
F1 Score: 0.8679
```

5. New Dataset Creation:
   - Data Collection: Recorded about 10 samples of each command.

Used data augmentation to enhance the dataset into 30 samples of each command word by time stretching, time shifting, noise addition.

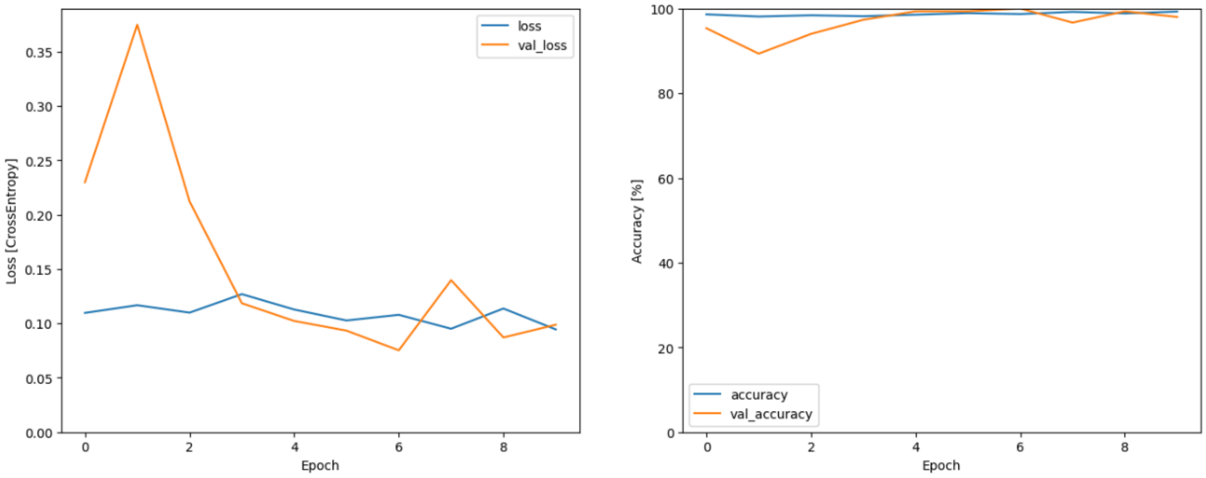
6. Fine tuning model:

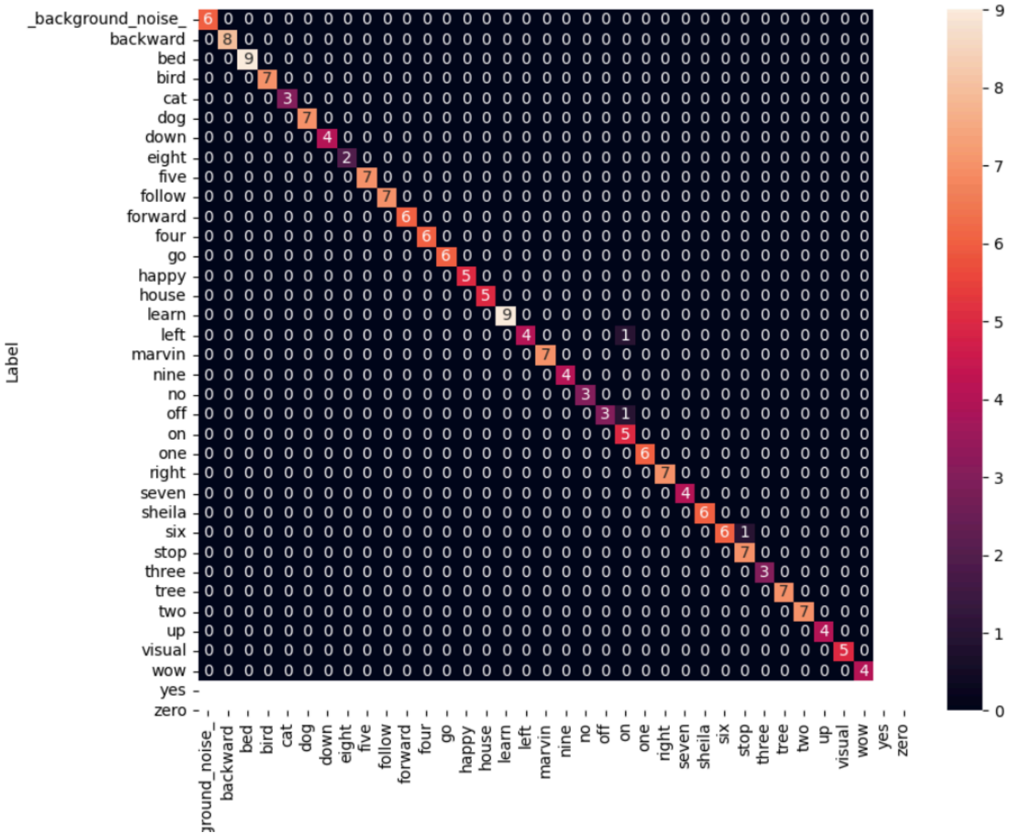
- The model was loaded again, first few layers were freezed and then model was fine tuned again on custom dataset



```
Model: "sequential"
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━┓
┃ Layer (type)                         ┃ Output Shape                ┃         Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━┩
│ resizing_1 (Resizing)                │ (None, 32, 32, 1)           │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ normalization_1 (Normalization)      │ (None, 32, 32, 1)           │               3 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ conv2d (Conv2D)                      │ (None, 30, 30, 32)          │             320 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ batch_normalization                  │ (None, 30, 30, 32)          │             128 │
│ (BatchNormalization)                 │                             │                 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ conv2d_1 (Conv2D)                    │ (None, 28, 28, 64)          │          18,496 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ batch_normalization_1                │ (None, 28, 28, 64)          │             256 │
│ (BatchNormalization)                 │                             │                 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ max_pooling2d (MaxPooling2D)         │ (None, 14, 14, 64)          │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dropout (Dropout)                    │ (None, 14, 14, 64)          │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ conv2d_2 (Conv2D)                    │ (None, 12, 12, 128)         │          73,856 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ batch_normalization_2                │ (None, 12, 12, 128)         │             512 │
│ (BatchNormalization)                 │                             │                 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ max_pooling2d_1 (MaxPooling2D)       │ (None, 6, 6, 128)           │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dropout_1 (Dropout)                  │ (None, 6, 6, 128)           │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ global_average_pooling2d             │ (None, 128)                 │               0 │
│ (GlobalAveragePooling2D)             │                             │                 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dense (Dense)                        │ (None, 256)                 │          33,024 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ batch_normalization_3                │ (None, 256)                 │           1,024 │
│ (BatchNormalization)                 │                             │                 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dropout_2 (Dropout)                  │ (None, 256)                 │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dense_1 (Dense)                      │ (None, 128)                 │          32,896 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dropout_3 (Dropout)                  │ (None, 128)                 │               0 │
├──────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dense_2 (Dense)                      │ (None, 36)                  │           4,644 │
└──────────────────────────────────────┴─────────────────────────────┴─────────────────┘
 Total params: 165,159 (645.16 KB)
 Trainable params: 4,644 (18.14 KB)
 Non-trainable params: 160,515 (627.02 KB)
```

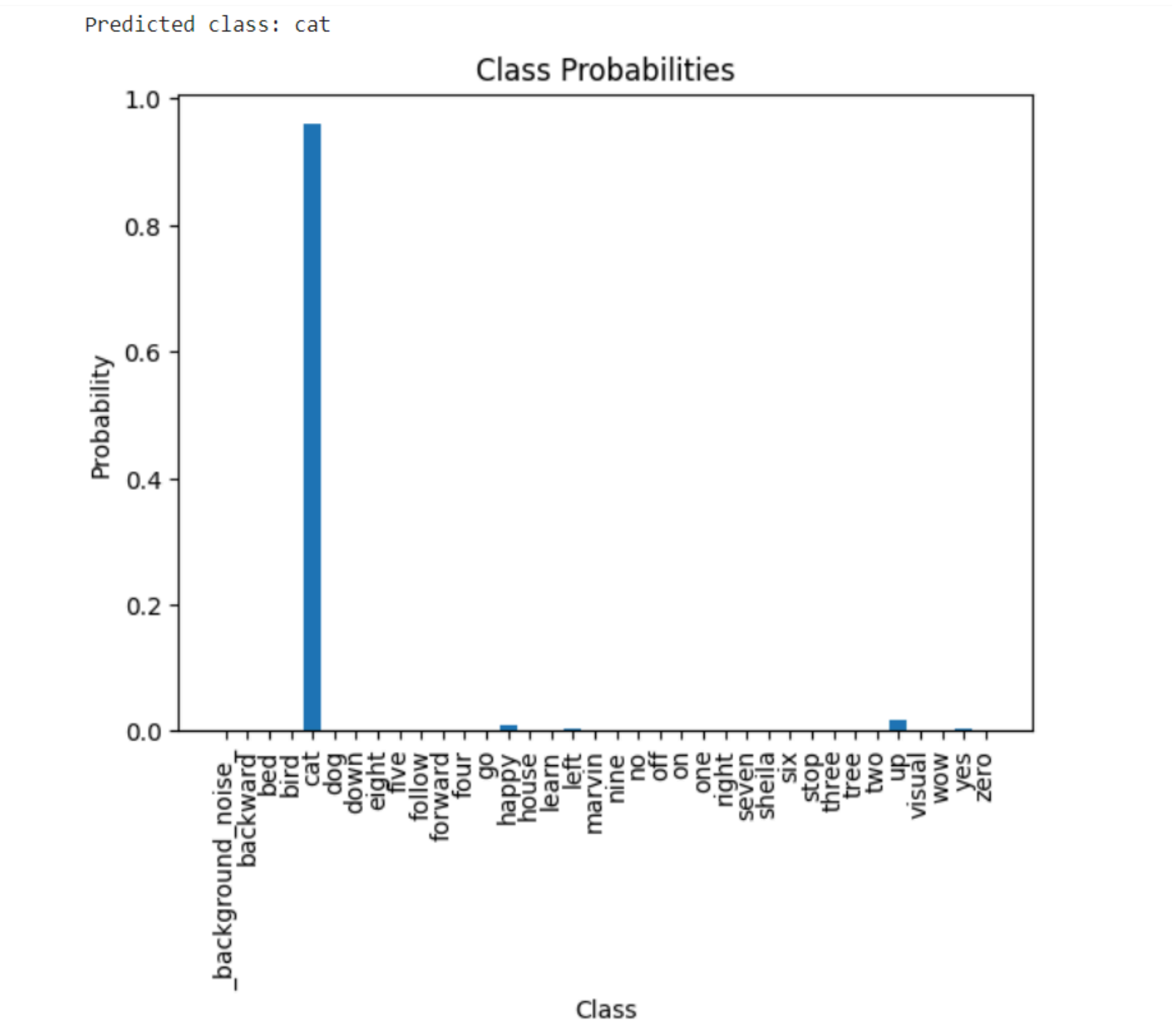
```
Text(0, 0.5, 'Accuracy [%]')
```





	ground_noise_	backward	bed	bird	cat	dog	down	eight	five	follow	forward	four	go	happy	house	learn	left	marvin	nine	no	off	on	one	right	seven	sheila	six	stop	three	tree	two	up	visual	wow	yes	zero
_background_noise_	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
backward	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
bed	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
bird	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
cat	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
dog	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
down	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
eight	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
five	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
follow	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
forward	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
four	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
go	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
happy	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
house	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
learn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
left	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0		
marvin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
nine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
no	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
off	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0		
on	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0		
one	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0		
right	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0		
seven	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0		
sheila	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0		
six	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	0	0	0	0	0	0		
stop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0		
three	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0		
tree	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0		
two	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0		
up	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0		
visual	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0		
wow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4		
yes																																				
zero																																				

```
Precision: 0.9880
Recall: 0.9844
F1 Score: 0.9846
```



```
Predicted class: cat
```