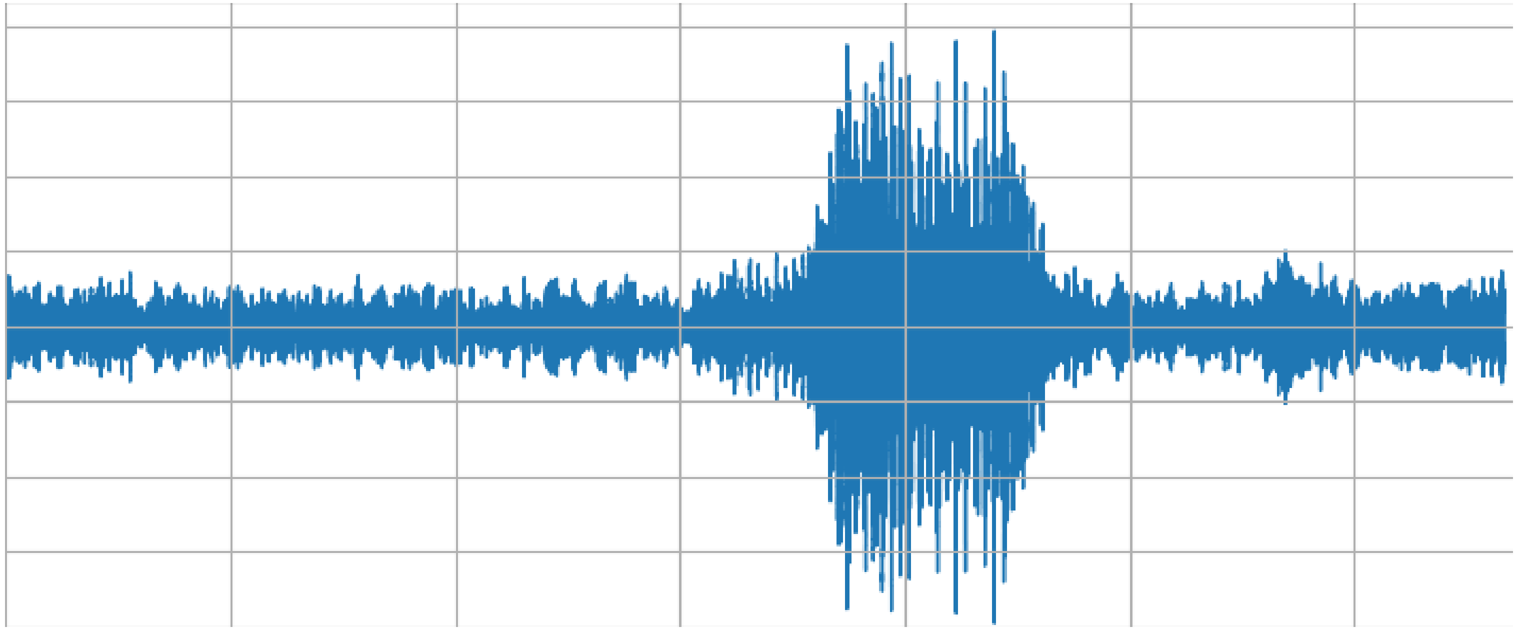
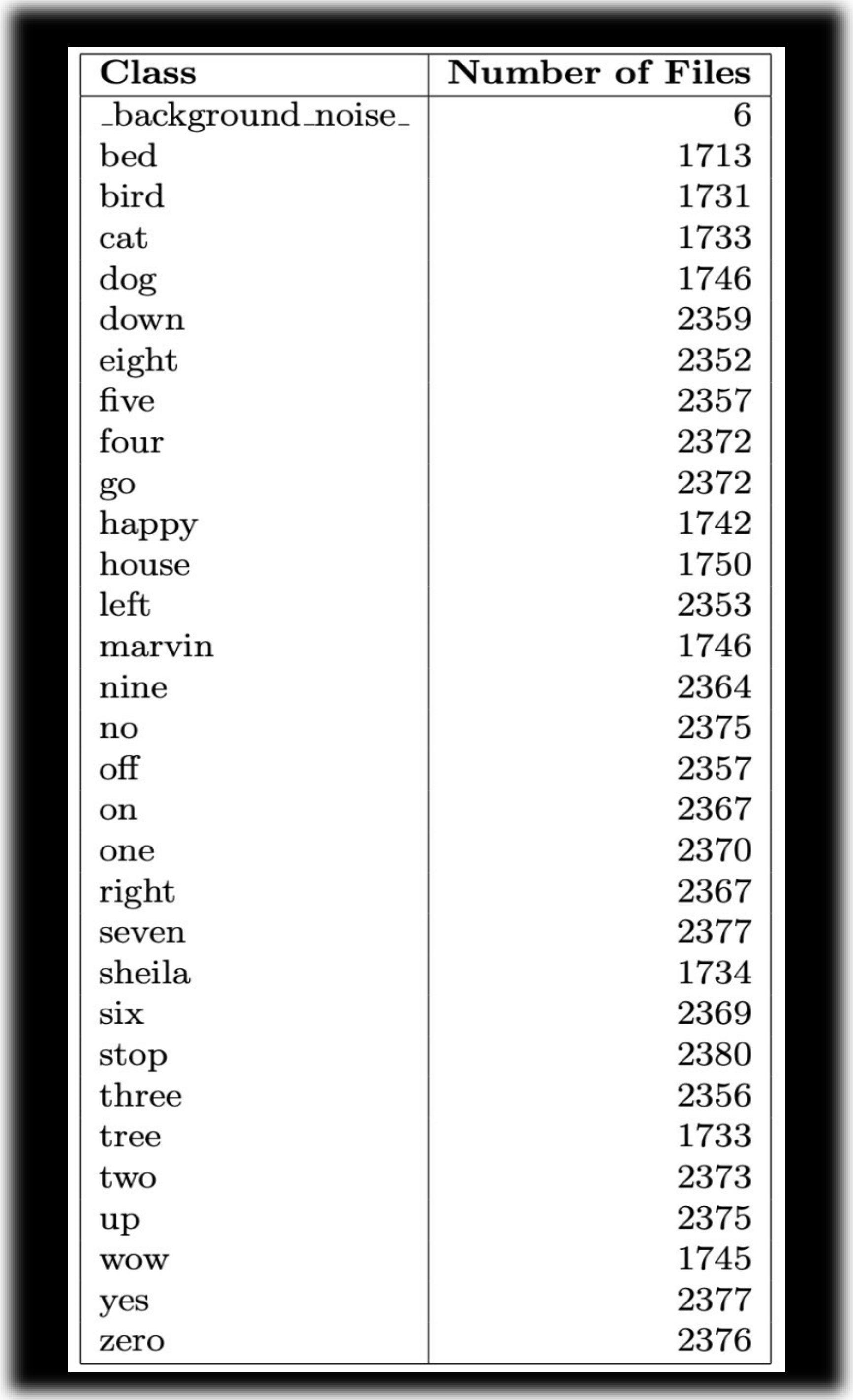


TENSORFLOW SPEECH RECOGNITION CHALLENGE

Paulina Kulczyk,

Jan Pogłód

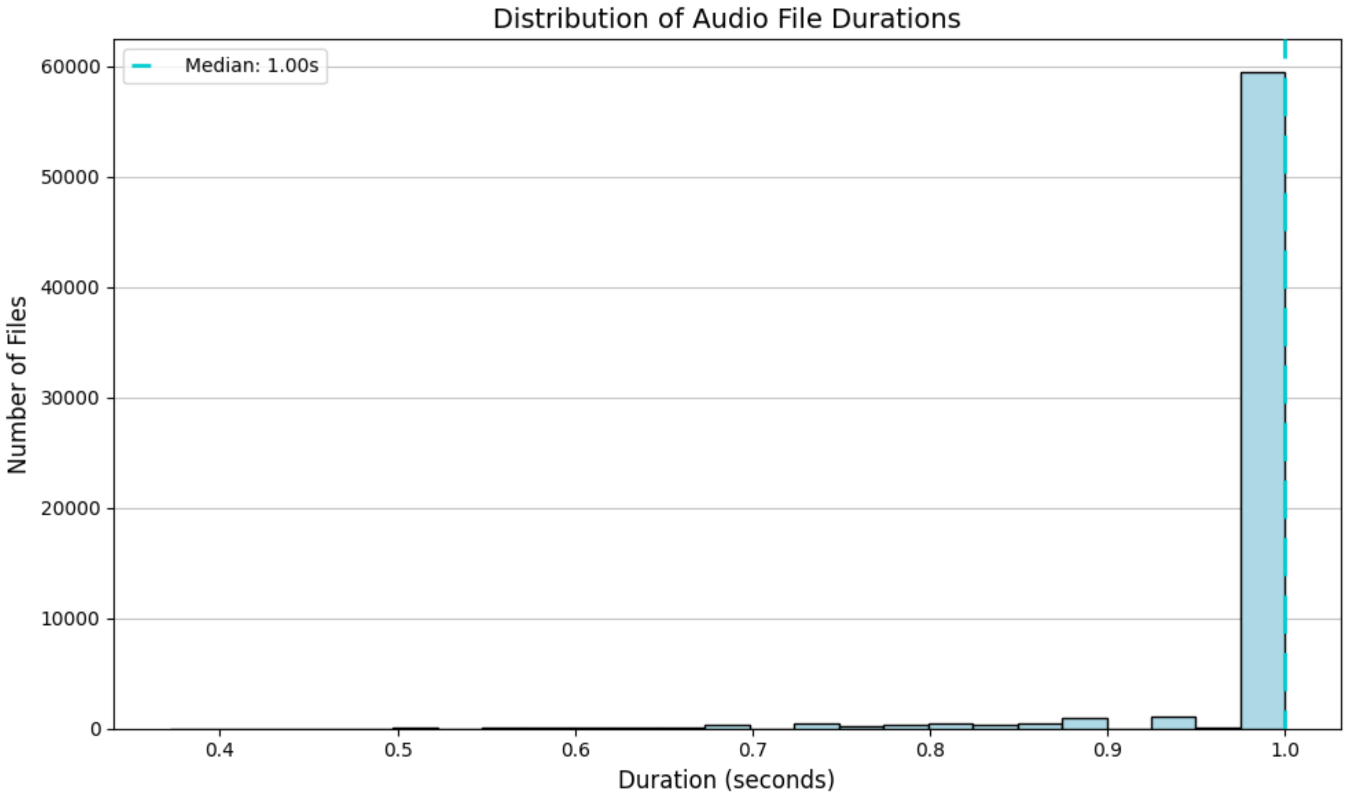


Class	Number of Files
_background_noise_	6
bed	1713
bird	1731
cat	1733
dog	1746
down	2359
eight	2352
five	2357
four	2372
go	2372
happy	1742
house	1750
left	2353
marvin	1746
nine	2364
no	2375
off	2357
on	2367
one	2370
right	2367
seven	2377
sheila	1734
six	2369
stop	2380
three	2356
tree	1733
two	2373
up	2375
wow	1745
yes	2377
zero	2376

DATA OVERVIEW

- **Dataset**: TensorFlow Speech Recognition Challenge form Kaggle
- **Twenty core command words** were recorded, with most speakers saying each of them five times: "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", "Zero", "One", "Two", "Three", "Four", "Five", "Six", "Seven", "Eight", and "Nine". **There are also ten auxiliary words**, which most speakers only said once: "Bed", "Bird", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila", "Tree", and "Wow". And 6 files with noise (silence).
- **Task**: classify words to one of the following classes: **"Down", "Go", "Left", "No", "Off", "On", "Right", "Stop", "Up", "Yes", "Unknown**" (with other words from Dataset) and Silence (where none word was recognized)

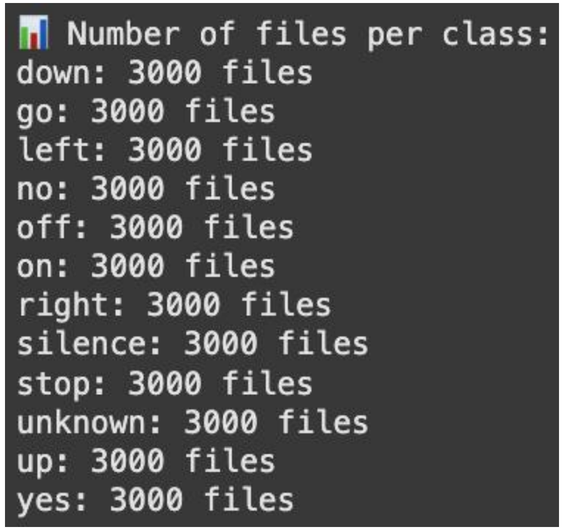
DURATION OF FILES



- **Number of files:** 64721
- **Min:** 0.37s**, Max:** 1.00s
- **Mean duration:** 0.98 sekundy
- **Sampling rate**: 16 000 Hz

PREPROCESSING

- Deletion of 308 files that last shorter then 0.6s
- Cutting 6 silence files to many shorter files which lasts about 1s (as our files from dataset)
- Mapping data to unknown and silence (randomly choosing 3000 files from uknown class with given seed)
- Bootstraping files per class and sampling from unknown and silence to have well balanced classes (3000 audio tracks per class)



```
📊 Number of files per class:
down: 3000 files
go: 3000 files
left: 3000 files
no: 3000 files
off: 3000 files
on: 3000 files
right: 3000 files
silence: 3000 files
stop: 3000 files
unknown: 3000 files
up: 3000 files
yes: 3000 files
```

SCIENTIFIC BACKGROUND

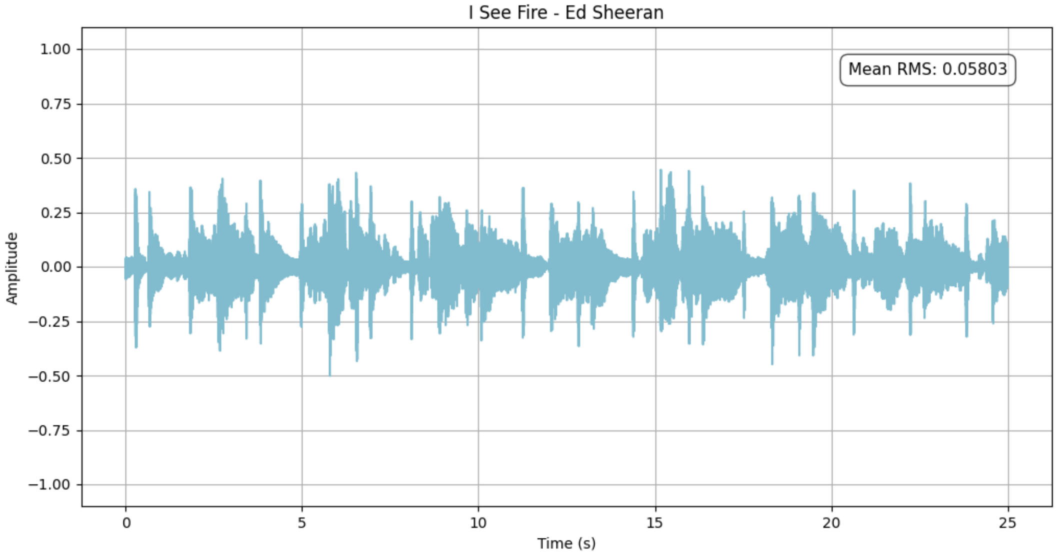
- Voice can be displayed as an **acoustic wave**
- Main measurements of waves that has impact to sound are: **amplitude and frequency**

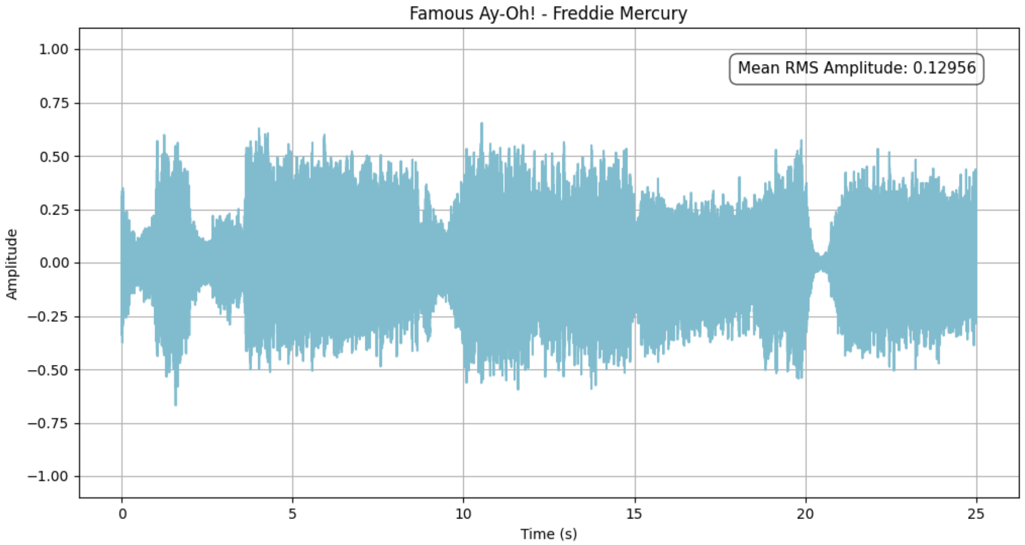
SCIENTIFIC BACKGROUND - MEASUREMENTS

- **Mean RMS Amplitude** – is the square root of the average of the squares of a series of measurements of values of sign of wave (almost amplitude).
- **Zero-crossing rate (ZCR)** - normally it measures how often the signal changes sign — meaning it crosses the horizontal axis (from positive to negative or negative to positive) — within a frame. But in our study to better capture fast waveform fluctuations, we modified the ZCR calculation. Instead of counting strict signal sign changes, we measure how often the waveform touches the zero axis, averaged over 0.25-second frames (as most files last 1s).
  This approach approximates the traditional ZCR while improving sensitivity to rapid oscillations.

SCIENTIFIC BACKGROUND – REFLECTED IN REAL LIFE EXAMPLES

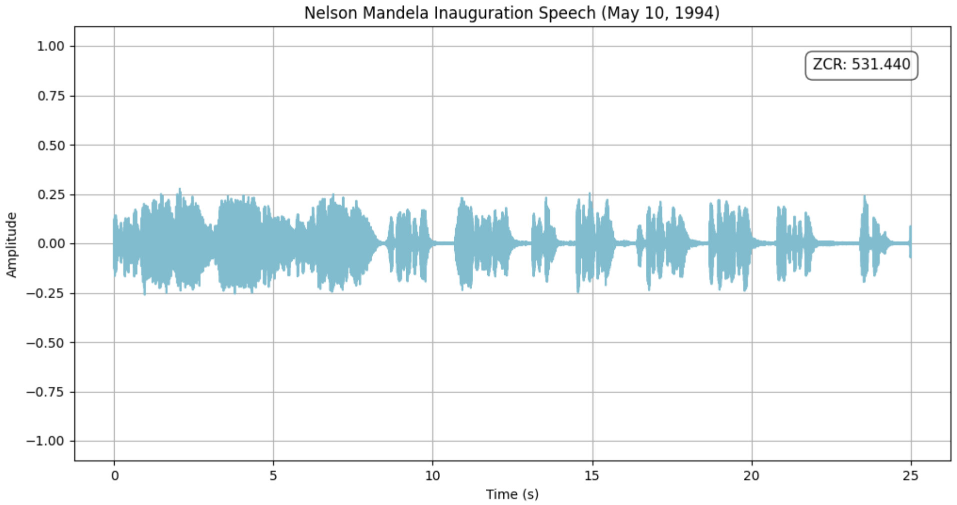
- The **loudness of a sound** is directly proportional to the quadrature of the **amplitude of the sound wave**

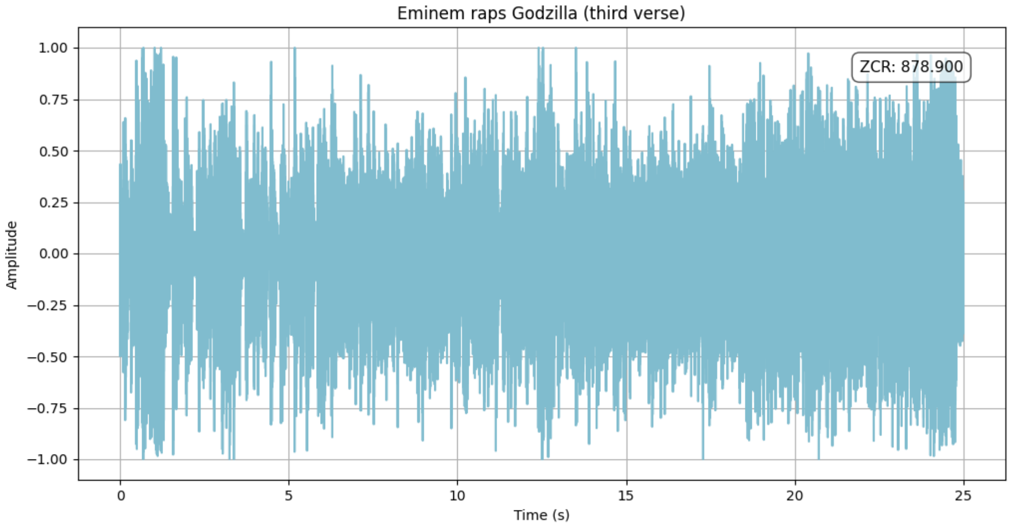




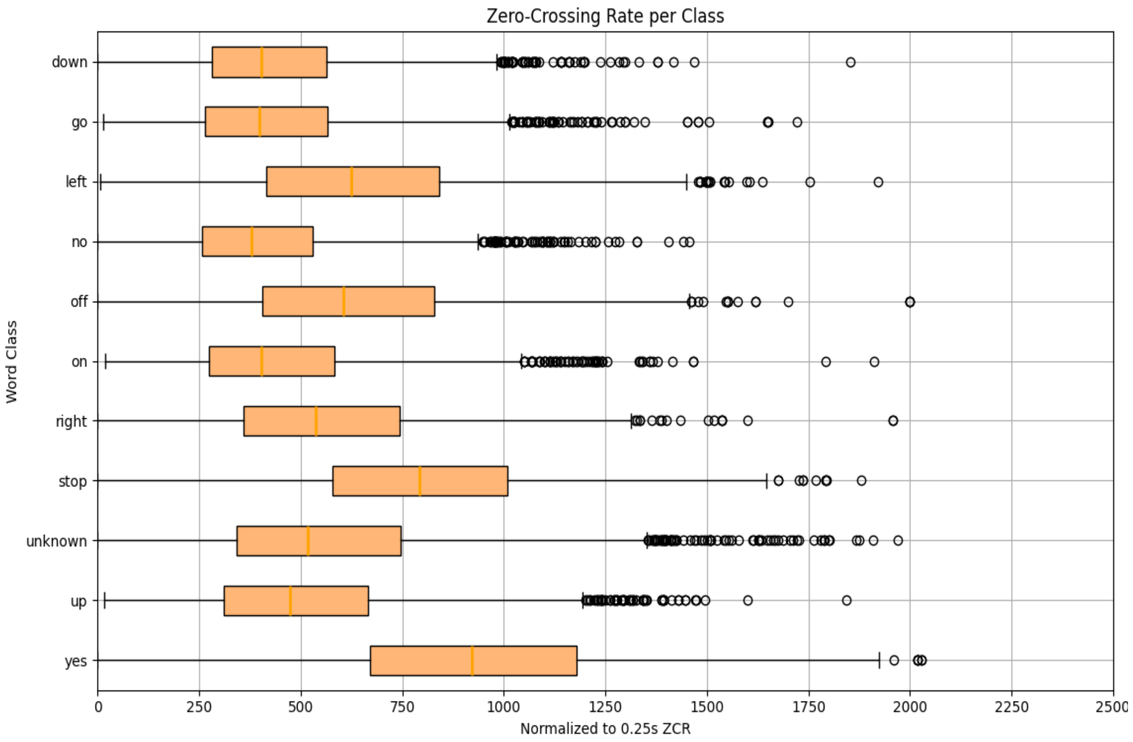
SCIENTIFIC BACKGROUND - REFLECTED IN REAL LIFE EXAMPLES

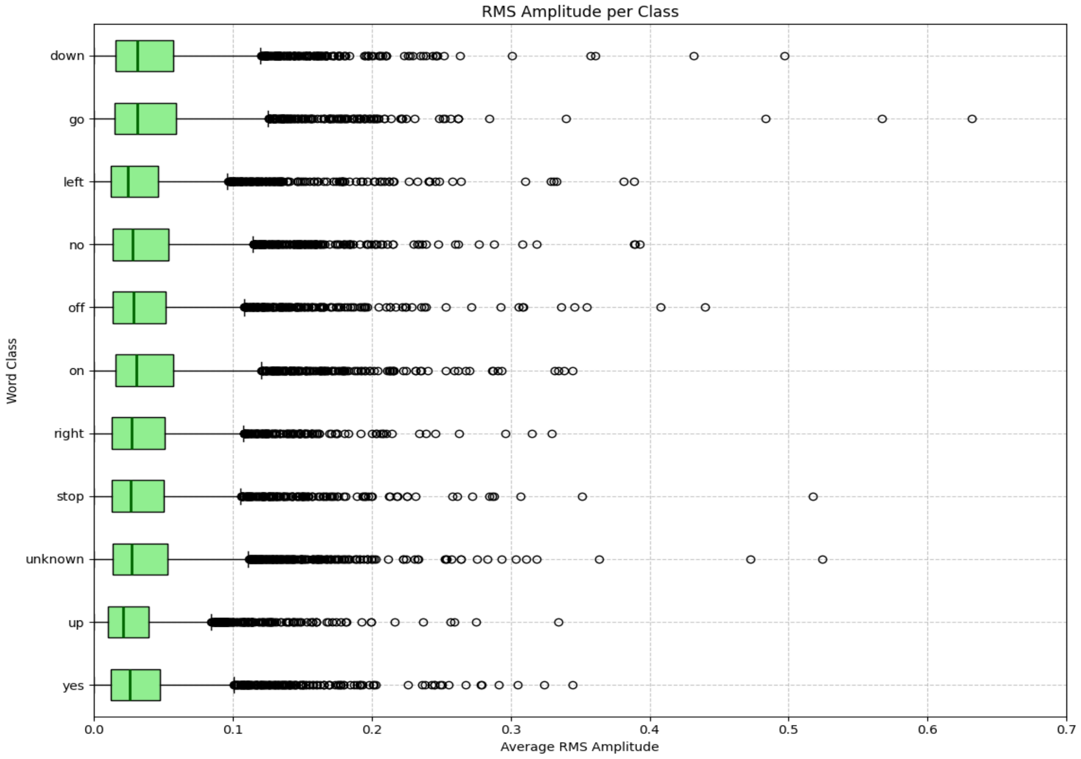
- Indirectly, **how fast we speak** or **how we stretch syllables has an effect on the frequency of the acoustic wave**





SCIENTIFIC BACKGROUND – OUR DATA


Zero-Crossing Rate per Class


RMS Amplitude per Class

1ST EXPERIMENT

BINARY CLASSIFICATION PROBLEM FOR RECOGNITION "YES" AND "NO"

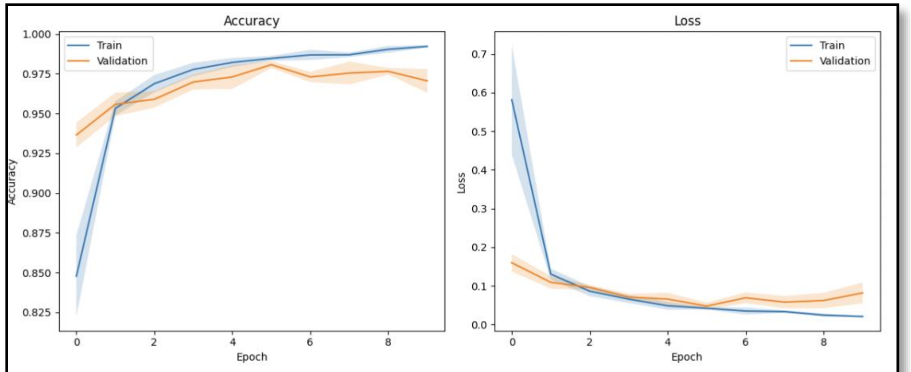


Figure 7: Learning process of the basic CNN architecture for binary problem

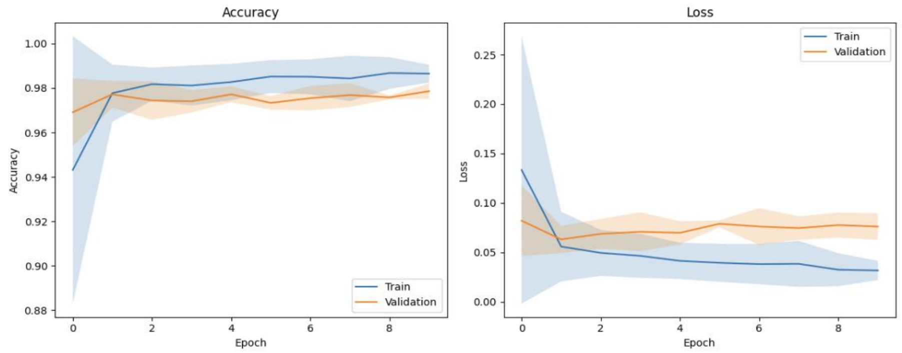


Figure 9: Learning process of the simple transformer architecture for binary problem

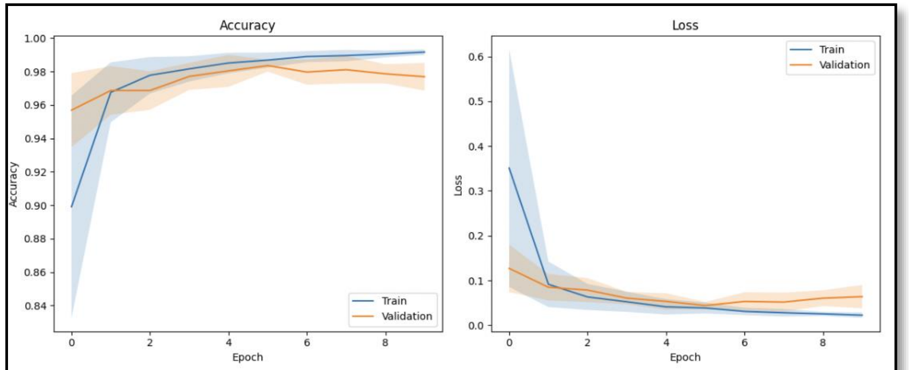


Figure 8: Learning process of the LSTM architecture for binary problem

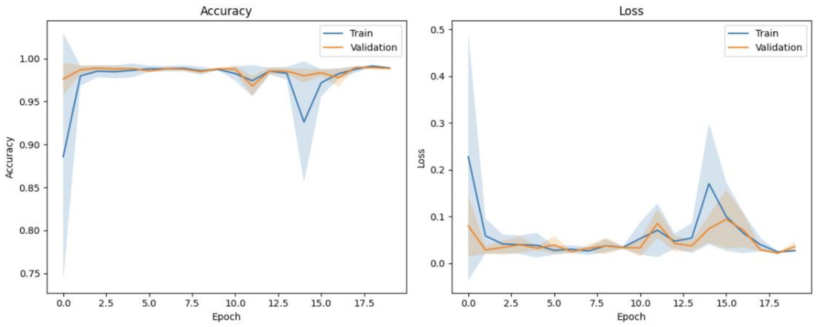
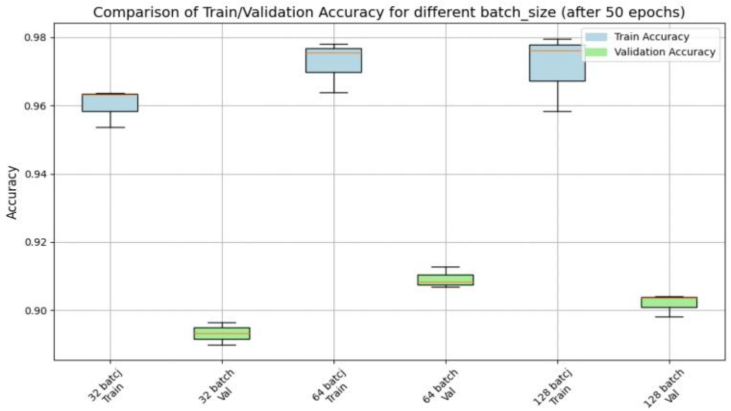


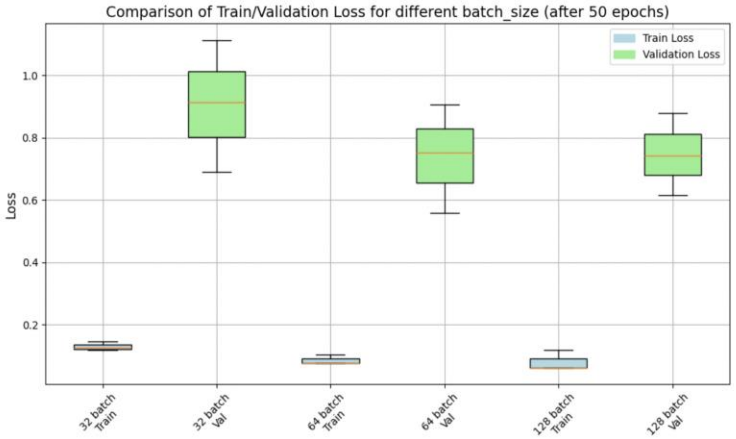
Figure 10: Learning process of the advanced architecture (CNN + transformer) for binary problem

Architecture	Valid accuracy (mean)
CNN	97.5%
LSTM	98,1%
Transformer	97.9%
CNN + Transformer	98.7%

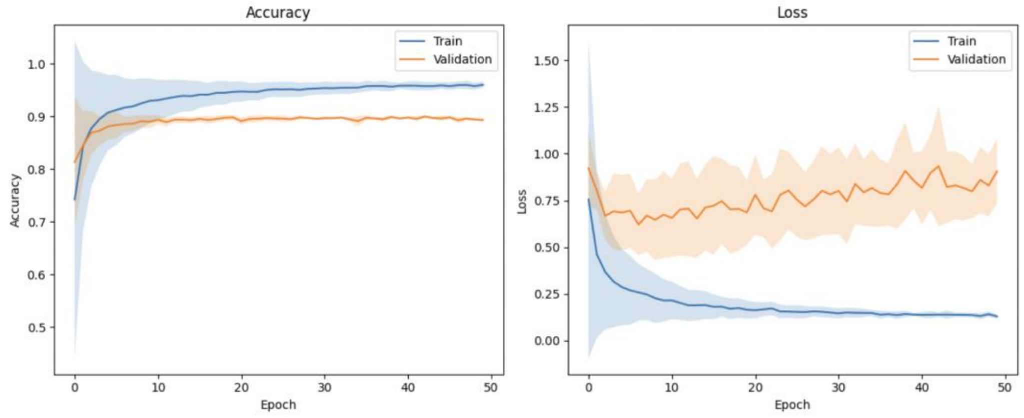
2ND EXPERIMENT

CNN COMPARISON OF PARAMETERS – BATCH SIZE

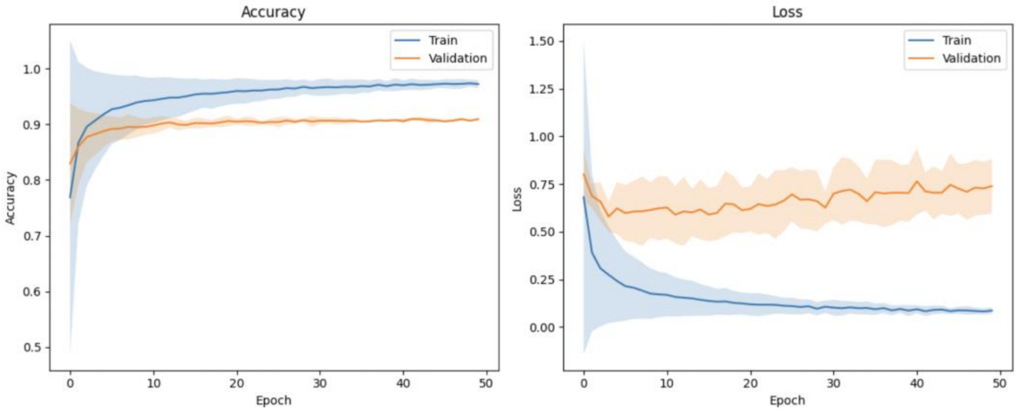




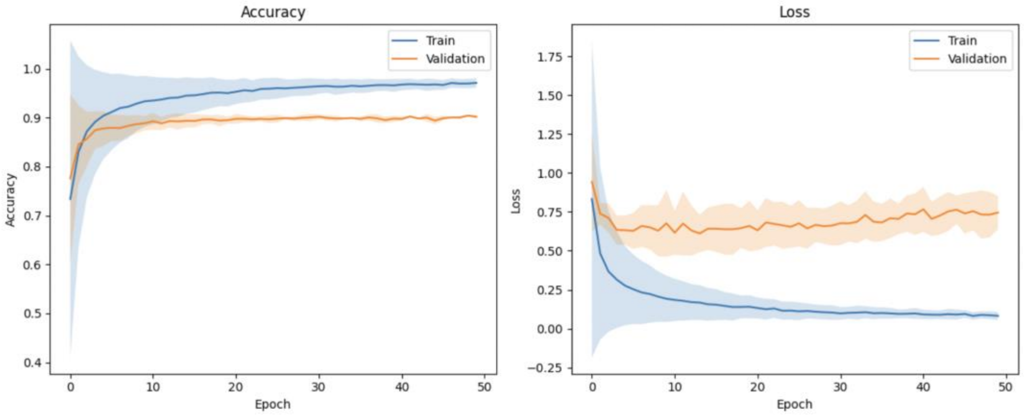
32



64

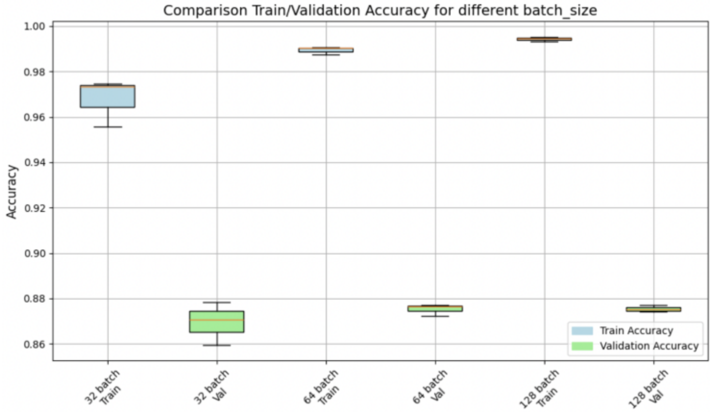


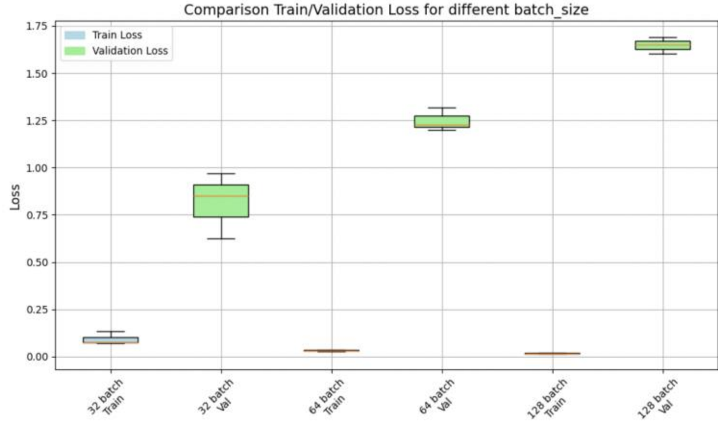
128



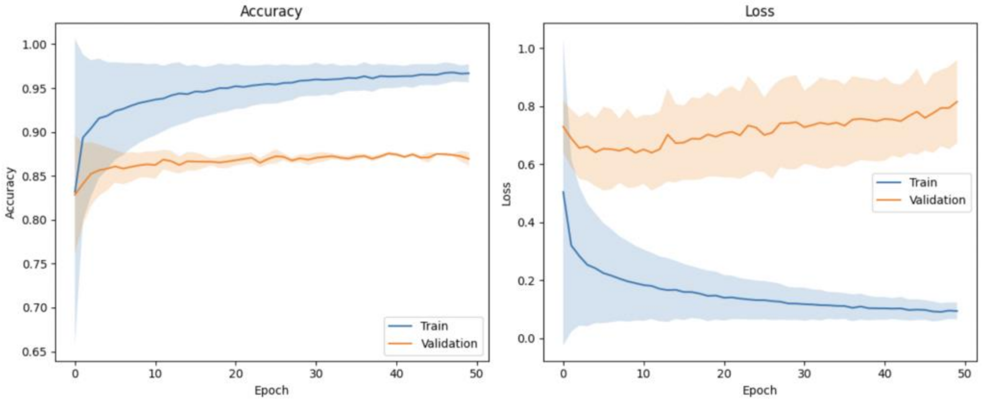
2ND EXPERIMENT

TRANSFORMER COMPARISON OF PARAMETERS – BATCH SIZE

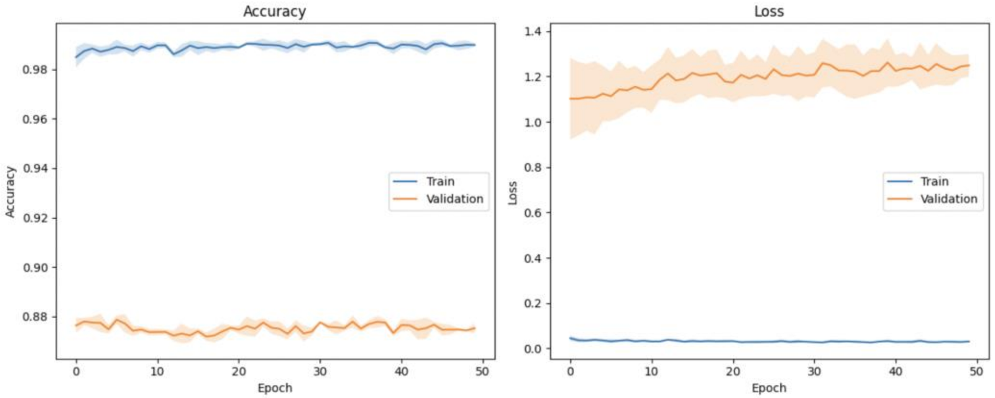




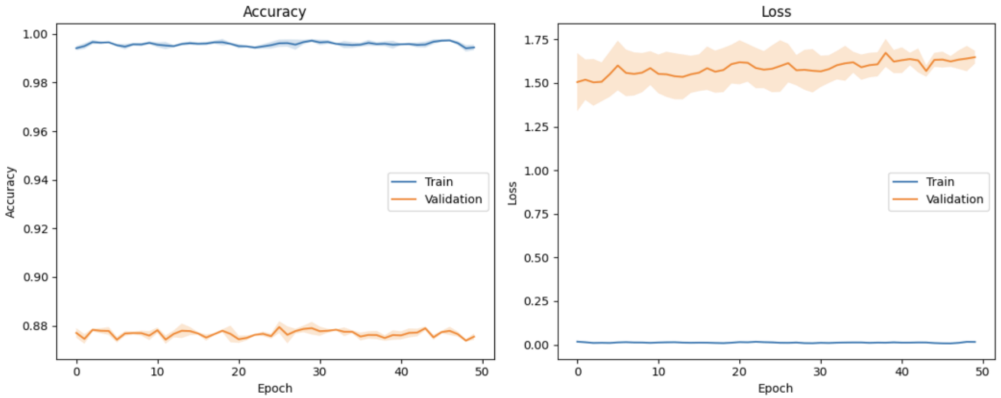
32



64

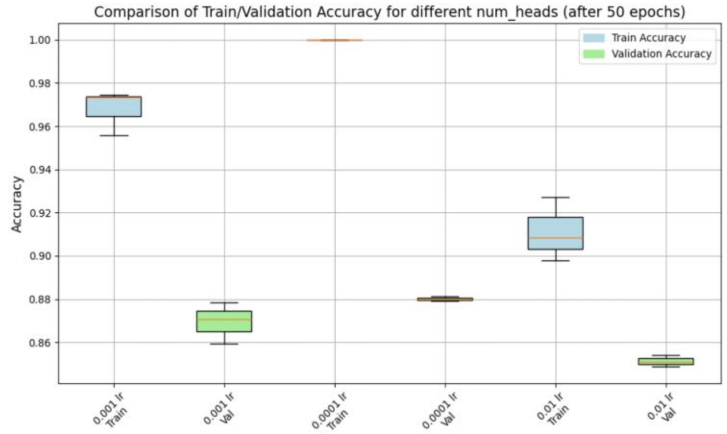


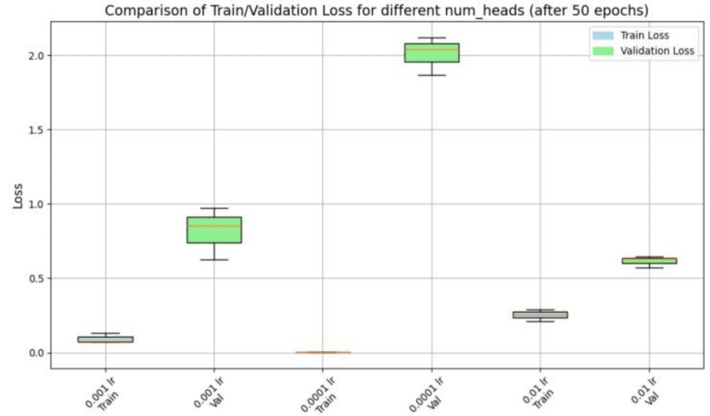
128



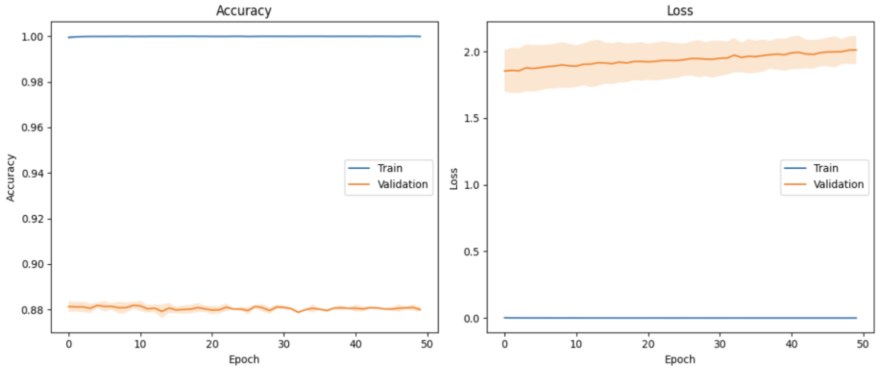
2ND EXPERIMENT

TRANSFORMER COMPARISON OF PARAMETERS – LEARNING RATE

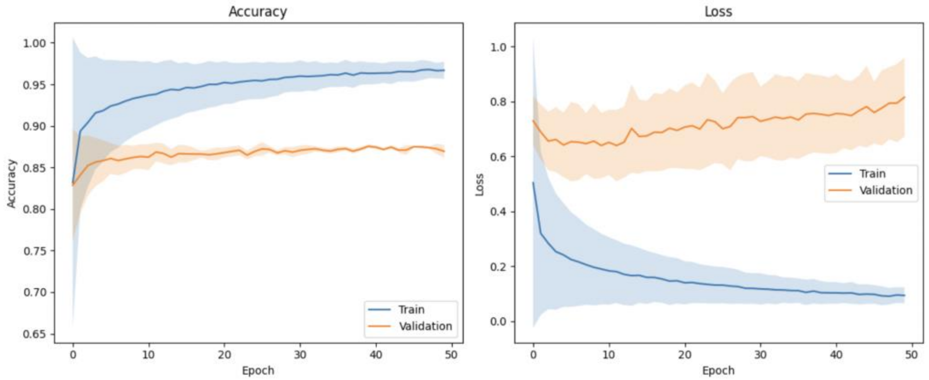




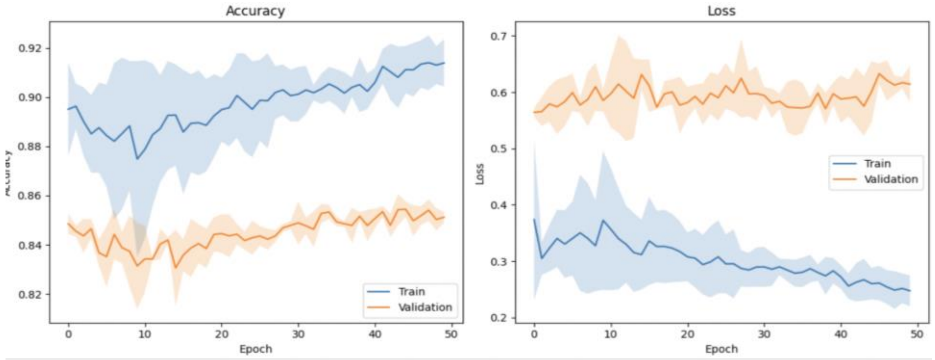
0.0001



0.001

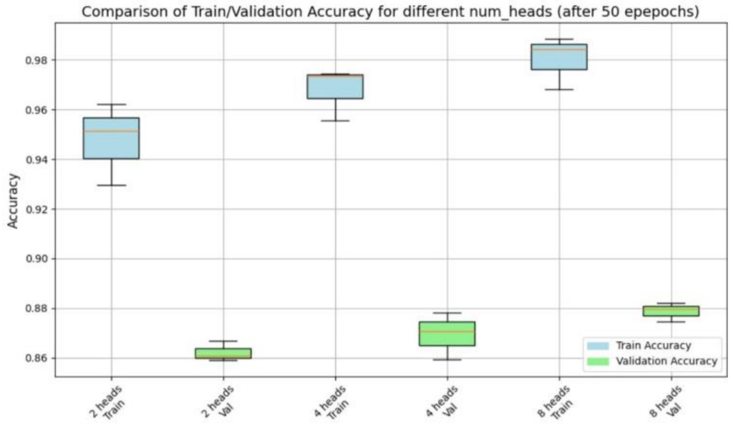


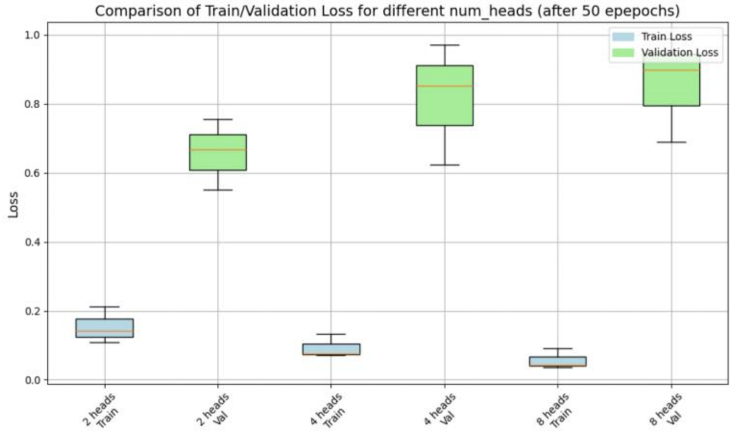
0.01



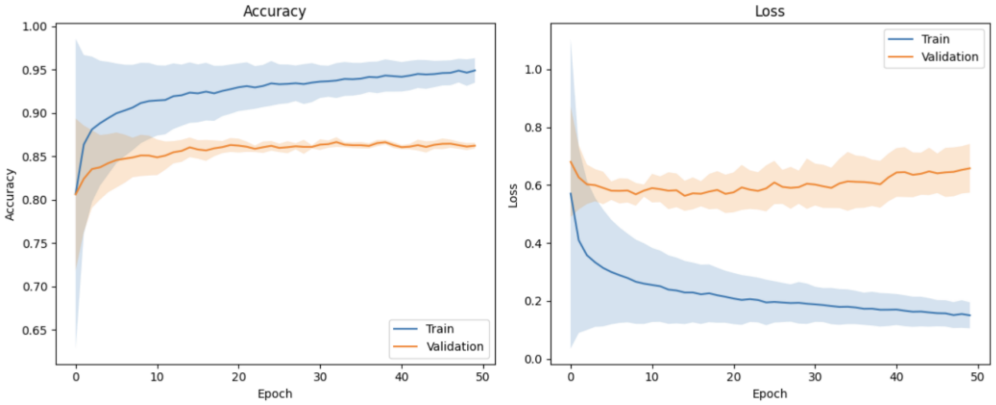
2ND EXPERIMENT

TRANSFORMER COMPARISON – NUMBER OF ATTENTION HEADS

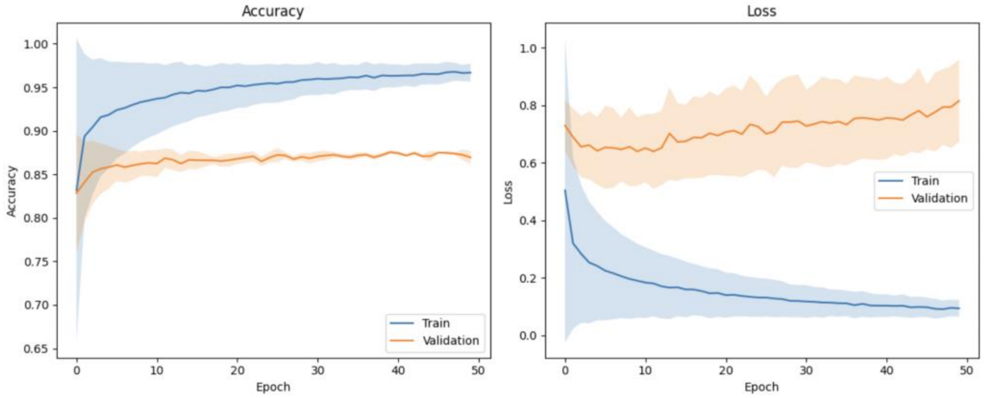




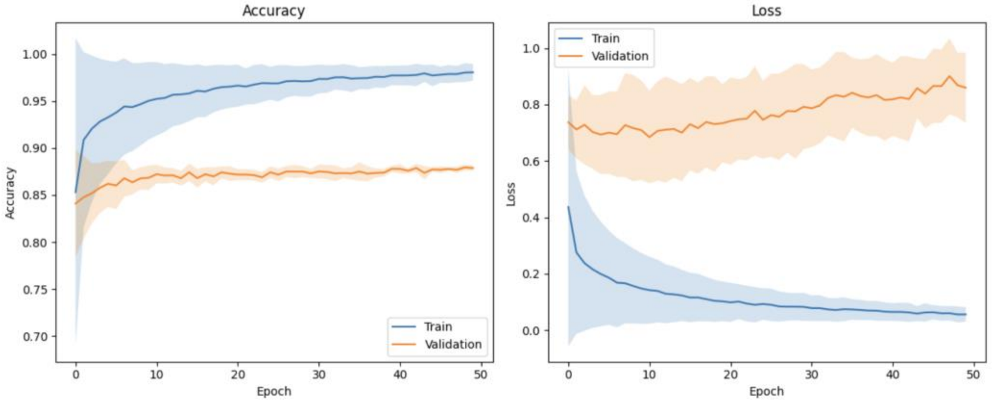
2



4

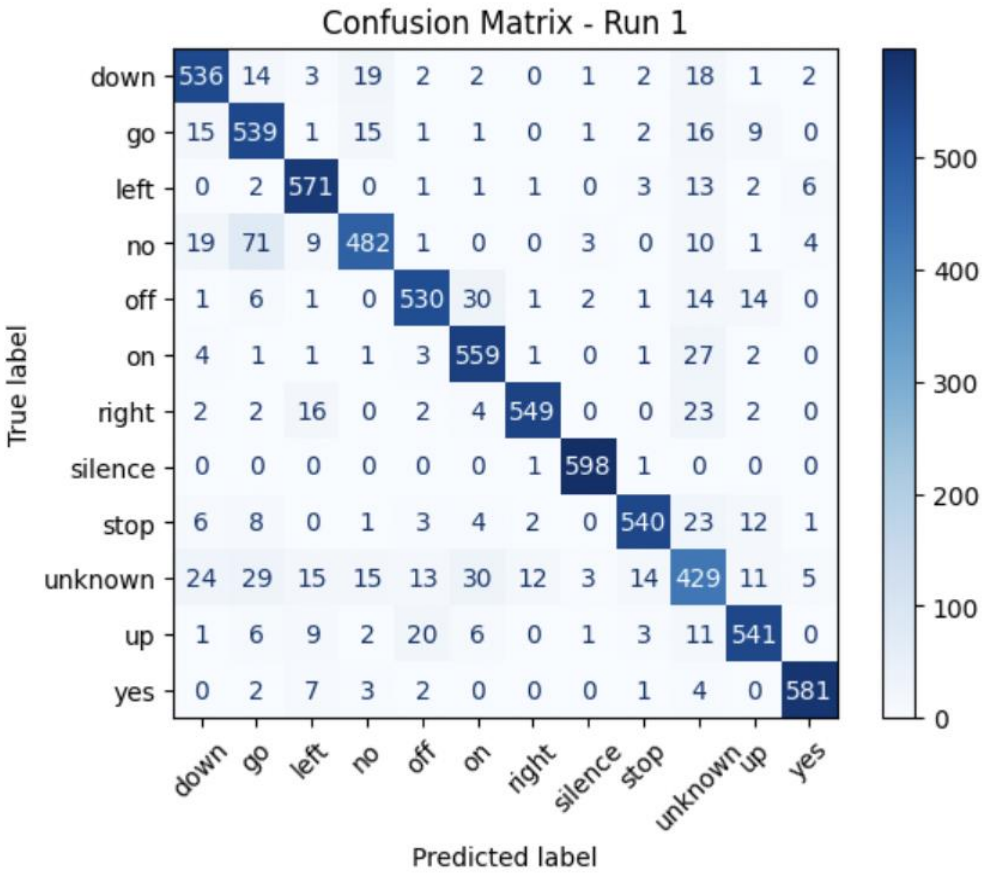


8

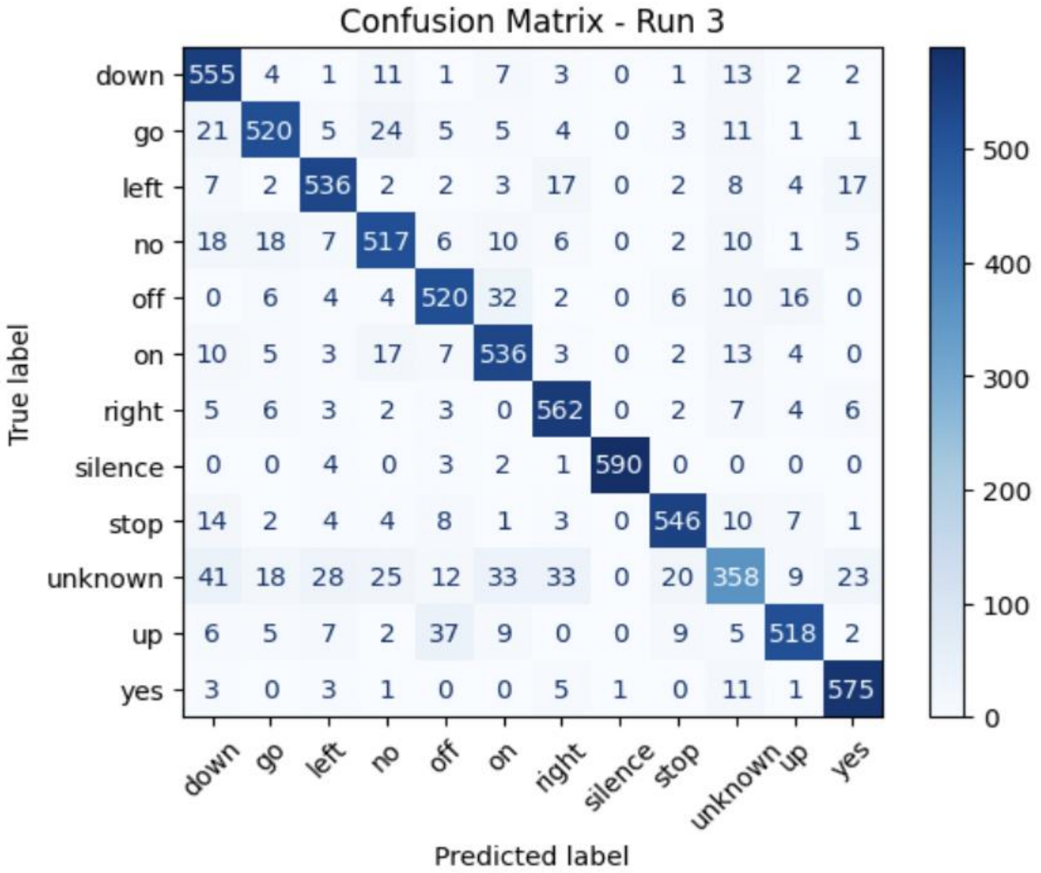


2ND EXPERIMENT

CNN VS TRANSFORMER



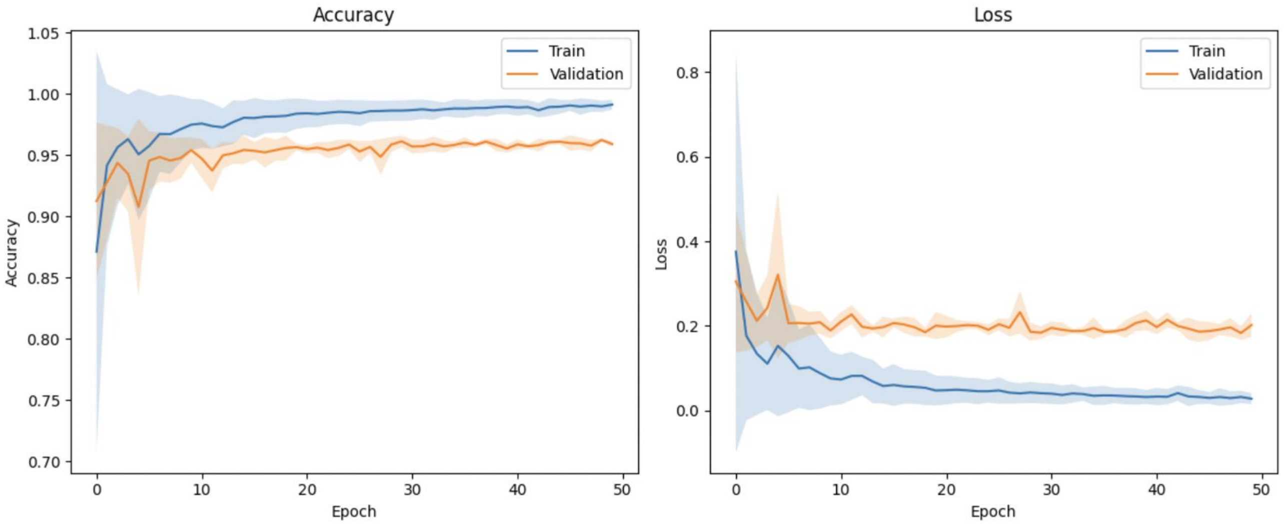
	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	536	14	3	19	2	2	0	1	2	18	1	2
go	15	539	1	15	1	1	0	1	2	16	9	0
left	0	2	571	0	1	1	1	0	3	13	2	6
no	19	71	9	482	1	0	0	3	0	10	1	4
off	1	6	1	0	530	30	1	2	1	14	14	0
on	4	1	1	1	3	559	1	0	1	27	2	0
right	2	2	16	0	2	4	549	0	0	23	2	0
silence	0	0	0	0	0	0	1	598	1	0	0	0
stop	6	8	0	1	3	4	2	0	540	23	12	1
unknown	24	29	15	15	13	30	12	3	14	429	11	5
up	1	6	9	2	20	6	0	1	3	11	541	0
yes	0	2	7	3	2	0	0	0	1	4	0	581

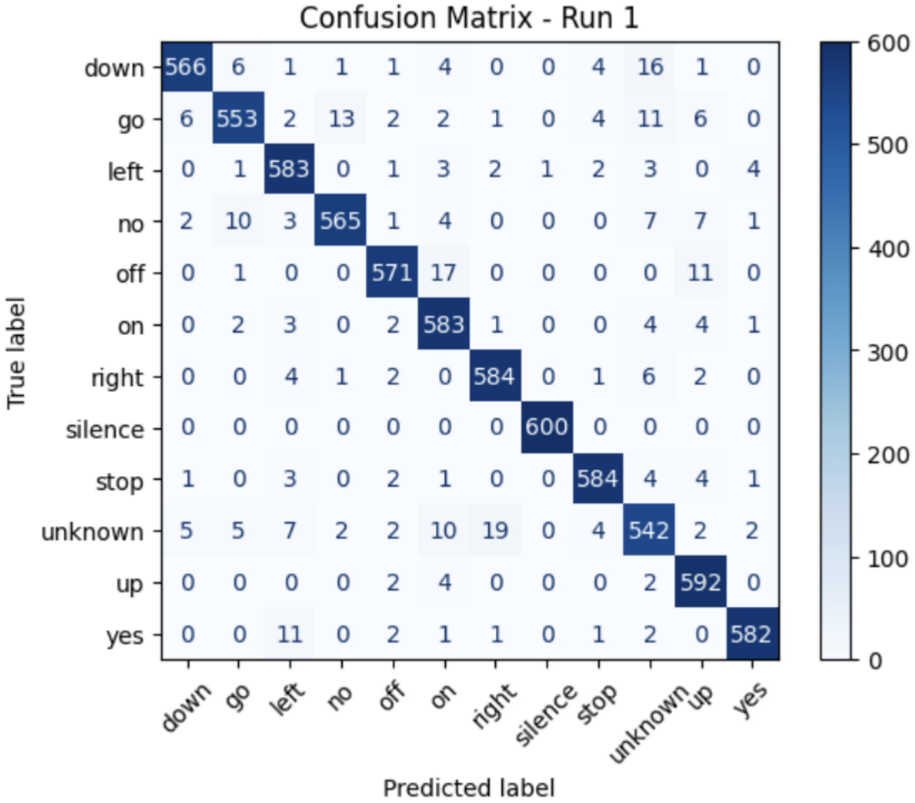


	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	555	4	1	11	1	7	3	0	1	13	2	2
go	21	520	5	24	5	5	4	0	3	11	1	1
left	7	2	536	2	2	3	17	0	2	8	4	17
no	18	18	7	517	6	10	6	0	2	10	1	5
off	0	6	4	4	520	32	2	0	6	10	16	0
on	10	5	3	17	7	536	3	0	2	13	4	0
right	5	6	3	2	3	0	562	0	2	7	4	6
silence	0	0	4	0	3	2	1	590	0	0	0	0
stop	14	2	4	4	8	1	3	0	546	10	7	1
unknown	41	18	28	25	12	33	33	0	20	358	9	23
up	6	5	7	2	37	9	0	0	9	5	518	2
yes	3	0	3	1	0	0	5	1	0	11	1	575

2ND EXPERIMENT

TRANSFORMER + CNN ARCHITECTURE





Confusion Matrix - Run 1

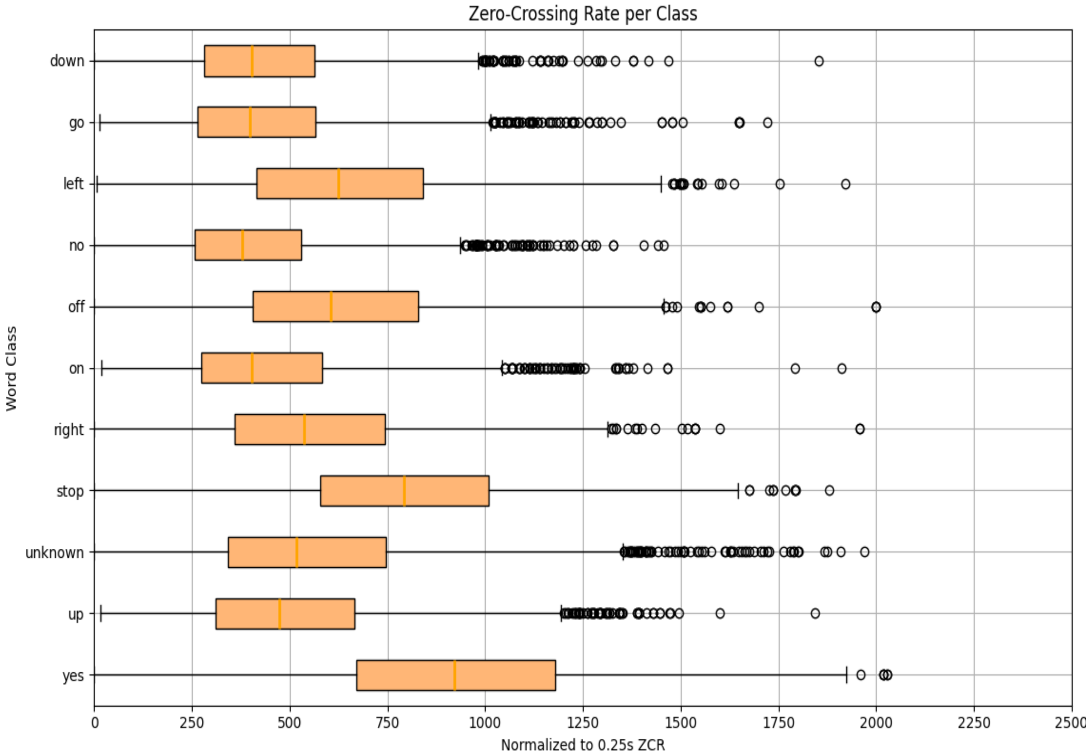
True label \ Predicted label	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	566	6	1	1	1	4	0	0	4	16	1	0
go	6	553	2	13	2	2	1	0	4	11	6	0
left	0	1	583	0	1	3	2	1	2	3	0	4
no	2	10	3	565	1	4	0	0	0	7	7	1
off	0	1	0	0	571	17	0	0	0	0	11	0
on	0	2	3	0	2	583	1	0	0	4	4	1
right	0	0	4	1	2	0	584	0	1	6	2	0
silence	0	0	0	0	0	0	0	600	0	0	0	0
stop	1	0	3	0	2	1	0	0	584	4	4	1
unknown	5	5	7	2	2	10	19	0	4	542	2	2
up	0	0	0	0	2	4	0	0	0	2	592	0
yes	0	0	11	0	2	1	1	0	1	2	0	582

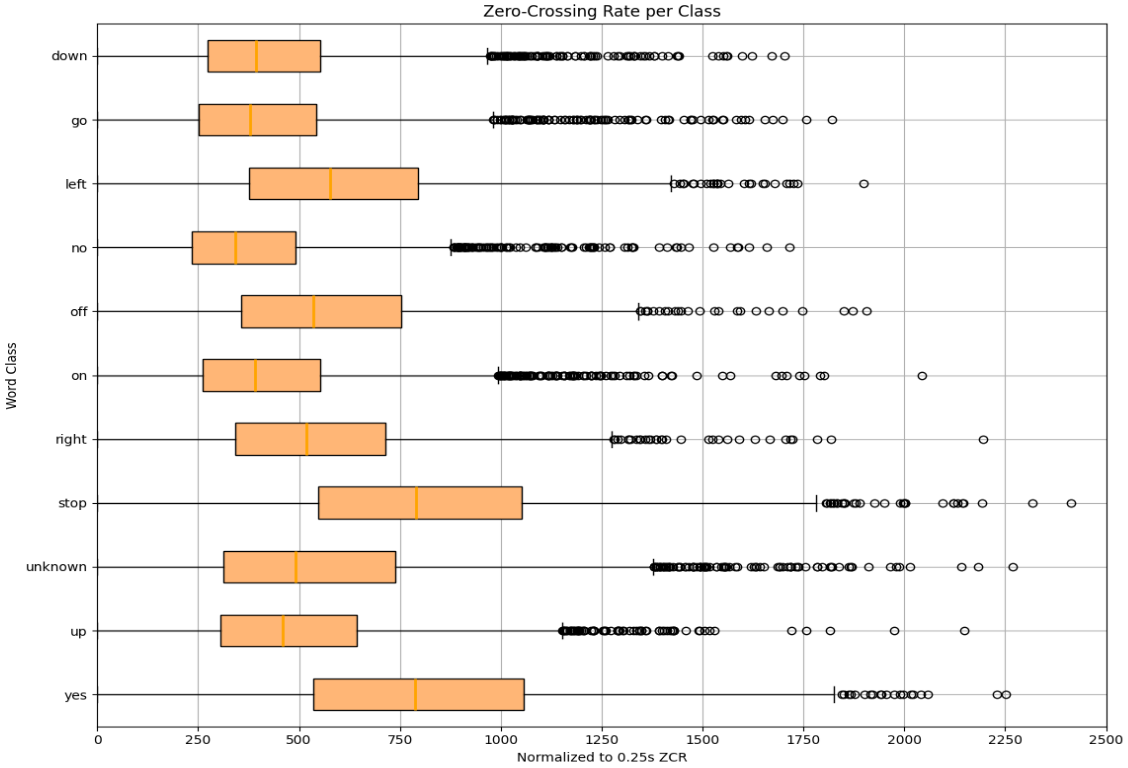
DATA AUGMENTATION -> BETTER GENERALIZATION

- We have large dataset but... we can still enlarge it with also changing a little bit our data to reflects on more ways how people accent words
- People can say word slower or faster, can accent something louder or ending word really quietly, or in general speaking quietly, can also speak in a noisy space
- We want to reflect all these situation by applying the following augmentations: stretching speech, changing volume, adding noise

3RD EXPERIMENT

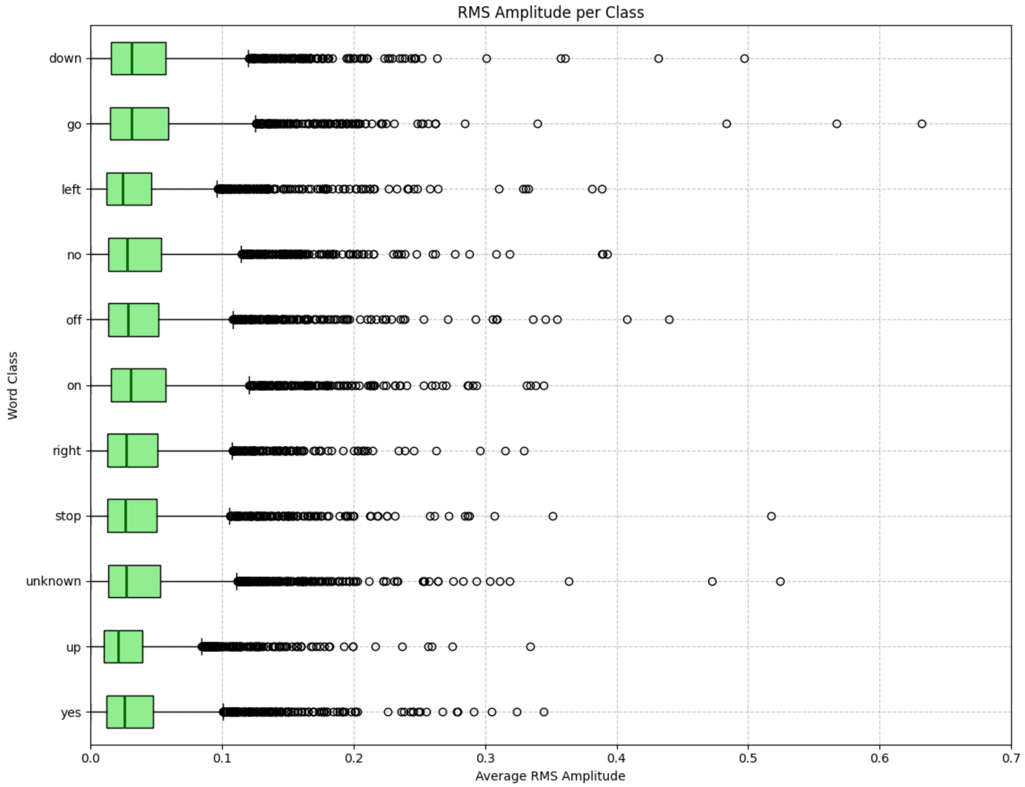
STRETCHING

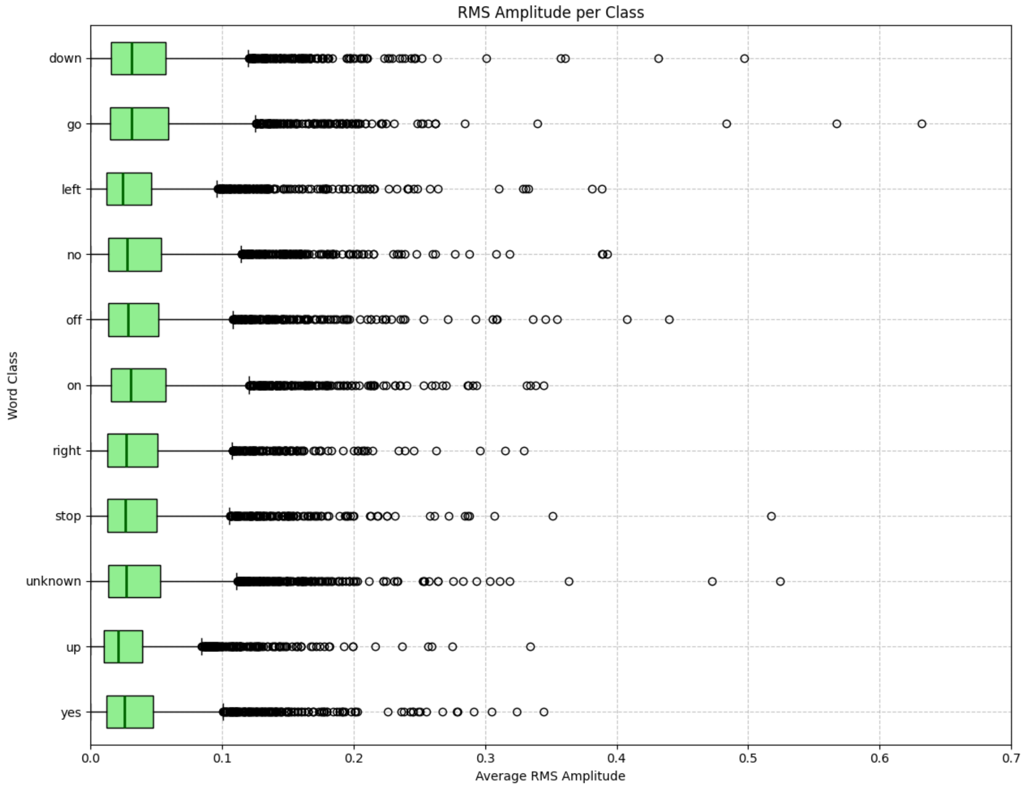




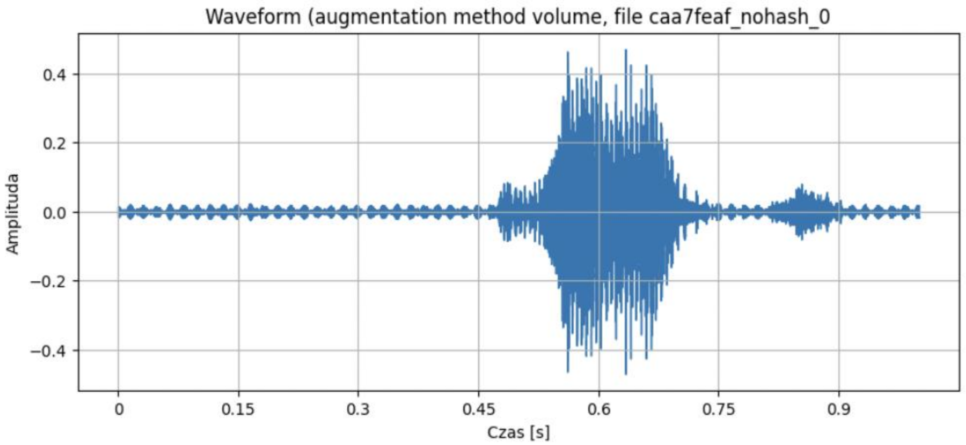
3RD EXPERIMENT

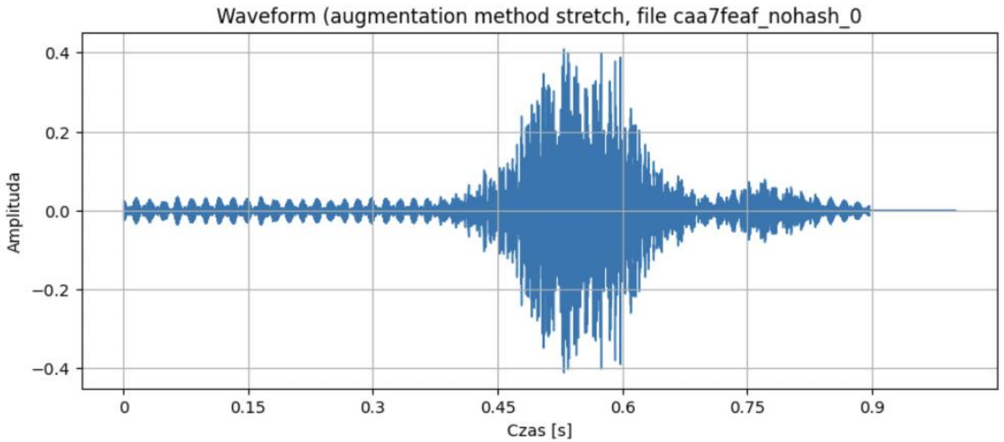
VOLUME CHANGING

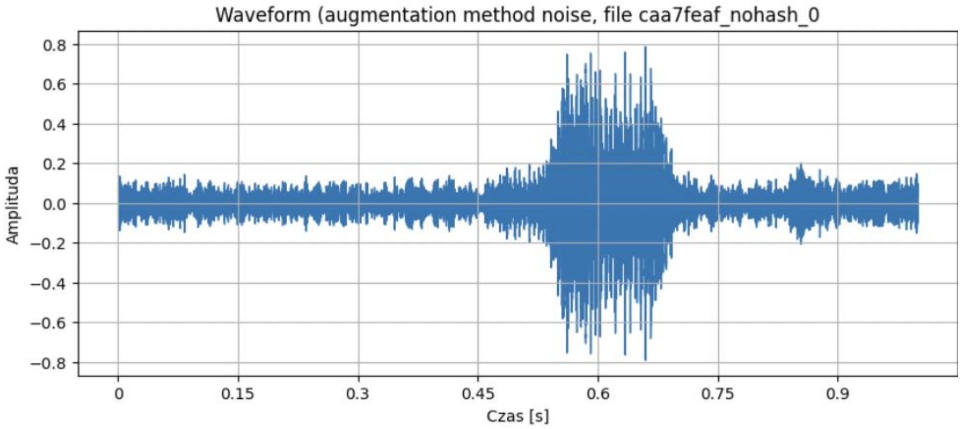


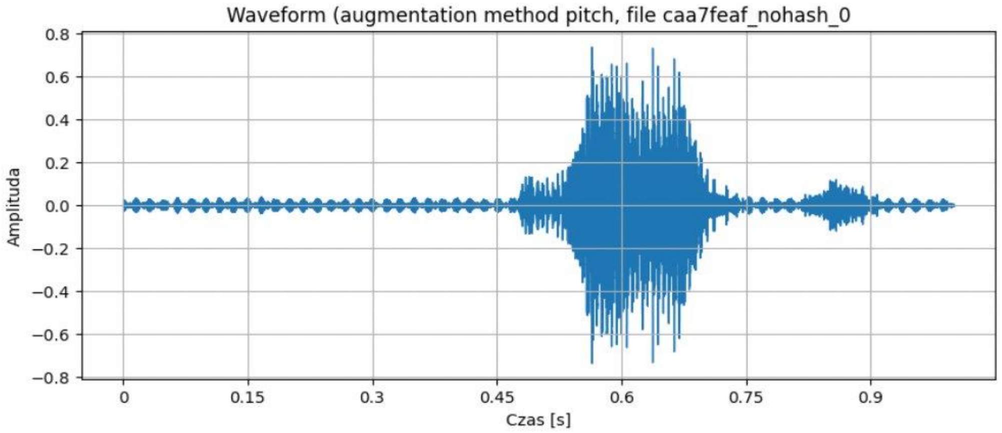


DATA AUGMENTATION TECHNIQUES



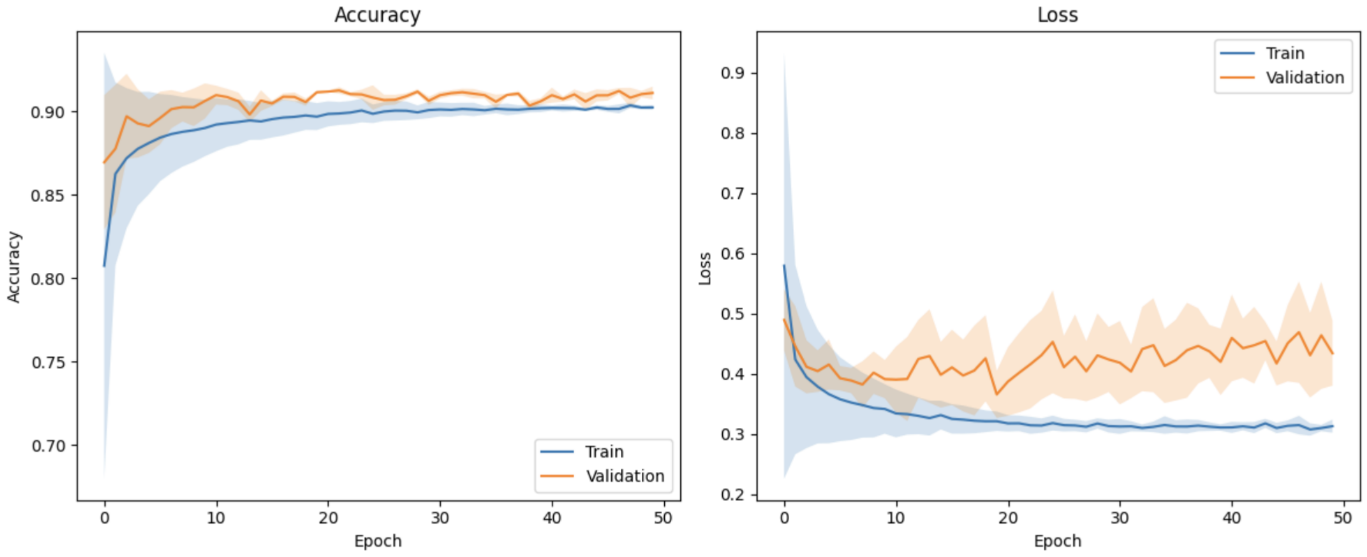


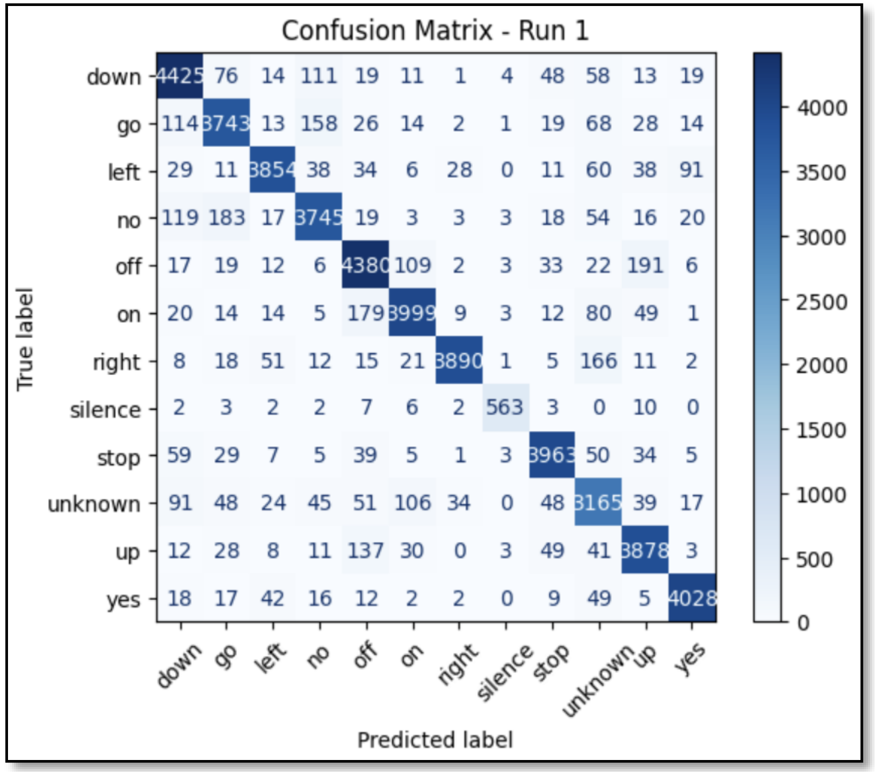




3RD EXPERIMENT

CNN FOR AUGMENTED DATA



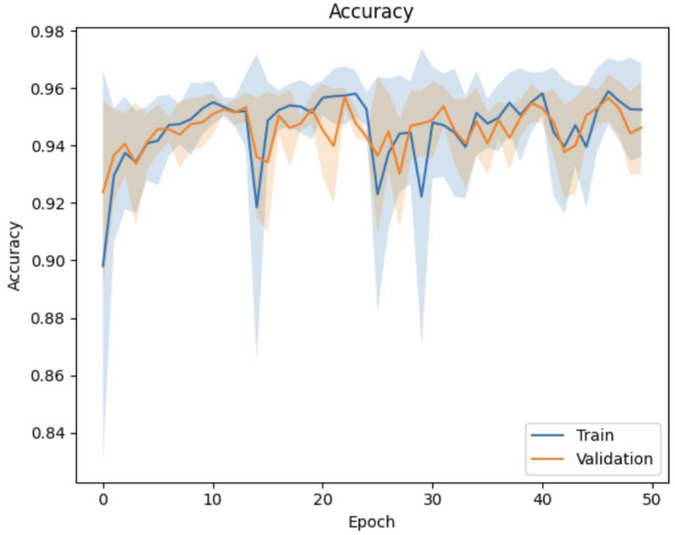

Confusion Matrix - Run 1

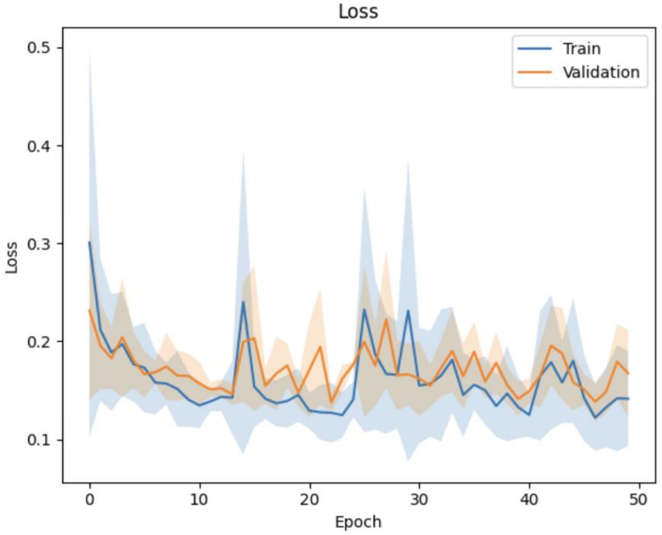
True label \ Predicted label	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	4425	76	14	111	19	11	1	4	48	58	13	19
go	114	3743	13	158	26	14	2	1	19	68	28	14
left	29	11	3854	38	34	6	28	0	11	60	38	91
no	119	183	17	3745	19	3	3	3	18	54	16	20
off	17	19	12	6	4380	109	2	3	33	22	191	6
on	20	14	14	5	179	3999	9	3	12	80	49	1
right	8	18	51	12	15	21	3890	1	5	166	11	2
silence	2	3	2	2	7	6	2	563	3	0	10	0
stop	59	29	7	5	39	5	1	3	3963	50	34	5
unknown	91	48	24	45	51	106	34	0	48	3165	39	17
up	12	28	8	11	137	30	0	3	49	41	3878	3
yes	18	17	42	16	12	2	2	0	9	49	5	4028

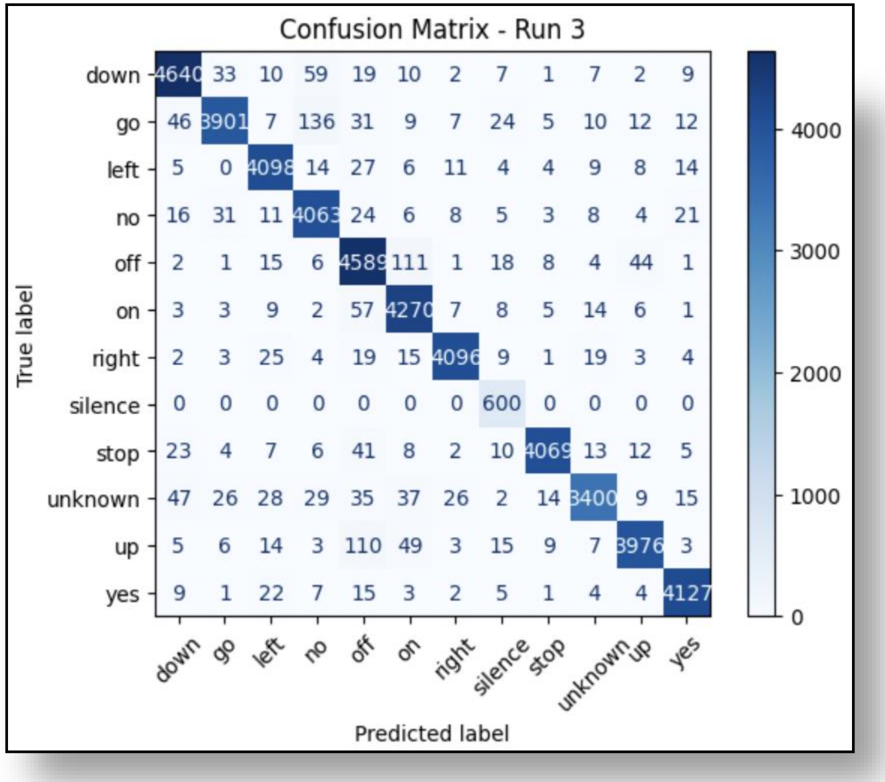
AUGMENTED DATA	ORIGINAL DATA
91,57%	89,65%

3RD EXPERIMENT

TRANSFORMER+CNN FOR AUGMENTED DATA






Confusion Matrix - Run 3

	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	4640	33	10	59	19	10	2	7	1	7	2	9
go	46	3901	7	136	31	9	7	24	5	10	12	12
left	5	0	4098	14	27	6	11	4	4	9	8	14
no	16	31	11	4063	24	6	8	5	3	8	4	21
off	2	1	15	6	4589	111	1	18	8	4	44	1
on	3	3	9	2	57	4270	7	8	5	14	6	1
right	2	3	25	4	19	15	4096	9	1	19	3	4
silence	0	0	0	0	0	0	0	600	0	0	0	0
stop	23	4	7	6	41	8	2	10	4069	13	12	5
unknown	47	26	28	29	35	37	26	2	14	3400	9	15
up	5	6	14	3	110	49	3	15	9	7	3976	3
yes	9	1	22	7	15	3	2	5	1	4	4	4127

AUGMENTED DATA	ORIGINAL DATA
96,17%	95,90%

FUTURE GOALS

- Explore more augmentation techniques
- Analyse different ensemble models (focus on classes like on/off)
- Explore different architectures of Transformer+CNN
- Test more hyperparameters
- Extend the range of possible labels

THANK YOU FOR YOUR ATTENTION!