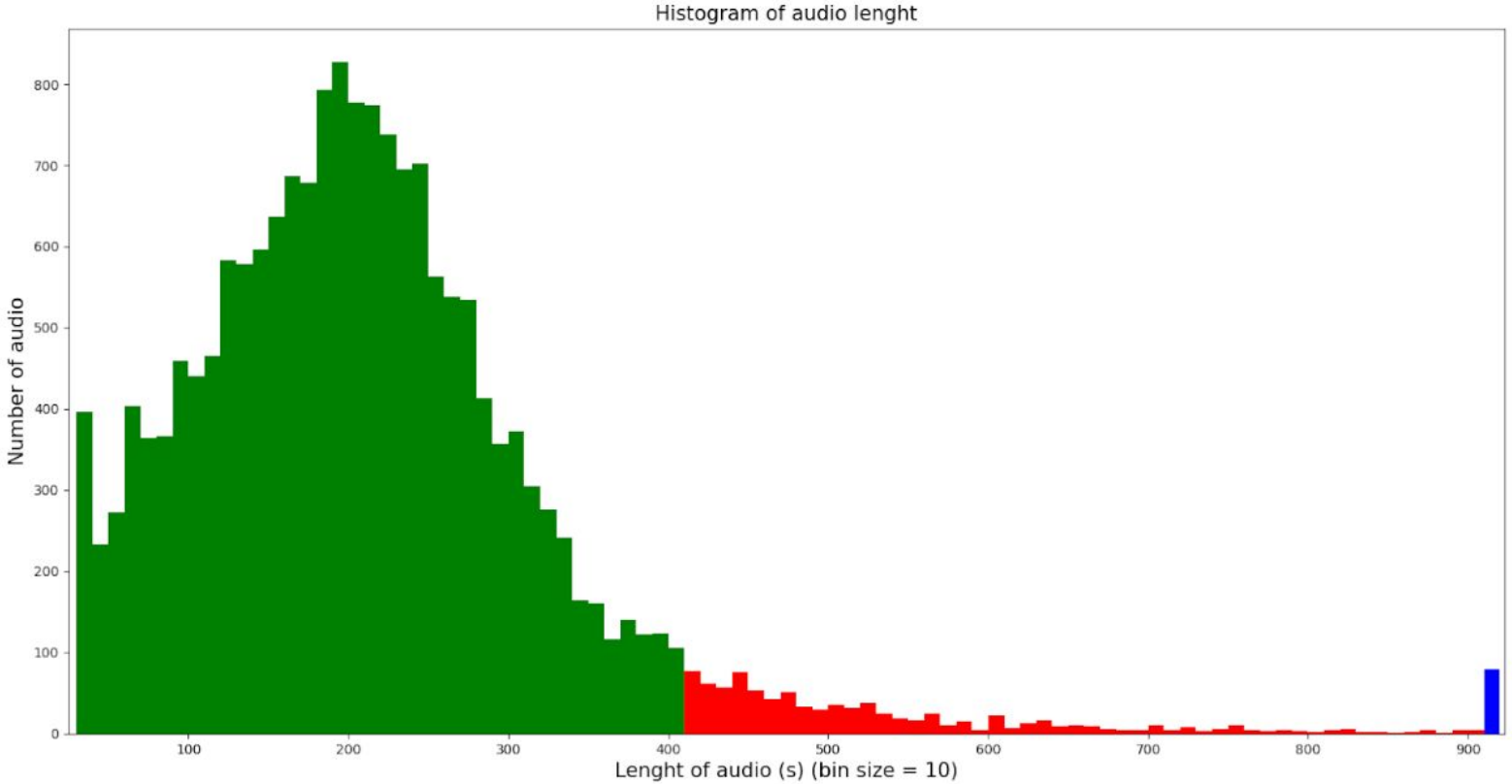
HCMUS at MediaEval 2020:

Emotion Classification Using Wavenet Features with SpecAugment and EfficientNet

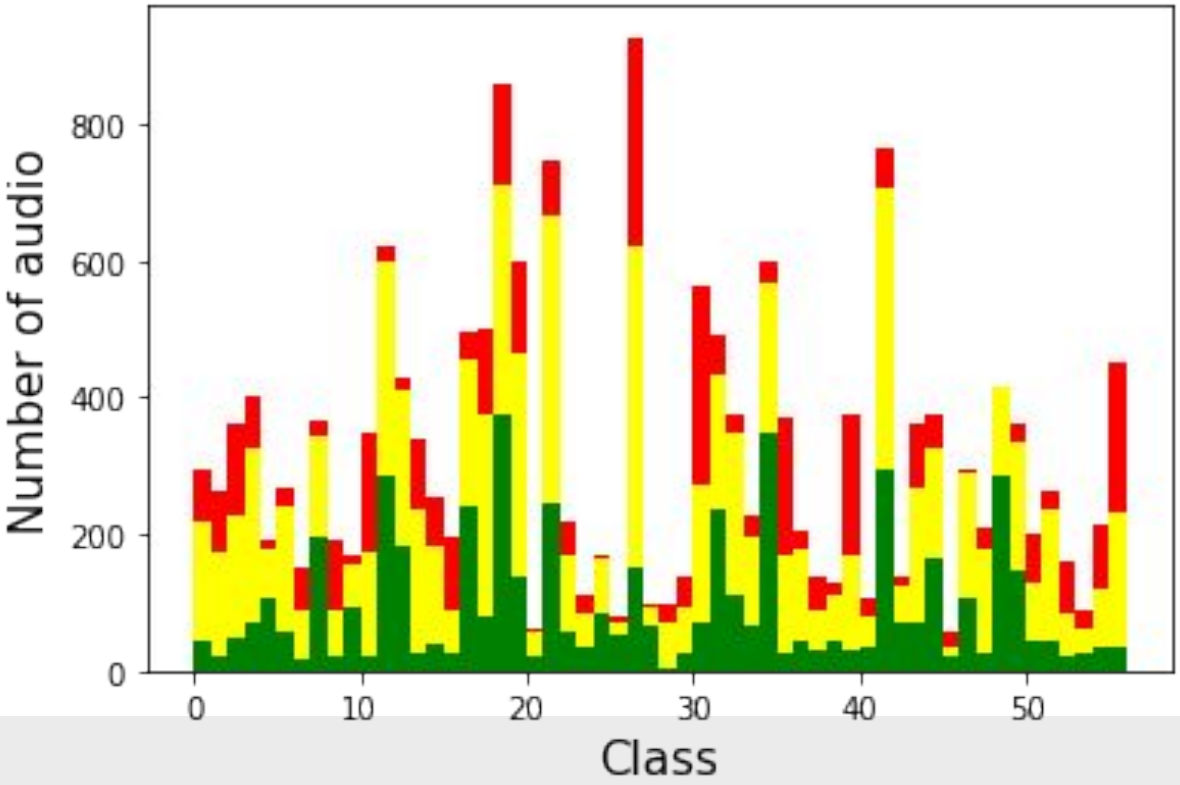
Tri-Nhan Do, Minh-Tri Nguyen, Hai-Dang Nguyen, Minh-Triet Tran, Xuan-Nam Cao

Data analysis

Histogram of Audio length

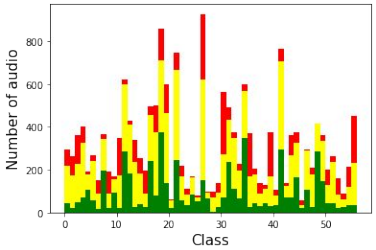


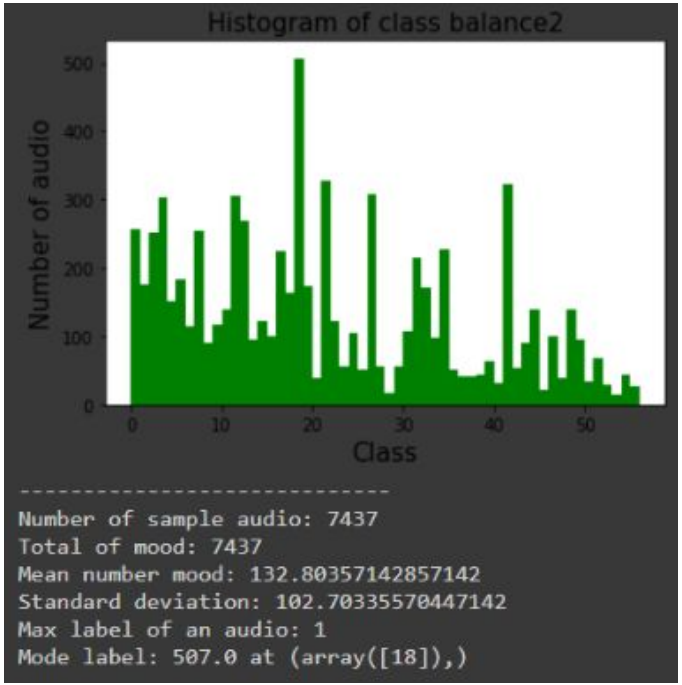
Histogram of mood and theme of training set

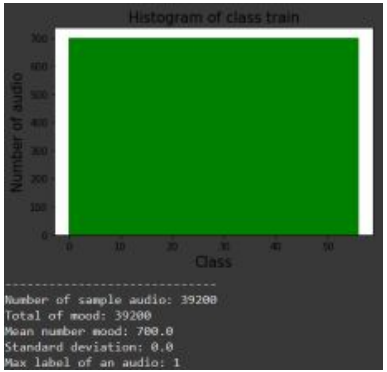


- The green part shows the audio with only one class
- The yellow part shows the audio with 2 to 3 classes,
- The red part shows the audio with more than 3 classes.
- The maximum number of moods of an audio is 8.
- Mood /theme that appears most is happy with 927 audios

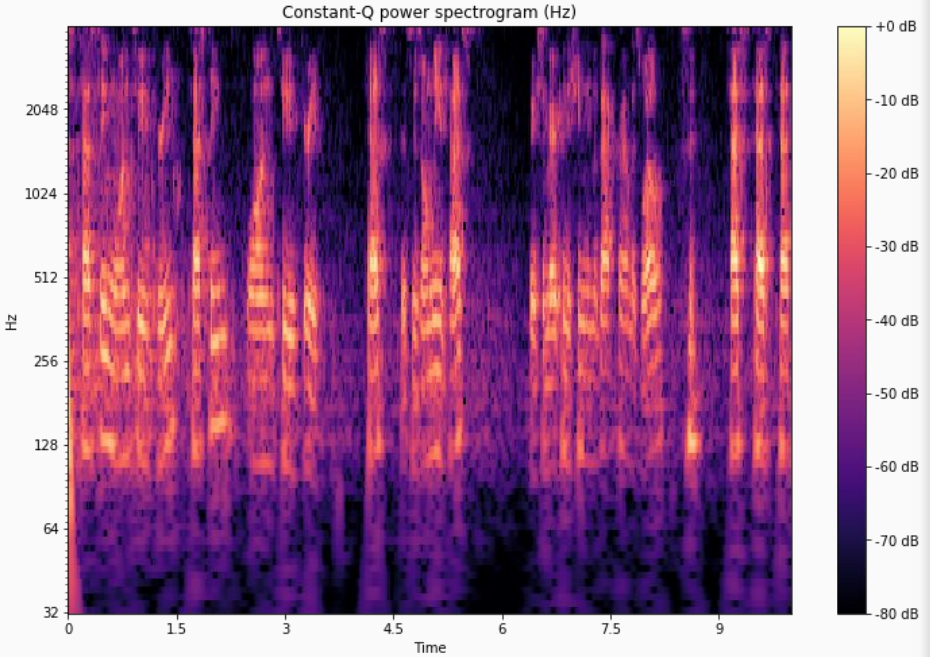
Data Preprocessing







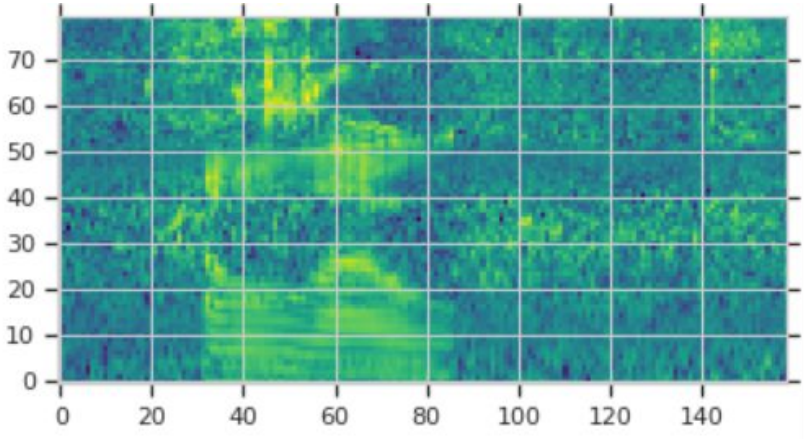
Features Processing

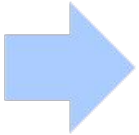


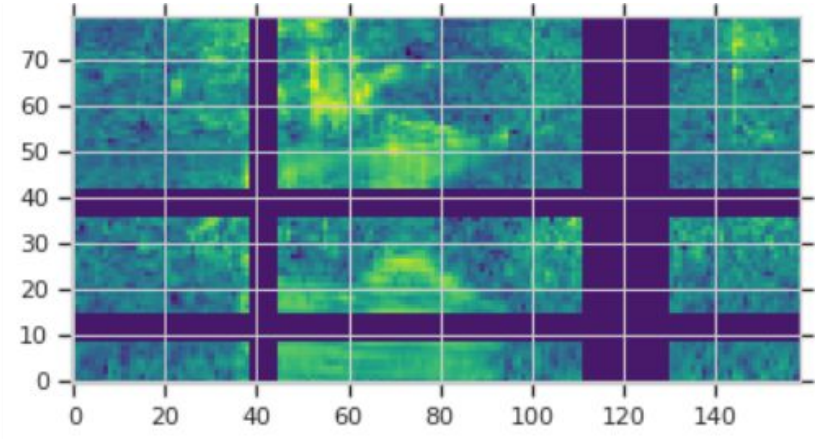
Mel-spectrogram

- Each sample feature has 96 channels
- Time frames are randomly cropped to 6950 after each epoch

SpecAugment







0.139 PR-AUC-macro

0.139 PR-AUC-macro

- Each input have 70% chance to be augmented by using SpecAugment
- Each mel-spectrogram will have two blocks of time masking and two blocks of channel masking.

RESIDUAL AND SKIP CONNECTIONS

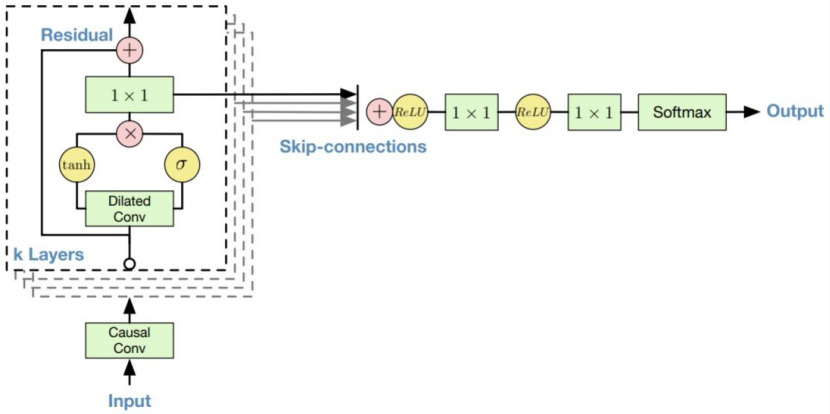


Figure 4: Overview of the residual block and the entire architecture.

Wavenet as Features for Classification

MUSIC ARTIST CLASSIFICATION WITH WAVENET CLASSIFIER FOR RAW WAVEFORM AUDIO DATA

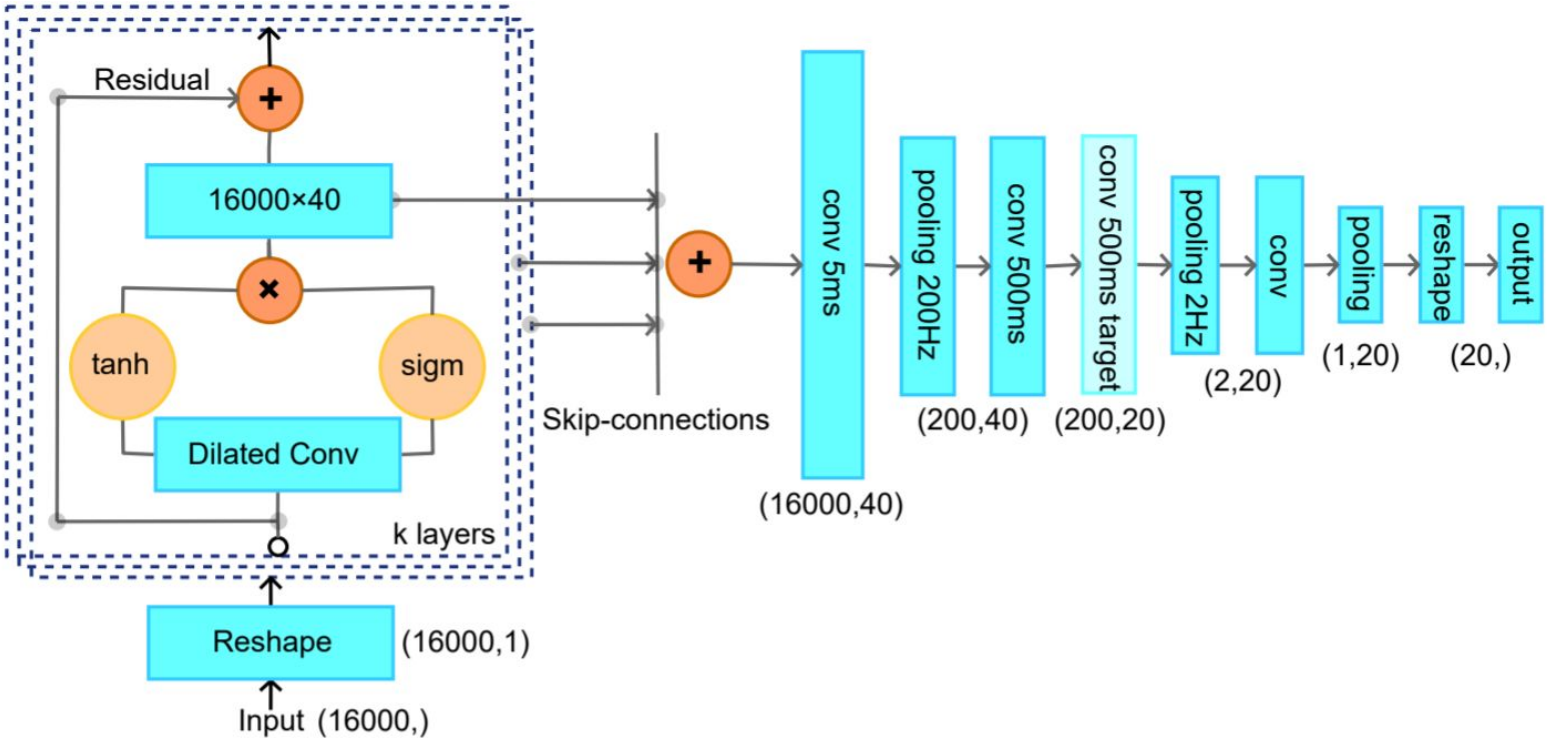


Figure 1: Overview of the WaveNet based deep model architecture. The left part is wavenet for encoder and the right part is CNN for the final classification.

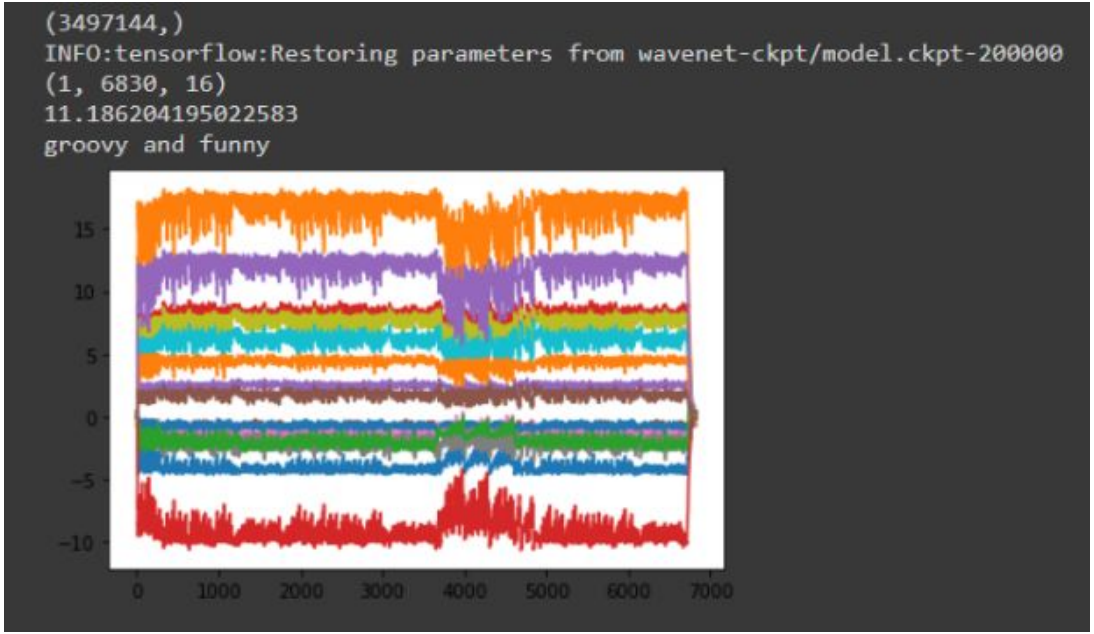
Emotion Recognition from Raw Speech using Wavenet

TABLE I

RECOGNITION ACCURACIES, NUMBER OF TRAINABLE PARAMETERS, TRAIN LOSS AND TEST LOSS OF WAVENET AND CNN+LSTM ARCHITECTURE FOR SER FROM RAW SPEECH USING EMO-DB DATASET.

Method	Parameter	Train loss	Test loss	Accuracy
Wavenet	**29,562**	0.4451	0.4024	**83.82%**
CNN+LSTM	16,736,324	0.0512	0.6217	73.52 %

Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders



- Use WaveNet-style autoencoder model
- This model was pretrained from high-quality dataset of musical notes Nsynth
- The output of a 30 seconds audio is 16 frames multiply with 937-time steps

Approach Overview

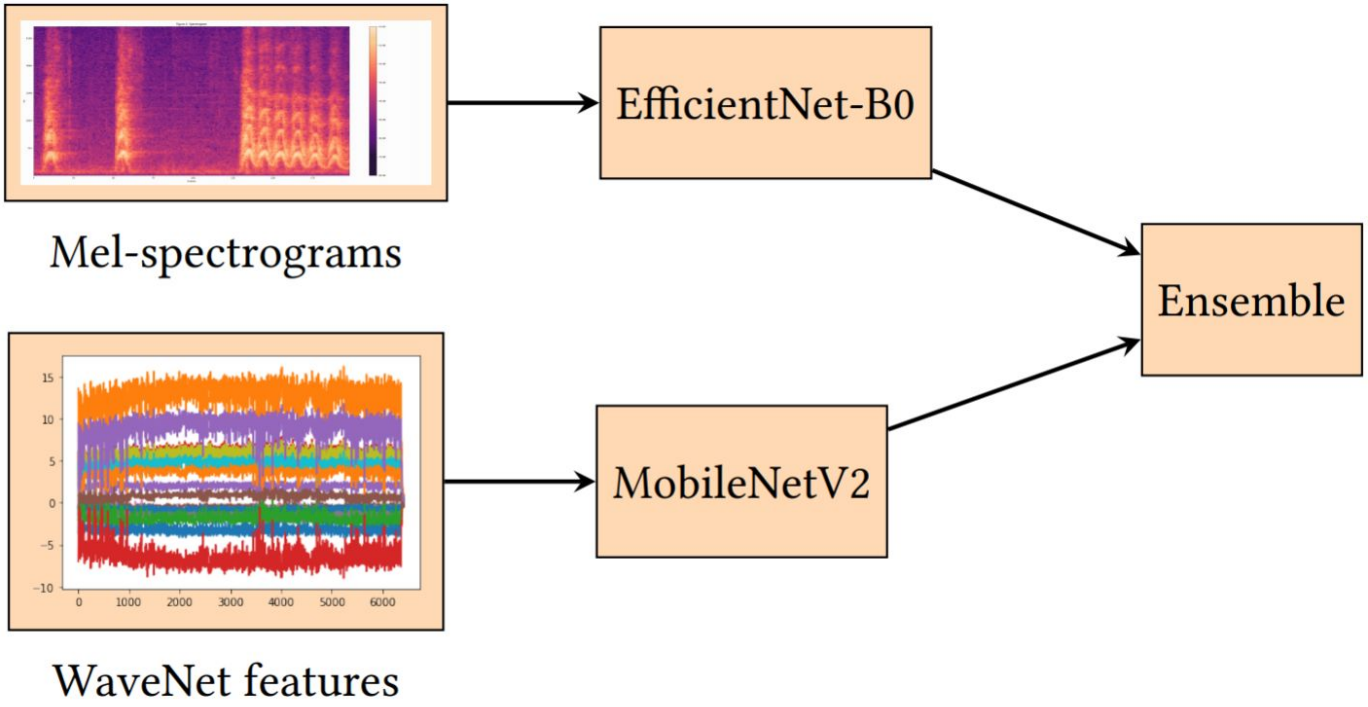
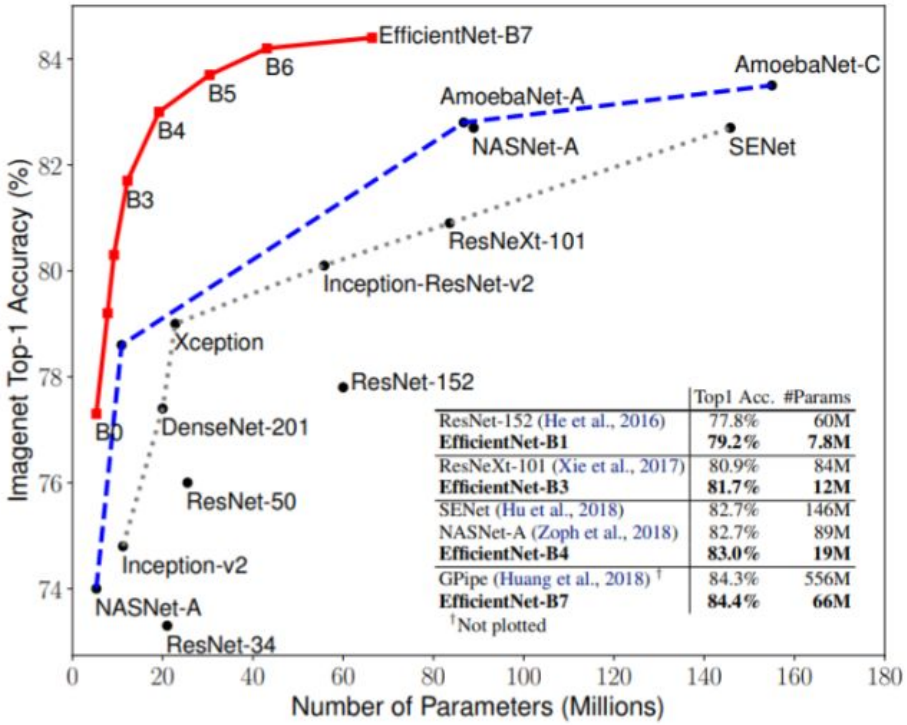


Figure 2: Overview of submission 1.

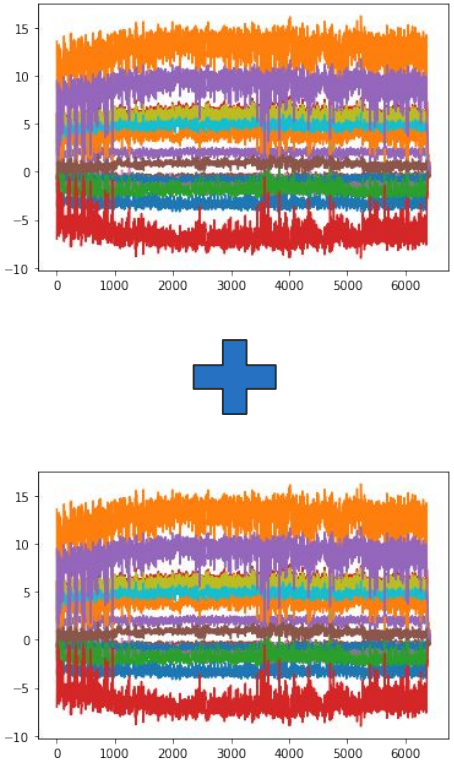
EfficientNet



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.2%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.7%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**83.0%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

† Not plotted

Try to use Efficient-B8 for Wavenet Feature

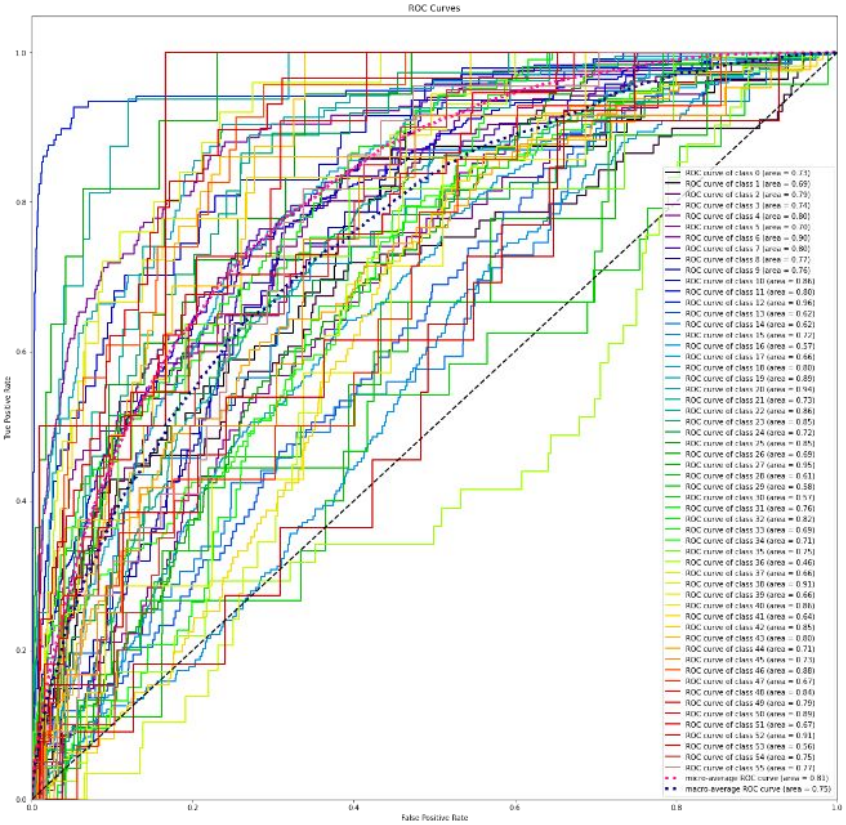


Shape (937,32)

Efficient-B8

Experiment results

Method	Features and Model	PR-AUC-macro
A	Mel-spectrogram EfficientNet-B0	0.127
B	Mel-spectrogram EfficientNet-B0 with data processing	0.134
C (run2)	Mel-spectrogram EfficientNet-B0 using augmentation	0.139
D	WaveNet MobileNetV2	0.102
E (run3)	WaveNet EfficientNet-B7	0.105
F (run1)	Ensemble C and D	0.1413
G (run4)	Ensemble C and E	0.1414



Conclusion

- The EfficientNet model was shown to be more efficient than previous models
- Wavenet can be considered as a features from signal, can extract other aspects of the dataset.