Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

ENVIRONMENTAL SOUND CLASSIFICATION USING DEEP LEARNING

Research and Development Project

March 8, 2021

Manoj Kolpe Lingappa

Prof. Dr. Paul G. Plöger
Dr. Anastassia Küstenmacher

Introduction

Information in everyday soundscapes

- Environmental sounds are everywhere

Figure 1: Savignyplatz street in berlin [1]

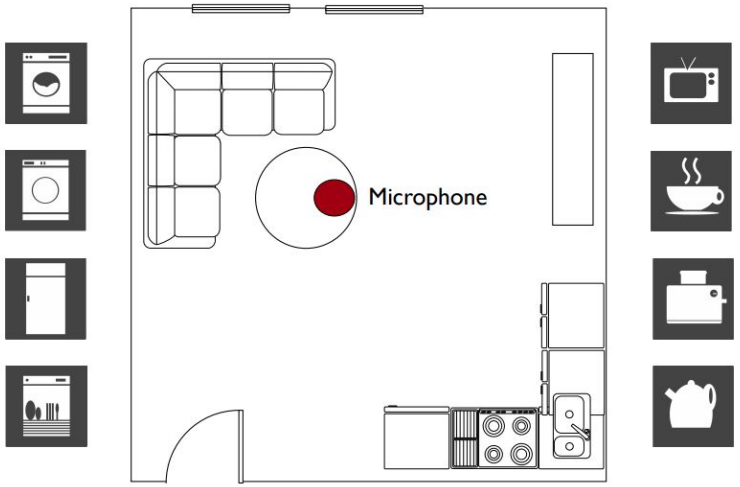


Figure 2: Sound sources in home [2]

Environmental Sound Classification

- Intelligent Sound Recognition (ISR) identifies sounds in the real environment [3]
- Analyzing human auditory and embed such percept ability in machines or robots
- Environmental sound classification (ESC) is the fundamental steps of ISR
- **CLASSIFICATION**: Describe sound event using a textual class label

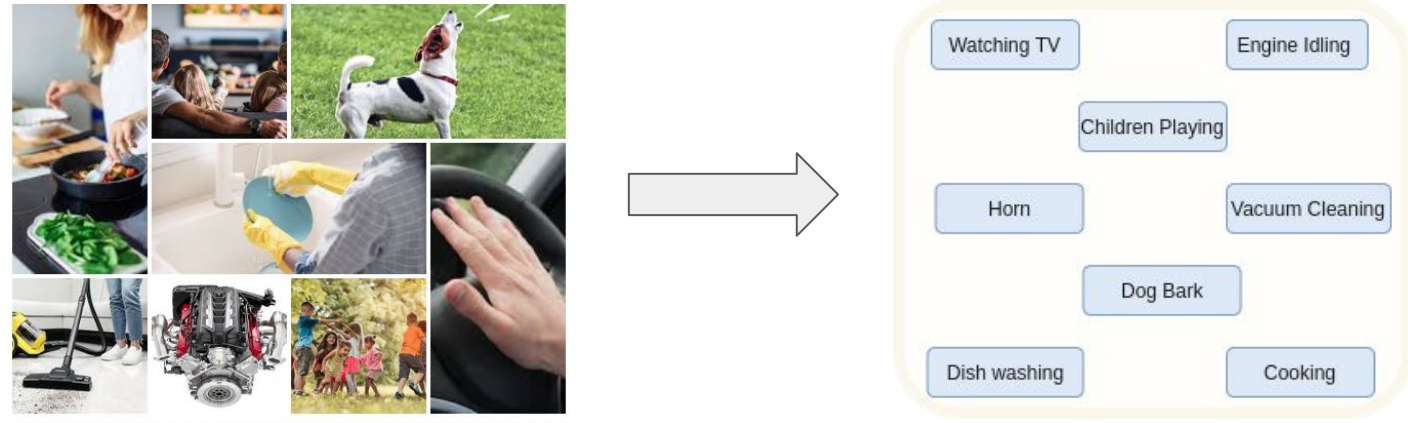


Figure 3: Sound events and class labels [4]

Potential Applications of Acoustic Event Detection

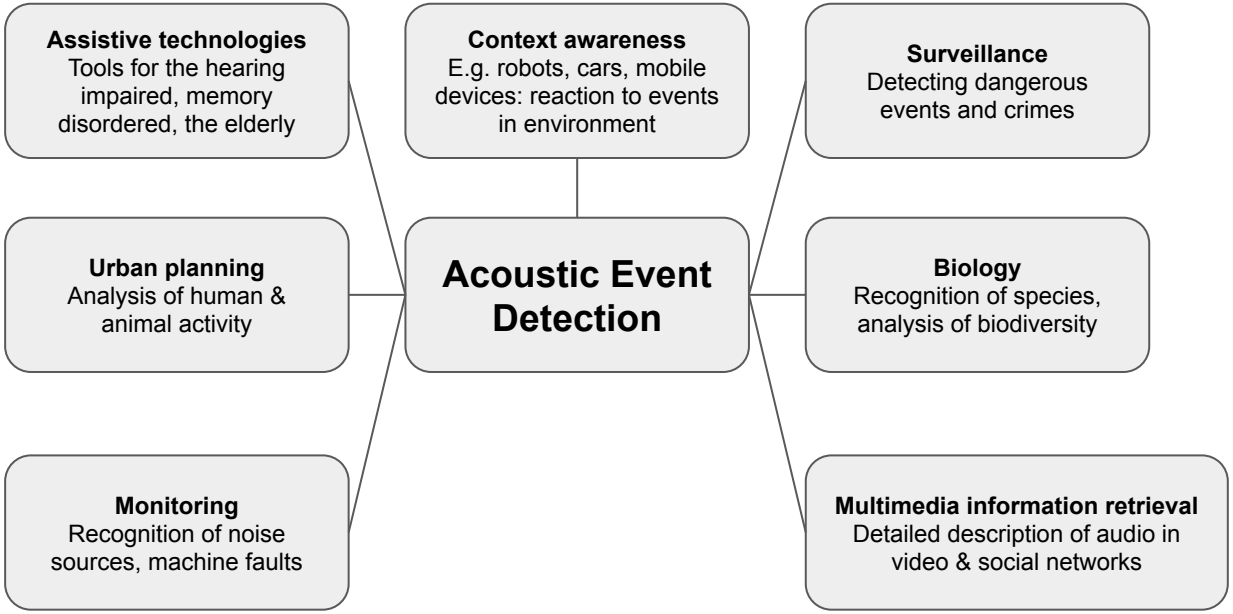
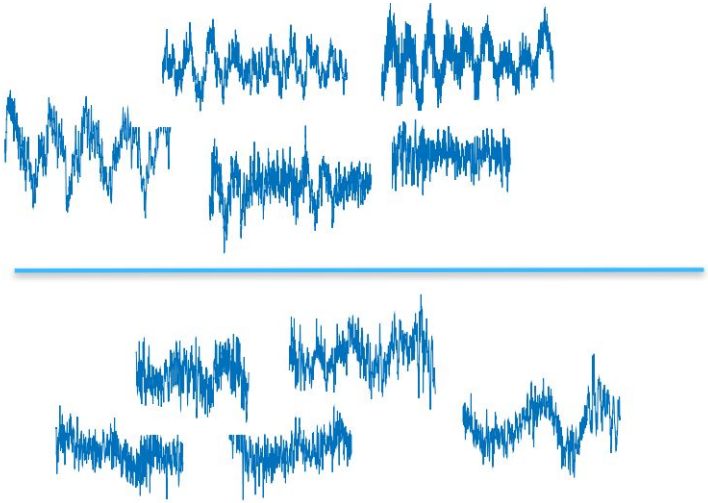
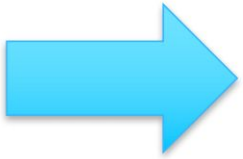


Figure 4: Potential application of acoustic event detection

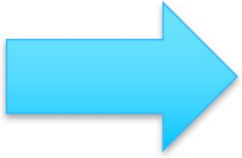
Supervised Machine Learning Approach

- Computer algorithms that find mapping between training examples and labels
- Traditional machine learning and deep learning based approaches





“Dog Barking”



“Vacuum cleaning”

Figure 5: Abstract representation of ESC

Acoustic Features

- Commonly used acoustic features
  - Mel Frequency Cepstral Coefficients (MFCCs) for traditional machine learning
  - Log Mel spectrograms for deep learning based image classification

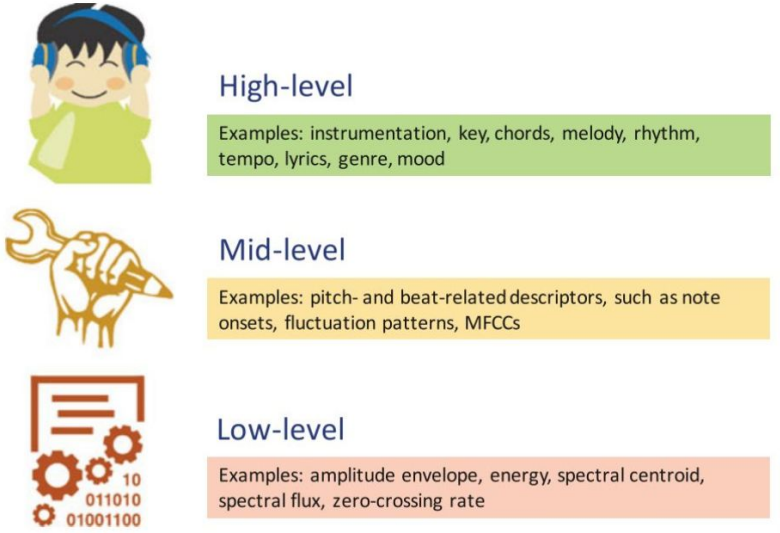


Figure 6: Acoustic features in different levels [6]

Mel Frequency Cepstral Coefficients (MFCCs)

- MFCCs are the short term power spectrum features of an acoustic signal

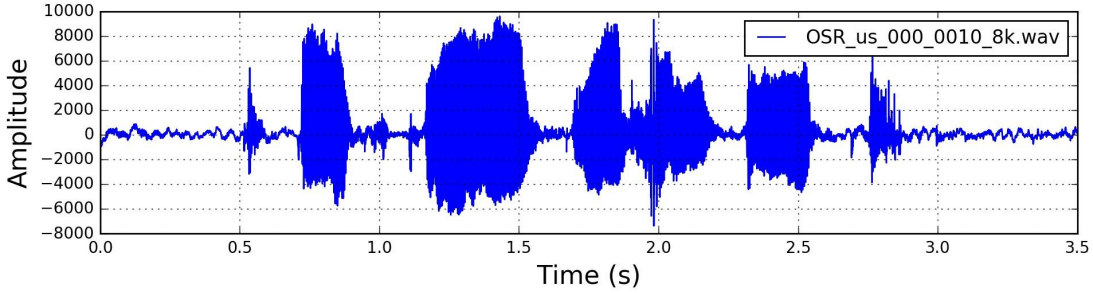


Figure 7: Sound signal

[0.24, 0.56, 0.86,.....,0.67]

Figure 8: MFCCs feature vectors

Log Mel Spectrogram

- A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale
- Raw spectrogram power are log scaled to decibels

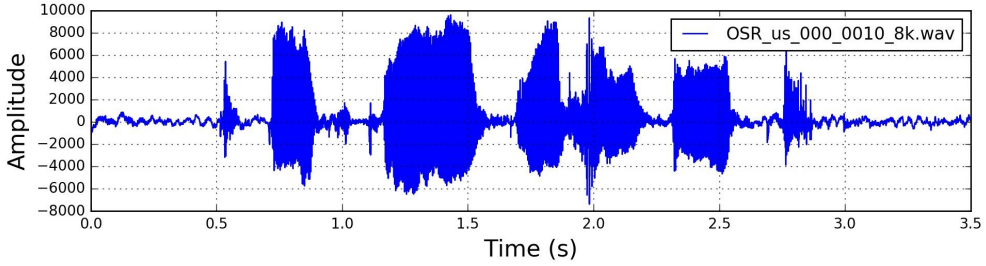


Figure 9: Children playing sound signal

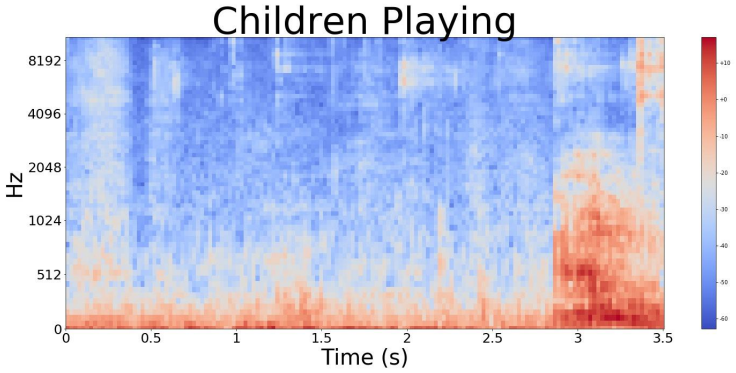


Figure 10: Log Mel spectrogram

Urbansound8K

Urbansounds

- This dataset contains 8732 labeled sound of ≈ 4 (s) length in .wav format at 22.05 kHz

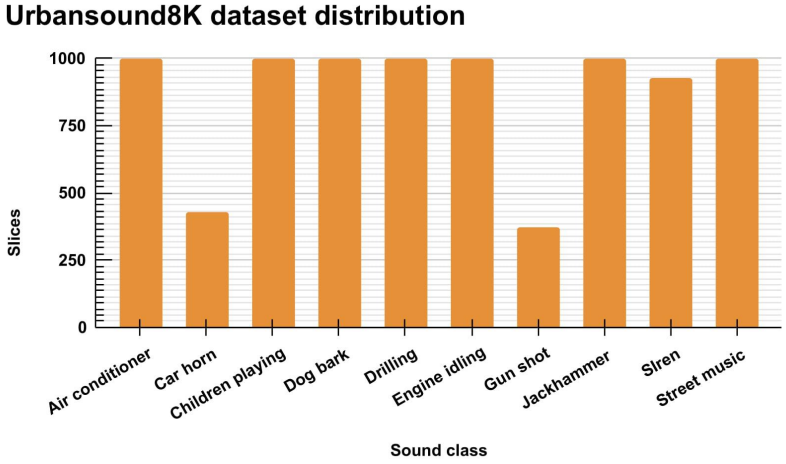


Figure 11: Urbansound8K dataset

DCASE2018

Monitoring of domestic activities based on multi-channel acoustics

- Total of 72984, 10 (s) audio dataset with 9 classes
- The continuous recordings in home environment were split into audio segments of 10 (s)

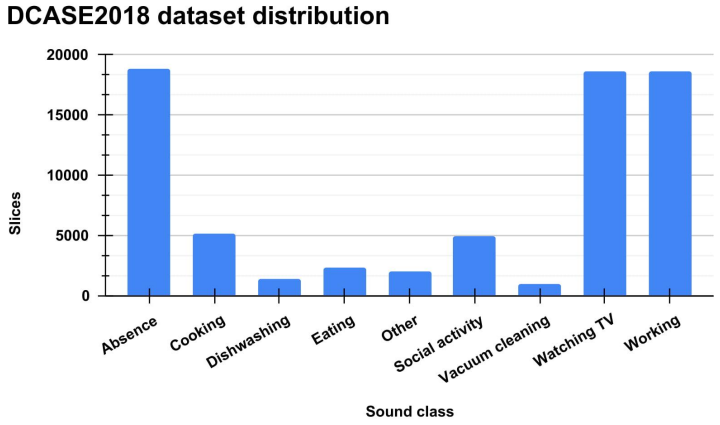


Figure 12: DCASE2018 dataset

Supervised Machine Learning Approach

- Sound event label (classes) defined in advance
- Traditional Machine Learning Techniques
  - K Nearest Neighbor (KNN)
  - Support Vector Machine (SVM)
  - Naive Bayes (NB)
  - Random Forest (RF)
  - Gradient Boosting (GB)
  - XGBoost (XGBoost)
- Convolutional Neural Network based Transfer Learning Model (VGGish)
  - Equal Splitting of Sound Signal
  - Random Samples of Sound Signal

Acoustic Feature Pre-Processing

- Mel Frequency Cepstral Coefficients
  - Re-sampled at 22.05 kHz
- Log-mel Spectrogram
  - Resampled at 22.05 kHz

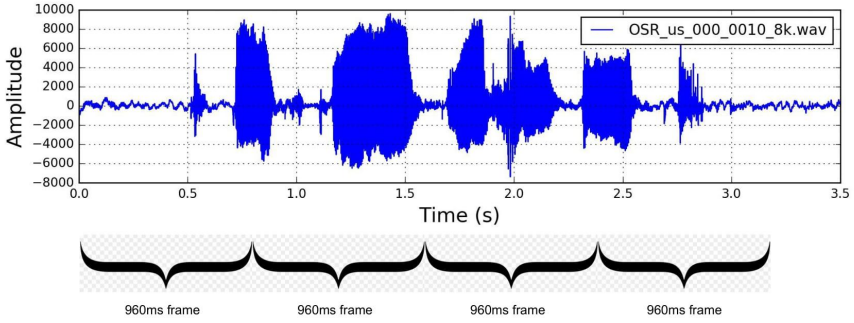


Figure 13: Equal splitting of sound data

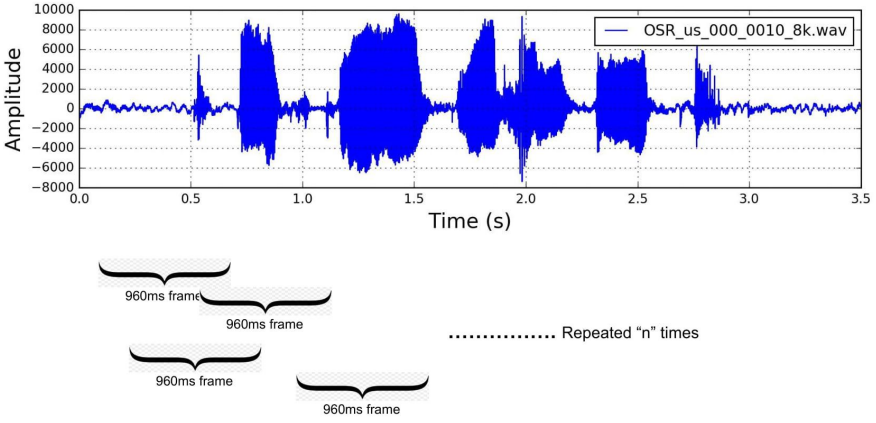


Figure 14: Random sampling of sound data

Acoustic Feature Extraction Pipeline

- Mel Frequency Cepstral Coefficients

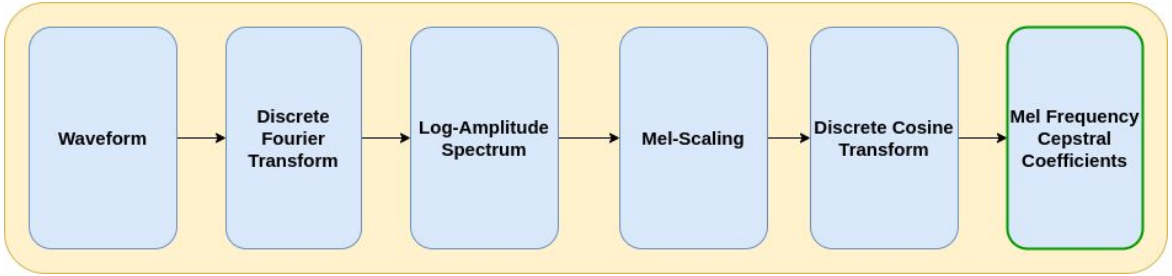


Figure 15: MFCCs extraction pipeline

- Log Mel spectrogram

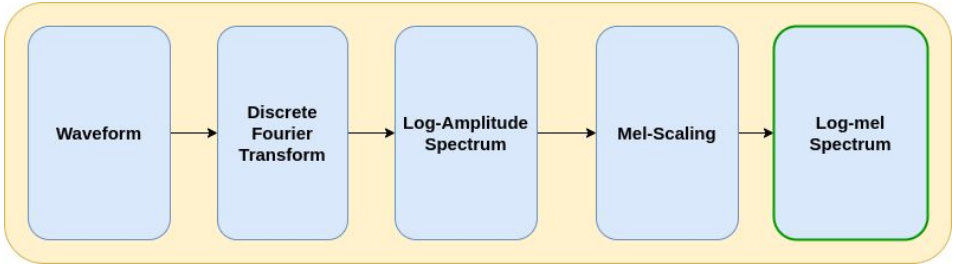


Figure 16: Log Mel spectrogram extraction pipeline

Mel Scale

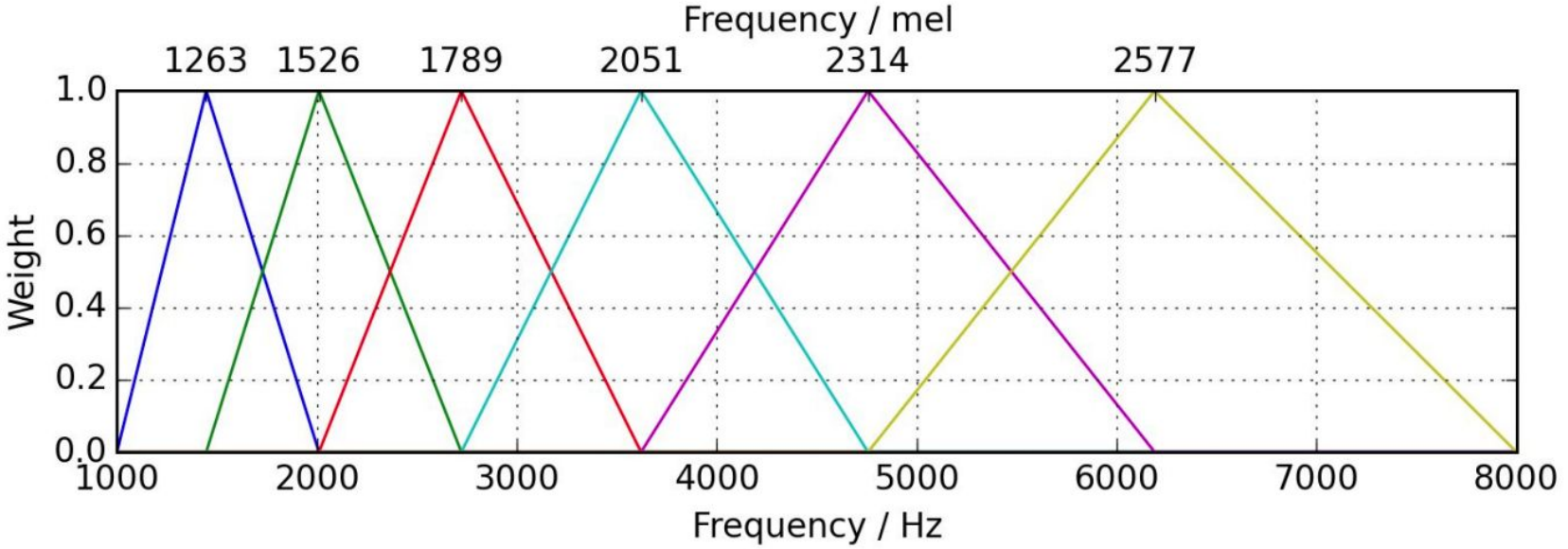


Figure 17: Mel scale

Traditional Machine Learning Pipeline

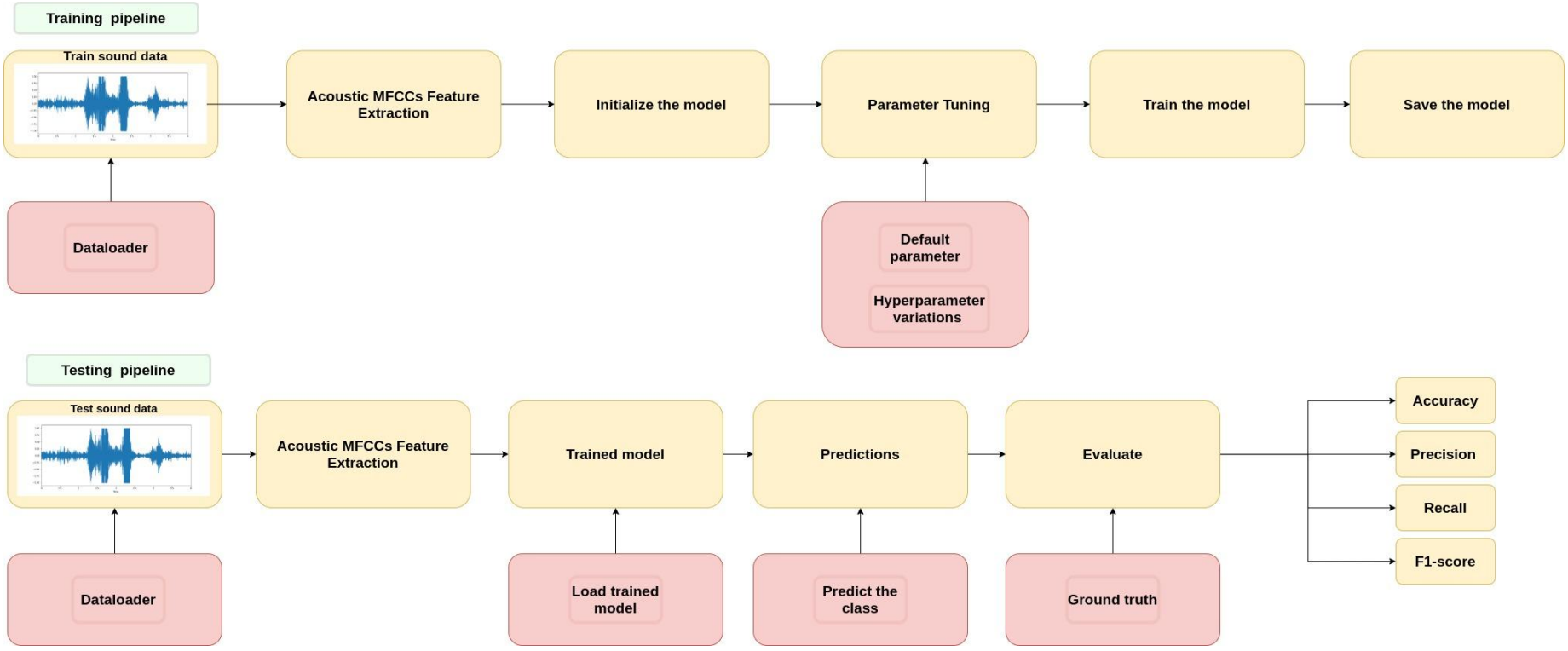


Figure 18: Traditional machine learning training and testing pipeline

Transfer Learning Pipeline

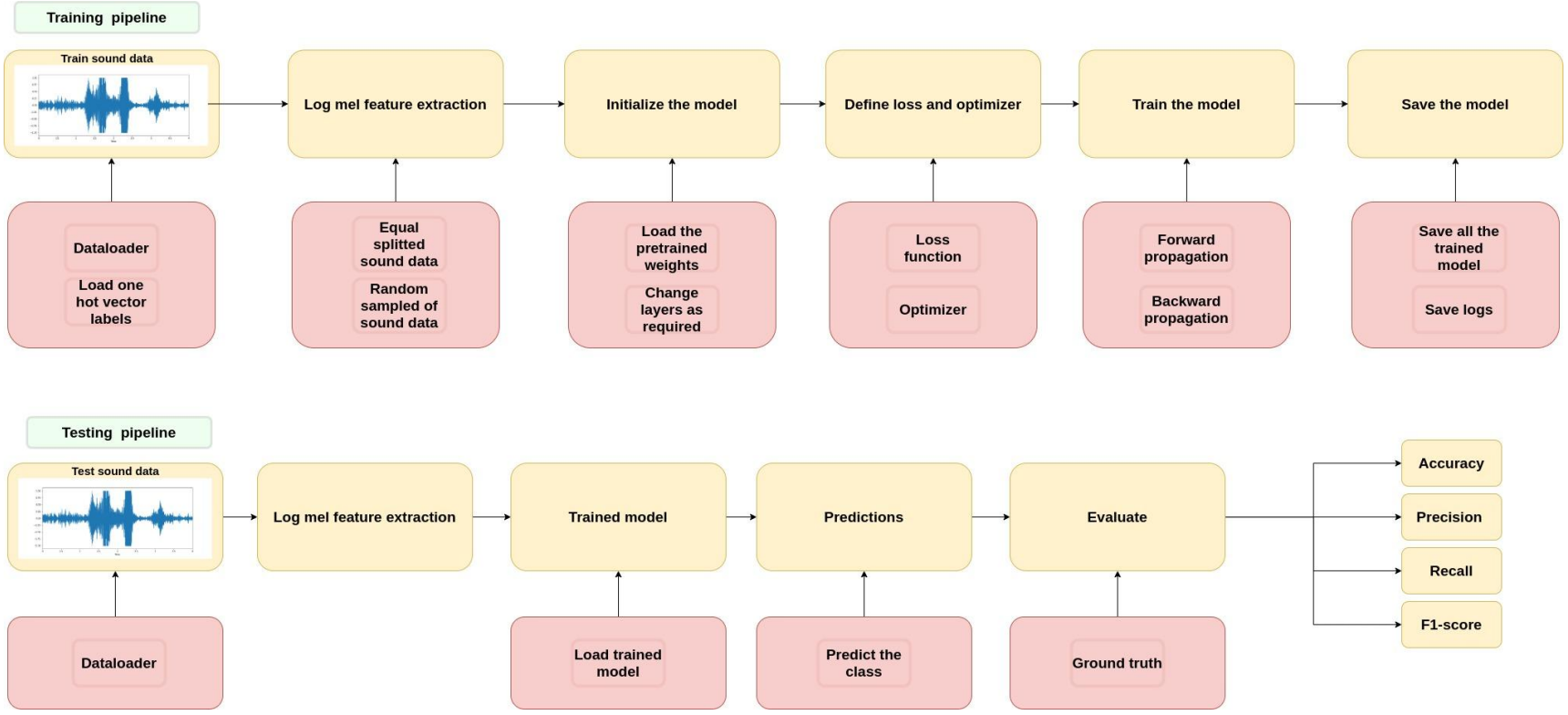


Figure 19: Transfer learning training and testing pipeline

VGGish Architecture

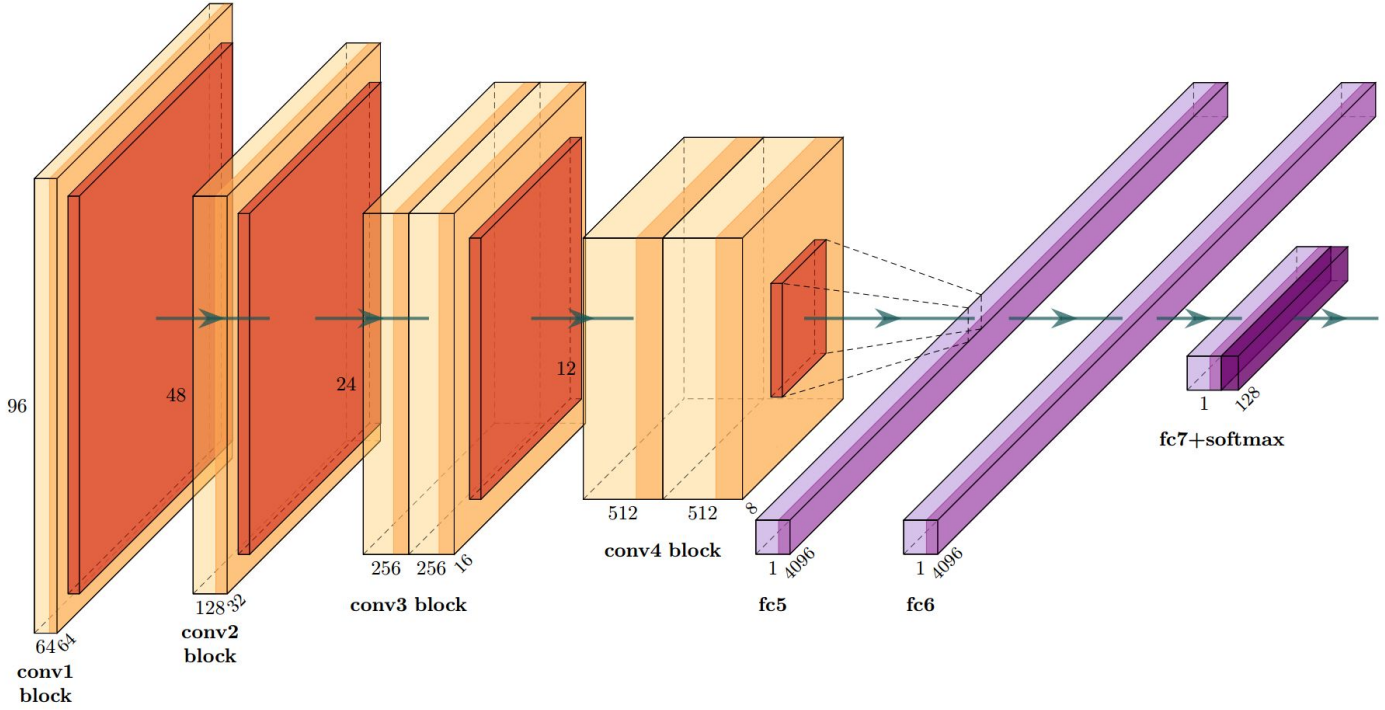


Figure 20: Original VGGish architecture. Generated from [27]

Modified VGGish Architecture

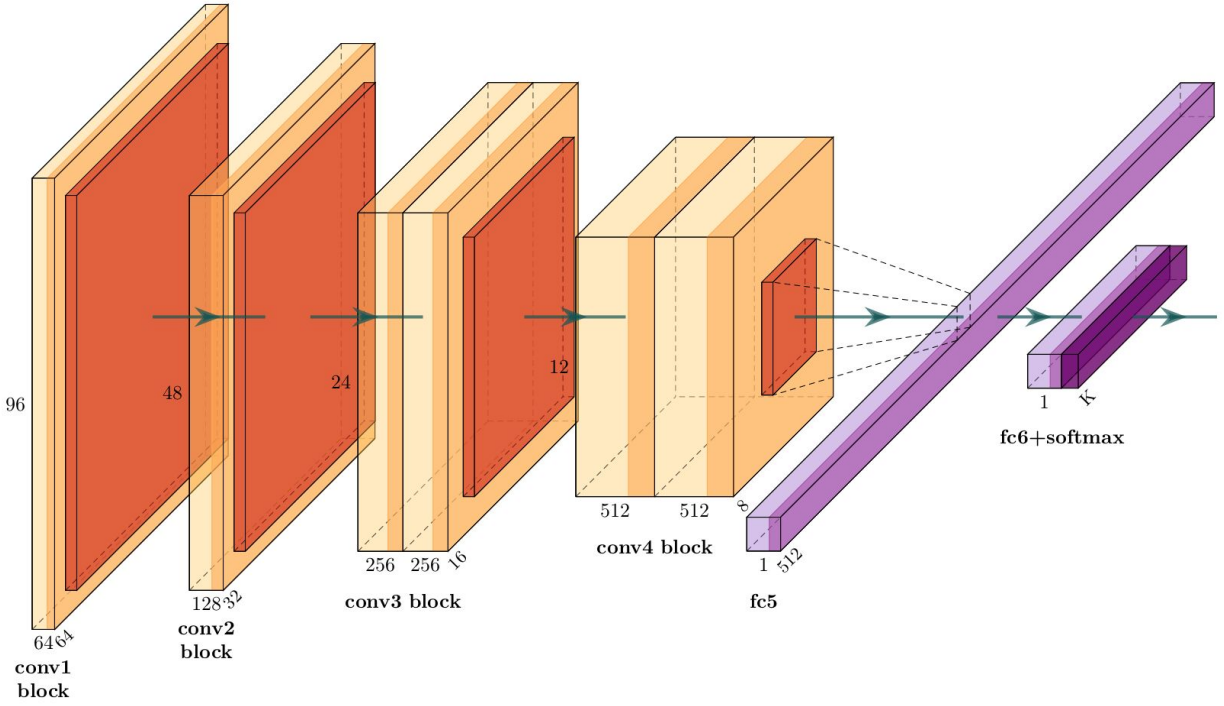


Figure 21: Modified VGGish architecture. Generated from [27]

Evaluation Metrics

Accuracy, Precision, Recall, and F1-Score

- Accuracy is the average across all instances
- Precision is the ratio of correct positives to the total number of positive results predicted by the classifier
- Recall is the ratio of true positive to the total number of all the relevant samples that should be predicted as positive
- F1-Score is the weighted average of recall and precision

Notation

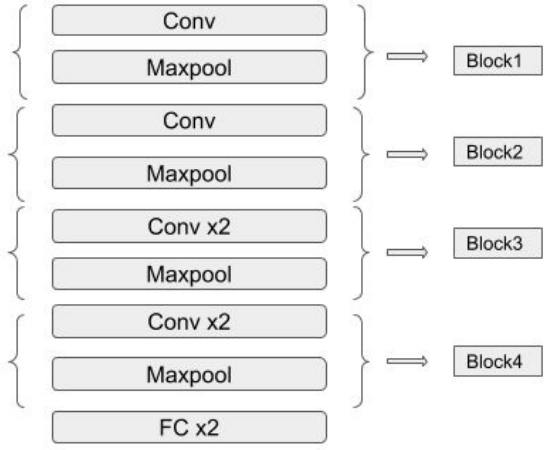


Figure 22: Splitting of VGGish layers to blocks

Classifier	Notation	Remarks
K Nearest Neighbour	KNN	
Support Vector Machine	SVM	
Naive Bayes	NB	
Random Forest	RF	
Gradient Boosting	GB	
XGBoost	XGBoost	
Modified VGGish	VGG01_entire_audio_clip	Frozen block 1
Modified VGGish	VGG02_entire_audio_clip	Frozen block 1, block 2
Modified VGGish	VGG03_entire_audio_clip	Frozen block 1, block 2 and block 3
Modified VGGish	VGG01_random_audio_clip	Frozen block 1
Modified VGGish	VGG02_random_audio_clip	Frozen block 1, block 2
Modified VGGish	VGG03_random_audio_clip	Frozen block 1, block 2 and block 3

Figure 23: Notations used in the results

Results for Urbansound8K Dataset

- Modified VGGish network outperformed traditional machine learning algorithms in classifying the sound data
- Random forest performed well in traditional machine learning category

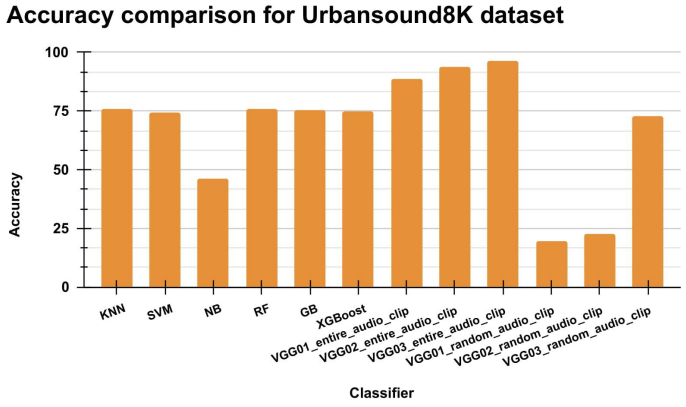


Figure 24: Traditional machine learning and deep learning accuracy comparison

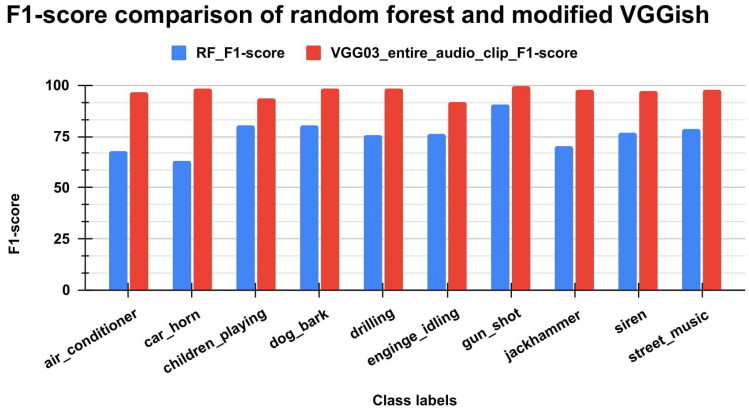


Figure 25: Best performing random forest and modified VGGish F1-score comparison

Results for Urbansound8K Dataset

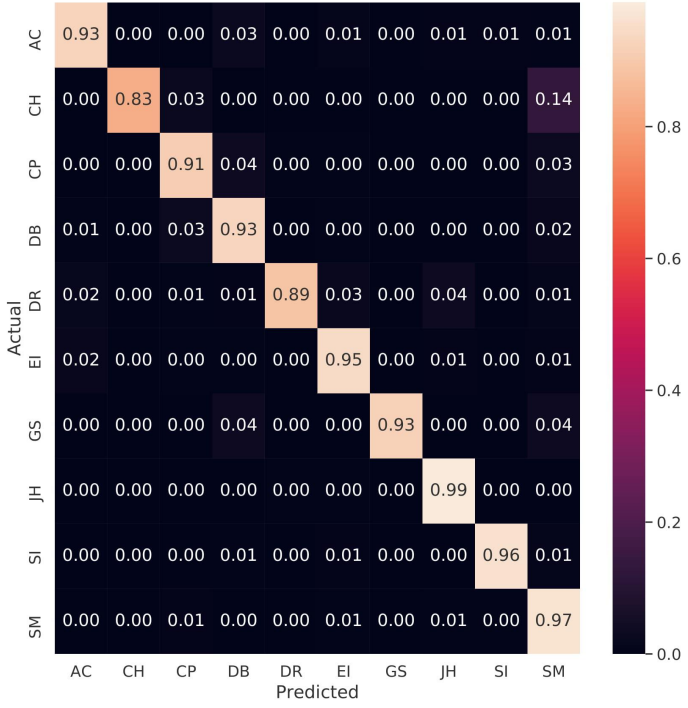


Figure 26: Normalised confusion matrix for VGG03 entire audio clip model in Urbansound8K dataset. [Note:“AC” = Air Conditioner, “CH”= Car Horn, “CP” = Children Playing, “DB” = Dog bark, “DR” = Drilling, “EI” = Engine Idling, “GS” = Gun Shot, “JH” = Jackhammer, “SI” = Siren, “SM” =Street Music]

Results for Urbansound8K Dataset

Model	Feature	Accuracy
Silva [7]	MFCC+7 acoustic features	54.91%
Piczak [8]	LM (Log-Mel)	72.70 %
Tokozume [9]	Raw data	78.30%
Zhang X [10]	Mel spectrogram	81.90%
Zhang Z [11]	LM-GS	83.70%
Li [12]	Raw data -LM	92.20%
Boddapati [13]	Spectrogram -MFCC -CRP	93.00%
LMCNet [14]	LM-C	95.20%
MCNet [15]	M-C	95.30%
Proposed approach	**Log-Mel spectrogram**	**96.56%**
TSCNN-DS [16]	MC and LMC	97.20%

Figure 27: Accuracy comparison with other models on Urbansound8K dataset

Results for DCASE2018 Dataset

- Gradient boosting outperformed the deep learning models in classifying the sound data

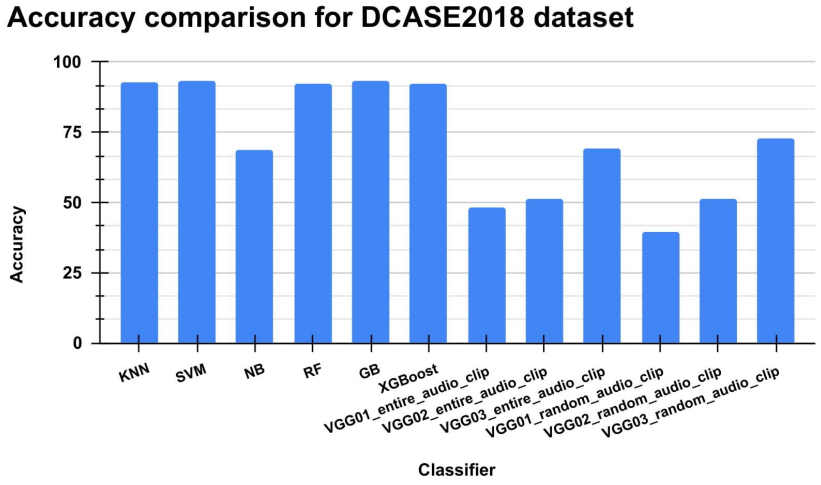


Figure 28: Traditional machine learning and deep learning algorithms accuracy comparison

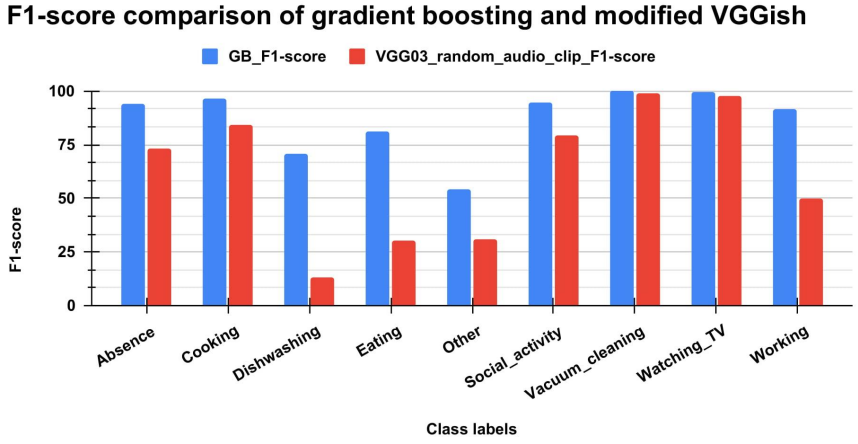


Figure 29: Best performing traditional gradient boosting and modified VGGish F1-score comparison

Results for DCASE2018 Dataset

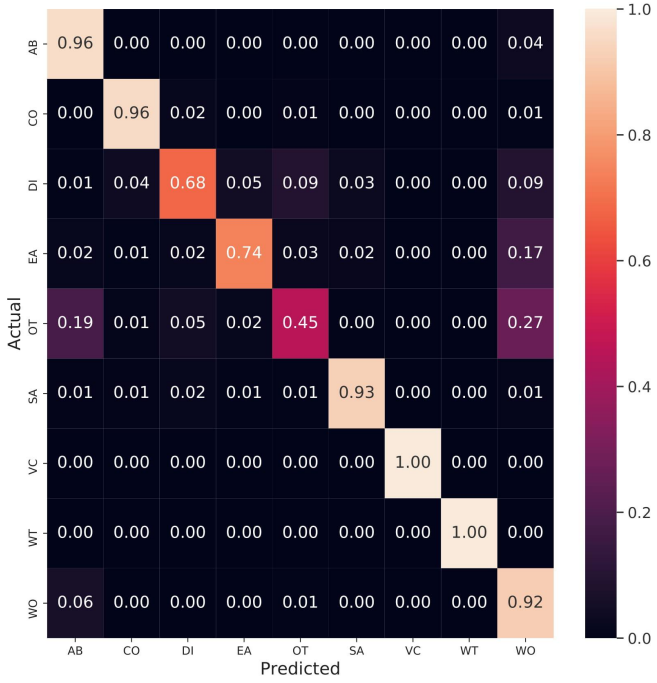


Figure 30: Normalised confusion matrix for GB model in DCASE2018 dataset. [Note: “AB” = Absence, “CO” = Cooking, “DI” = Dishwashing, “EA” = Eating, “OT” = Other, “SA” = Social Activity, “VC” = Vacuum Cleaner, “WT” = Watching TV, “WO” = Working]

Results for DCASE2018 Dataset

Model	Features	Averaged F1-score
Inoue_IBM_task5_1[17]	log-mel energies	88.40 %
Tanabe_HIT_task5_1[18]	log-mel energies +MFCC	88.40 %
Inoue_IBM_task5_2[19]	log-mel energies	88.30 %
Liu_THU_task5_1[20]	log-mel energies +MFCC	87.50 %
Liu_THU_task5_2[21]	log-mel energies +MFCC	87.40 %
Proposed approach -GB	**MFCC**	**86.97 %**
Liu_THU_task5_3[22]	log-mel energies +MFCC	86.80 %
Liao_NTHU_task5_1[23]	log-mel energies	86.70 %
Tanabe_HIT_task5_3[24]	log-mel energies +MFCC	86.30 %
Zhang_THU_task5_3[25]	log-mel energies+Time-Frequency Cepstral	86.00 %

Figure 31: F1-score comparison with other models on DCASE2018 dataset

Contribution

- Analysis of traditional machine learning algorithms performance
- Implementation of six traditional classifiers using MFCC features and analysis of performance
- Evaluating the classifier performance by tuning the hyperparameter
- Execution of log-Mel spectrogram based transfer learning approach
- Analysis of transfer learning performance by freezing different layers of the architecture
- Evaluation of the machine learning method on two types of environmental sound dataset

Future Work

- Data augmentation
- Injection of noise
- Different acoustic features
- Ensemble learning

Learning

- Transfer learning improves the classification performance
- It is important to find events duration in the sound data
- Freezing more layers results in improved performance

References

[1] https://www.alamy.com/friedrichstrasse-street-in-berlin-germany-image63427459.html

[2] Dimitrov, S., Britz, J.,Brandherm, B., Frey, J. (2014, November). Analyzing sounds of home environment for device recognition. In European Conferenceon Ambient Intelligence (pp. 1-16). Springer, Cham.

[3] Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors, 19(7), 1733.

[4] https://www.autoweek.com/news/technology/g30986992/11-of-the-longest-lived-engine-families-you-can-buy-new-today/

[5] https://online.kitp.ucsb.edu/online/hearing17/virtanen2/pdf/Virtanen2_Hearing17_KITP.pdf

[6] Knees, P., & Schedl, M. (2016). Music similarity and retrieval: an introduction to audio-and web-based strategies (Vol. 36). Springer.

References

[7] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In Ismir, volume 270, pages 1–11, 2000

[8] Karol J Piczak. Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6. IEEE, 2015

[9] Yuji Tokozume and Tatsuya Harada. Learning environmental sounds with end-to-end convolutional neural network. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2721–2725. IEEE, 2017.

[10] Xiaohu Zhang, Yuexian Zou, and Wei Shi. Dilated convolution neural network with leakyrelu for environmental sound classification. In 2017 22nd International Conference on Digital Signal Processing (DSP), pages 1–5. IEEE, 2017.

References

[11] Zhichao Zhang, Shugong Xu, Shan Cao, and Shunqing Zhang. Deep convolutional neural network with mixup for environmental sound classification. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pages 356–367. Springer, 2018.

[12] Shaobo Li, Yong Yao, Jie Hu, Guokai Liu, Xuemei Yao, and Jianjun Hu. An ensemble stacked convolutional neural network model for environmental event sound recognition. Applied Sciences, 8 (7):1152, 2018

[13] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. Classifying environmental sounds using image recognition networks. Procedia computer science, 112:2048–2056, 2017

[14] Yu Su, Ke Zhang, Jingyu Wang, and Kurosh Madani. Environment sound classification using a two-stream cnn based on decision-level fusion. Sensors, 19(7):1733, 2019

References

[15] Gert Dekkers. Monitoring of domestic activities based on multi-channel acoustics, 2020. URL http://dcase.community/challenge2018/task-monitoring-domestic-activities#citation. Accessed on: 2020-11-20. [Online]

[16] Yu Su, Ke Zhang, Jingyu Wang, and Kurosh Madani. Environment sound classification using a two-stream cnn based on decision-level fusion. Sensors, 19(7):1733, 2019

[17] Tadanobu Inoue, Phongtharin Vinayavekhin, Shiqiang Wang, David Wood, Nancy Greco, and Ryuki Tachibana. Domestic activities classification based on cnn using shuffling and mixing data augmentation. DCASE 2018 Challenge-Task 5: Monitoring of domestic activities based on multichannel acoustics, Tokyo, Japan, Technical, 2018

[18] Ryo Tanabe, Takashi Endo, Yuki Nikaido, Takeshi Ichige, Phong Nguyen, Yohei Kawaguchi, and Koichi Hamada. Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling. In DCASE 2018 Challenge, 2018

References

[19] Tadanobu Inoue, Phongtharin Vinayavekhin, Shiqiang Wang, David Wood, Nancy Greco, and Ryuki Tachibana. Domestic activities classification based on cnn using shuffling and mixing data augmentation. DCASE 2018 Challenge-Task 5: Monitoring of domestic activities based on multichannel acoustics, Tokyo, Japan, Technical, 2018

[20] Huaping Liu, Feng Wang, Xinzhu Liu, Di Guo, and Fuchun Sun. An ensemble system for Domestic activity recognition. Technical report, DCASE2018 Challenge, Tech. Rep, 2018.

[21] Huaping Liu, Feng Wang, Xinzhu Liu, Di Guo, and Fuchun Sun. An ensemble system for domestic activity recognition. Technical report, DCASE2018 Challenge, Tech. Rep, 2018.

[22] Huaping Liu, Feng Wang, Xinzhu Liu, Di Guo, and Fuchun Sun. An ensemble system for domestic activity recognition. Technical report, DCASE2018 Challenge, Tech. Rep, 2018.

[23] H Liao, J Huang, S Lan, T Lee, Y Liu, and M Bai. Dcase 2018 task 5 challenge technical report: Sound event classification by a deep neural network with attention and minimum variance distortionless response enhancement. Technical report, DCASE2018 Challenge, Tech. Rep, 2018

References

[24] Ryo Tanabe, Takashi Endo, Yuki Nikaido, Takeshi Ichige, Phong Nguyen, Yohei Kawaguchi, and Koichi Hamada. Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling. In DCASE 2018 Challenge, 2018

[25] Yu-Han Shen, Ke-Xin He, and Wei-Qiang Zhang. Home activity monitoring based on gated convolutional neural networks and system fusion. 2018

[26] Salamon, J., Jacoby, C., & Bello, J. P. (2014, November). A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 1041-1044).

[27] https://www.overleaf.com/project/5ff4fb1f2d753d4259fe78ef