Vehicle type recognition based on audio data

Dariusz Kobiela
GUT ETI DSE
dariusz.kobiela@pg.edu.pl

Michał Hajdasz
GUT ETI DMS
s172156@student.pg.edu.pl

Mateusz Erezman
GUT ETI DMS
s171675@student.pg.edu.pl

Karolina Nurzyńska
SUT ACECS DAS
Karolina.Nurzynska@polsl.pl

Szymon Zaporowski
GUT ETI DMS
szyzapor@pg.edu.pl

Adam Kurowski
GUT ETI DMS
adakurow@multimed.org

Paweł Weichbroth
GUT ETI DSE
pawel.weichbroth@pg.edu.pl

GUT - Gdańsk University of Technology, Gabriela Narutowicza 11/12, Gdańsk 80-233, Poland
SUT - Silesian University of Technology, Akademicka 16, Gliwice 44-100, Poland
ACECS - Faculty of Automatic Control, Electronics and Computer Science
ETI - Faculty of Electronics, Telecommunication and Informatics
DAS - Department of Algorithmics and Software
DMS - Department of Multimedia Systems
DSE - Department of Software Engineering

Abstract

Identifying different vehicle types can help manage traffic more efficiently, reduce congestion, and improve public safety. This study aims to create a classification model that can recognize vehicle types based on the sound of passing vehicles. To achieve this, a database of raw audio files containing 1763 samples from two sources was assembled. The time-domain signals were converted to a time-frequency representation using the short-time Fourier transform to generate Mel Spectrograms. Mel-frequency Cepstral Coefficients (MFCCs) were also generated using the discrete cosine transform. In our experiments we compared these approaches. Since the data was imbalanced we applied online augmentation. Based on the literature review, we chose a Convolutional Neural Network (CNN) classifier because it is particularly well suited for analyzing large datasets due to its automatic feature extraction, parameter sharing and sparsity. The results showed that Mel Spectrograms were more effective for audio data preprocessing in this particular use case, achieving the highest accuracy of 0.875 and the highest f1-score of 0.877 compared to MFCCs.

Keywords: vehicle type recognition, vehicle type detection, sound, acoustics, mfcc, mel-frequency cepstral coefficient, spectrogram.

1. Introduction

Vehicle type detection is a crucial technology in the traffic scene today. With its ability to identify different types of vehicles, it can support managing traffic more efficiently, reducing congestion, and improving public safety. So far, there are many approaches investigated for vehicle identification based on different kinds of signals (shown in Kobiela et al., 2024; Suhao et al., 2018; Wu et al., 2020, 2022). The most promising approach for vehicle identification is the one that is based on acoustic signals since moving vehicles emit characteristic sounds. Deep learning techniques are the most used ones (as in H. Chen et al., 2020; Luo et al., 2021; Wieczorkowska et al., 2018). Amongst them, the Convolutional Neural Network (CNN) is the most promising one (used in H. Chen et al., 2020; Dong et al., 2015; Kurowski et al., 2020). The obtained results are compared with the second most used model, the Support Vector Machine (SVM), used in the paper prepared by Czyżewski et al., 2019. This work is the continuation of the aforementioned paper in which we focus on

URI: https://hdl.handle.net/10125/108983
978-0-9981331-8-8

classification of data coming from multiple different sensors. To our knowledge, this is a considerably less explored topic related to acoustic analysis of road traffic. It is however, an important problem because merging already existing datasets is one of viable ways of obtaining datasets large enough to produce high performing acoustic vehicle type classifiers. We would like to address aforementioned research gap by proposing a vehicle type classifier architecture which can be trained on a heterogeneous dataset comprised of audio recordings coming from:

- the data used in Kurowski et al., 2020 which were obtained with a sound probe employing microphones based on microelectromechanical systems (MEMS),
- and data from Bazilinskyy et al., 2018 which was obtained by using a smartphone and a 8-microphone array.

It is also worth mentioning that data from Kurowski et al., 2020 was collected by recording a real-life road traffic. Due to this fact, the data contain class imbalance. In this paper, we propose a simple oversampling-based approach which could help to mitigate this imbalance. Knowledge on how to mitigate such a problem is important especially for researchers and engineers who intend to design robust machine learning algorithms which are employing continual, reinforcement and other machine learning paradigms in which input training data tend to be imbalanced. Examples of such algorithms are intrusion detection algorithm for computer networks employing adversarial reinforcement learning (as in Ma and Shi, 2021), object detection in very high resolution remote sensing images (as in X. Chen et al., 2023), or underwater acoustic target classification (as in Pala et al., 2023).

2. Background and related work

Today, as a global society, we recognize the problems of an ever-changing environment and global warming (Radziszewski et al., 2021). The emission of greenhouse gases, mainly caused by the use of private automobiles for transportation, is one of the major causes of global warming (Romero et al., 2024). In this view, vehicle recognition is an important and timely research topic for at least two reasons.

First, it is critical to emissions enforcement, helping to identify and penalize high-polluting vehicles, thereby reducing their contribution to global warming (A. A. Ahmed et al., 2023). It helps law enforcement identify vehicles that violate environmental regulations, contributing to the global effort to combat climate change (L. Chen et al., 2023). Vehicle recognition helps optimize traffic flow by directing different types of vehicles to appropriate routes, minimizing overall emissions in urban areas (Huang et al., 2020). The data collected by vehicle recognition systems can inform policy decisions aimed at reducing the environmental impact of transportation on a global scale.

Second, in light of the Smart Cities research agenda (Mora et al., 2023), by recognizing vehicle types, cities can implement differentiated tolls and taxes that encourage the use of low-emission vehicles, thereby supporting sustainable development goals (Palomares et al., 2021). By identifying vehicle types, cities can better plan and develop infrastructure that supports sustainable transportation, such as dedicated lanes for electric vehicles and bicycles (C. Chen et al., 2020). Such technologies can identify non-compliant vehicles in low emission zones, ensuring that only clean vehicles contribute to urban traffic. Besides, the ability to accurately identify vehicles enables better traffic management in smart cities, reducing congestion and associated carbon emissions (L. Zhang et al., 2020). In addition, vehicle recognition can improve the efficiency of public transportation systems by prioritizing buses and electric vehicles, contributing to cleaner air and more sustainable urban mobility (Ceder, 2021). For example, recent advances in computer vision technology are enabling dynamic pricing for parking based on vehicle type, encouraging the use of smaller, more fuel-efficient cars and reducing the carbon footprint of cities.

Having said that, the current literature review aimed to learn about models used to recognize vehicles by their sound. Furthermore, it was desired to see the advantages and disadvantages of those models in the selected field of interest and also find out what results should be expected from the used implementation. It was found that the most common models used in vehicle type recognition based on sound are CNN, SVM, DNN (Deep Neural Networks) and LSTM (Long short-term memory networks), as shown in the Table 1.

Furthermore, these models are often combined to create more advanced network models and achieve better results. Through the Systematic Literature Review (SLR), it was also discovered that extracting the sound features that go into the network input is as necessary or even more important than the model used. Some articles focused on developing a complex network, while others focused on the best features to extract from recorded vehicle sound. The last twenty articles used for data extraction are: (Abdul Rahim et al., 2011; A. Ahmed et al., 2021; Anwar et al., 2019;

Becker et al., 2020; H. Chen et al., 2020; Jakubowski & Jackowski, 2021; Kurowski et al., 2019, 2020; S. Lee et al., 2017; Li et al., 2020; Luo et al., 2021; Montino & Pau, 2019; Scarpiniti et al., 2021; Sunu et al., 2018; Vij & Aggarwal, 2020; L. Wang & Roggen, 2019; Q. Wang et al., 2021; Wieczorkowska et al., 2018; Wu et al., 2020, 2022).

Note that the detailed report of the performed literature review can be found in the attached GitHub repository.

Table 1: Identified vehicle recognition models and the number of occurrences of their use in research.

Model	Count
Convolutional Neural Network (CNN)	9
Support Vector Machine (SVM)	7
Deep Neural Network (DNN)	7
Long-Short Term Memory (LSTM)	4
Recurrent Neural Network (RNN)	2
k-Nearest Neighbours (*k*-NN)	2
Gated Recurrent Unit (GRU)	1
Probabilistic Neural Network (PNN)	1
Siamese Neural Network (SNN)	1
Hybrid: S-CRNN (CNN + RNN)	1
Hybrid: LSTM+CNN	1
Random Forest	1
Decision tree	1
Naive Bayes	1

3. Input Data

Input data for the project is the raw audio file with the sound of the approaching vehicle. The dataset consisted of 1763 samples from two sources gathered by Kurowski et al., 2020 and by Bazilinskyy et al., 2018. The dataset was partially already tagged and partially self-tagged by the authors of this article. Raw audio files were cut into fixed 6 seconds length recordings. Most of the sample length were about 4 seconds, as shown in Figure 1, so the additional time was filled with silence up to 6 seconds for every sample. In the case of audio samples longer than 6 seconds, the centre of every audio sound was found, and then +/– 3 seconds of the audio sound center were taken. After data preprocessing, all the source videos were saved in the WAV (Waveform Audio Format) shown in Figure 2a, 2b, 2c. WAV is an audio file format that offers lossless audio quality. Additionally, the video in AVI (Audio Video Interleave) format of the vehicles was provided (shown in Figure 2d, 2e, 2f). AVI is a commonly used file format developed by Microsoft for storing video data. The video data was used to tag the audio data into one of the three classes: car, motorcycle or truck. According to the pilot study, the ”truck” class contained vehicles such as trucks, buses, and vans.

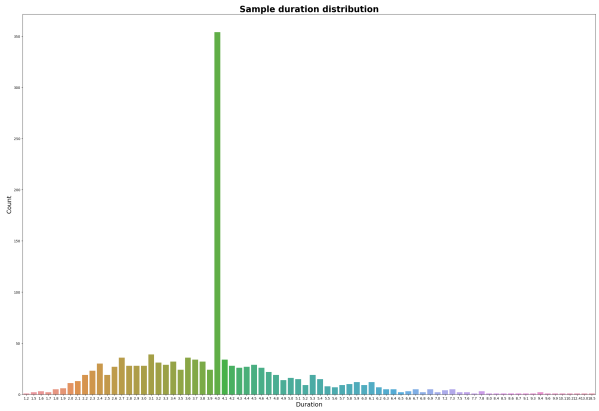
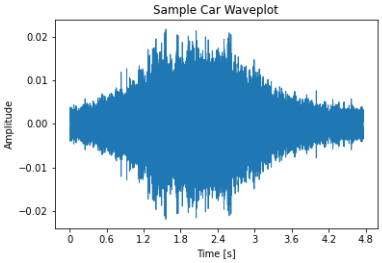


Figure 1: Histogram of samples duration distribution

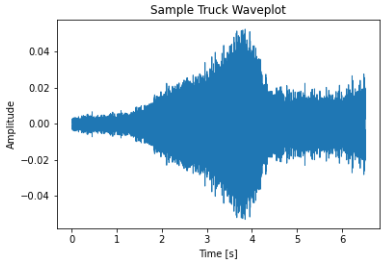
To convert signals from the time domain into a time-frequency representation, there were used the Short-Time Fourier Transform (STFT), with the aim of obtaining mel-spectrograms that served as a model input (as shown in Figure 2j, 2k, 2l). Short-time Fourier transform is a sequence of Fourier transforms of a windowed signal. STFT provides the time-localized frequency information for situations in which frequency components of a signal vary over time, whereas the standard Fourier transform provides the frequency information averaged over the entire signal time interval (Kehtarnavaz, 2008). Second data transformation was made using Discrete Cosine Transform (DCT) in order to create Mel-frequency cepstral coefficients (MFCCs, shown in the Figures 2g, 2h, 2i). Both representation types (mel-spectrograms and MFCCs) were used as model input. The main problem with the data was the imbalanced number of learning examples for every class - 1174 cars, 467 trucks and 122 motorcycles. It was decided to use online data augmentation techniques to upsample fewer classes (truck and motorcycle). Augmentation consisted of time shift: random shift of the raw input audio by a selected time from a range between 0 and 2 seconds. Change was made randomly back or forth. As a result, the fixed 6 seconds audio window contained more silence at the beginning or end of the window. Augmentation also consisted of masking out mel-spectrograms. Vertical black bars masked the selected time of the record, while horizontal black bars masked selected frequencies.

4. Research design and pilot study

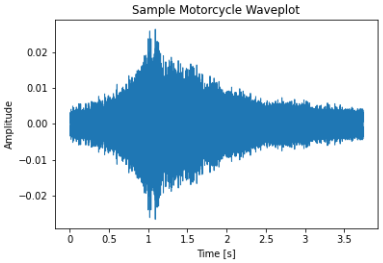
Pilot study experiments were executed on a small subset of data. Experiments were conducted sequentially, one after another, starting from choosing the model type. The best approach from the previous



(a) sample car waveplot



(b) sample truck waveplot

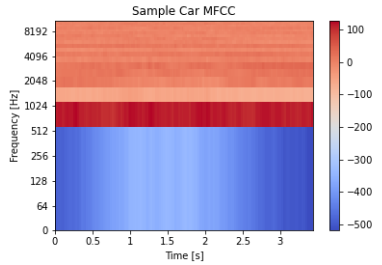


(c) sample motorcycle waveplot

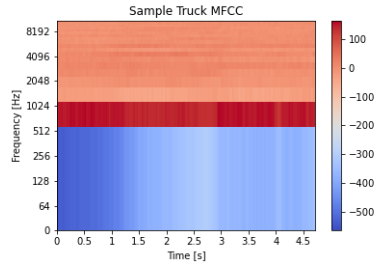
(d) sample car (screenshot from video)

(e) sample truck (screenshot from video)

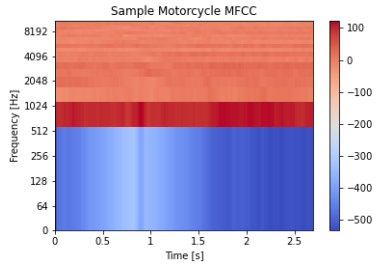
(f) sample motorcycle (screenshot from video)



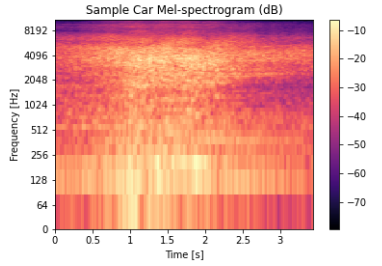
(g) sample car MFCC



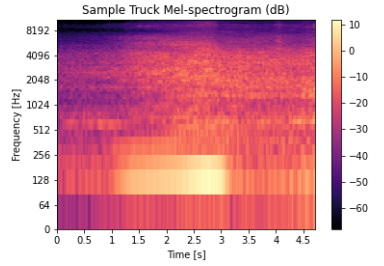
(h) sample truck MFCC



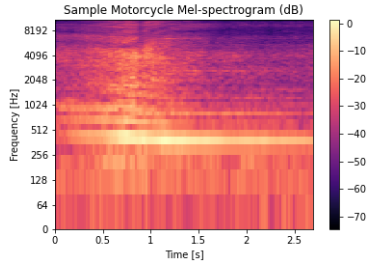
(i) sample motorcycle MFCC



(j) sample car mel-spectrogram



(k) sample truck mel-spectrogram



(l) sample motorcycle mel-spectrogram

Figure 2: Sample car, track and motorcycle data processing: from audio and movie tagging to MFCC and mel-spectrograms

phase was used for the next experiment. Each training was executed ten times to ensure the excellent reliability of the experiments. Results in each epoch were the mean value of 10 executions. During all experiments, the number of learning epochs was set to 500 with an early stopping parameter equal to 25 epochs (according to validation loss). The research questions were:

(a) What is the most optimal architecture type? Choose from Dense, 2x Dense, 3x Dense, Convolutional, 2xConvolution, 3xConvolutional, Dense+Convolutional.

(b) What is the optimal number of kernels (units) in each layer? Choose from 10, 30, 50, 80, 100.

(c) What is the best optimizer? Choose from RMSprop, Adadelta, Adagrad, Adam, SGD, Ftrl, Nadam, and Adamax.

(d) What is the best activation function? Choose from tanh, sigmoid, ReLu, softmax.

The best hyperparameters found after experiments are Adam and RMSprop optimizer (which allows for reduced training length), ReLu activation function, and 4xCNN architecture with 100 kernels in each layer. It took, on average, 150 epochs to train the model. A detailed report about research design and pilot study can be found in the attached GitHub source code repository. Final experiments were performed on multiple variants of data: with and without online data augmentation, with different numbers of hop_length parameter, with the usage of MFCCs and mel-spectrograms. Models were implemented using PyTorch and Tensorflow frameworks to eliminate the framework's influence and compare the results achieved. The dataset was divided into train and test sets using an 80:20 proportion.

5. Model

According to the SLR and pilot study results, the model of choice was CNN. It was implemented in both PyTorch and Tensorflow frameworks. CNN is a Deep Learning algorithm that can take in an input image (in this case - mel-spectrogram or MFCC), assign importance (learnable weights and biases) to various aspects or objects in the image, and be able to differentiate one from the other. The preprocessing required in a ConvNet is much lower than other classification algorithms. While in primitive methods, filters are hand-engineered, with enough training, ConvNets can learn these filters or characteristics (Saha, n.d.). The chosen architecture of the model can be seen in Table 2. It consists of 25,147 trainable parameters (also total parameters), and the estimated total size of the architecture is 2.92 MB. Output data from the model is the class of the given audio file. Models return one out of three classes, where 0 indicates car, 1 indicates truck, and 2 indicates motorcycle. Details about the model architecture can be found in the source code repository.

Table 2: CNN model architecture

Layer (type)	Output Shape	Parameters
Conv2d-1	[-1, 8, 32, 235]	408
ReLU-2	[-1, 8, 32, 235]	0
BatchNorm2d-3	[-1, 8, 32, 235]	16
Conv2d-4	[-1, 16, 16, 118]	1,168
ReLU-5	[-1, 16, 16, 118]	0
BatchNorm2d-6	[-1, 16, 16, 118]	32
Conv2d-7	[-1, 32, 8, 59]	4,640
ReLU-8	[-1, 32, 8, 59]	0
BatchNorm2d-9	[-1, 32, 8, 59]	64
Conv2d-10	[-1, 64, 4, 30]	18,496
ReLU-11	[-1, 64, 4, 30]	0
BatchNorm2d-12	[-1, 64, 4, 30]	128
AdaptiveAvgPool2d-13	[-1, 64, 1, 1]	0
Linear-14	[-1, 3]	195

6. Results

Results obtained by the models are as follows. Table 3, Figures 3 and 4 shows the results for the model with the usage of mel-spectrograms (the highest values of metrics), while Table 4 and Figure 5 shows the results for model with the usage of MFCCs. Table 5 shows the compared results obtained by Czyżewski et al., 2019 (SVM model).

Table 3: Best model metrics (usage of mel-spectrograms)

Vehicle	Precision	Recall	F1-score
car	0.94	0.89	0.91
truck	0.75	0.86	0.80
motorcycle	0.70	0.78	0.74
macro avg	0.80	0.84	0.82
weighted avg	0.88	0.88	**0.88**

Table 4: Second model metrics (usage of MFCCs)

Vehicle	Precision	Recall	F1-score
car	0.89	0.88	0.88
truck	0.66	0.71	0.69
motorcycle	1.00	0.71	0.83
macro avg	0.85	0.77	0.80
weighted avg	0.83	0.83	**0.83**

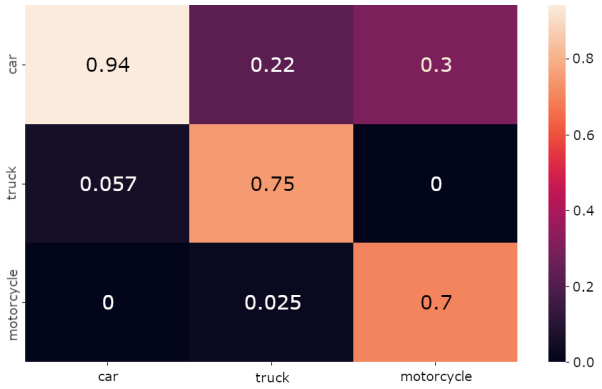


Figure 3: Confusion matrix for the best model (usage of mel-spectrograms)

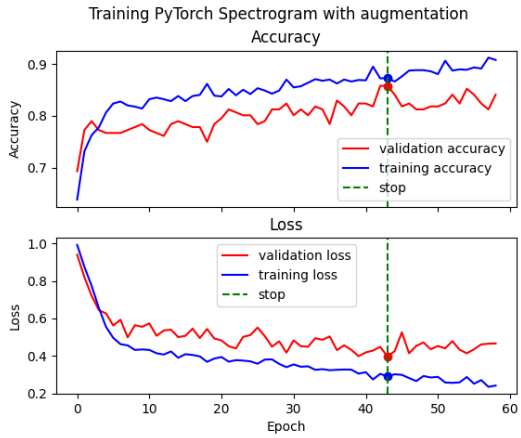


Figure 4: Best model accuracy and loss during the training of the model (usage of mel-spectrograms)

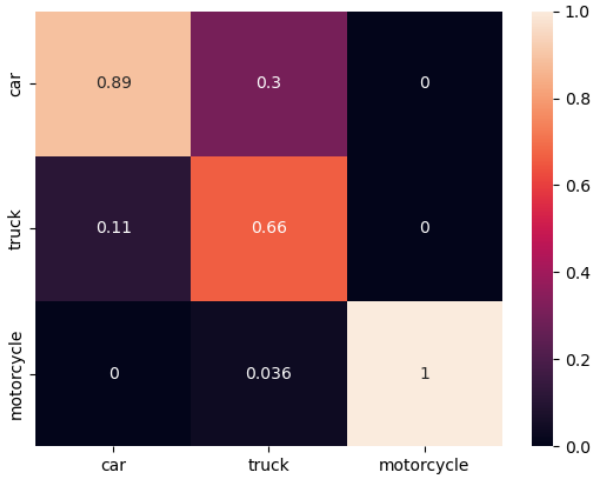


Figure 5: Confusion matrix for the model with the usage of MFCCs

Table 5: SVM model results (obtained by Czyżewski et al., 2019)

Vehicle	Precision	Recall	F1-score
car	0.706	0.703	0.704
truck	0.136	0.375	0.200
macro avg	0.421	0.539	**0.452**

7. Discussion

The Tensorflow framework was more effortless in the implementation, more user-friendly thanks to Keras API. Results achieved using both frameworks were pretty similar, the difference was neglectible. The results showed the superiority of mel-spectograms over MFCCs in the given use case (see Tables 3 and 4). The highest accuracy equal to **0.875**, along with the highest f1-score equal to **0.887** (see Table 3), was achieved by the model implemented in PyTorch with the usage of mel-spectrograms and with the augmented data, although the results obtained by the model without augmentation were not much lower. As it was decided to use time shift and masking out with horizontal and vertical black bars (time and frequency) as augmentation techniques, we think that better up-sampling methods are needed. The confusion matrix can provide additional information about the results (see Figure 3). The model did pretty well in classifying cars (almost no mistakes, 94% correct answers), while it sometimes mixed tracks with cars - 75% correct classifications (which is pretty predictable because tracks, vans and busses were put into one class). A similar problem occurred with motorcycles (70% correct classifications) - here, it may come with too few learning examples (the collection of cars was much more extensive than the collection of motorcycles). The learning of the model went as predicted during the pilot study (see Figure 4). The model was trained until the validation loss increased ($\geq$ 15 epochs). Then, the early stopping worked, and the saved best model was taken (from the 43rd training epoch). The comparison of data preprocessing methods showed that the usage of mel-spectrograms gives better results (see Table 3) in the data modelling phase in terms of accuracy and f1-score than the usage of MFCCs. A described model using MFCC metrics has an accuracy of **0.83** and f1-score equal to **0.83** (see Table 4). Obtained results are also shown in the confusion matrix (see Figure 5). Because of the small support of the class "motorcycle", the confusion matrix (100% correct classifications) results are unreliable. In comparison to the SVM model used by Czyżewski et al., 2019, which obtained macro average accuracy at the level of **0.929**

and macro average F1-score at the level of **0.452**, the tested solution (CNN architecture) is only 5% lower in terms of accuracy, but about 43% higher in terms of F1-score. We hypothesize that mel-spectrograms provided better performance due to the fact that they do not focus on periodic structures present in the analyzed sound spectrum unlike the cepstrum-based MFCCs. It is likely that while such focus on signal spectrum periodicity is beneficial for tasks such as speech signal processing, it may be at the same time detrimental for classification of acoustic spectra produced by passing vehicles.

The approach shown in this paper is an example of a relatively simple way of dealing with the imbalance present in the training data. Hence, it is likely that despite of a relatively large size of the input dataset (comprised of 1763 samples), the imbalance still has a considerable effect on classification accuracy of less represented classes. It is a limitation of our study. However, one should remember that in applications such as monitoring of real-live road traffic the problem of data imbalance is often hard to avoid. Therefore, in the future we would like to investigate more sophisticated ways of addressing the input data imbalance and improving the overall performance of the proposed architecture. Firstly, we would like to train the machine learning model architecture presented in this paper on extended version of the dataset described in this study. This extended dataset which would contain additional examples for currently underrepresented classes. Secondly, it is possible to use more advanced oversampling techniques such as the synthetic minority oversampling technique (SMOTE) which is described in detail in e.g. Chawla et al., 2002. The SMOTE algorithm was successfully employed for tasks such as classification of imbalanced hyperspectral images (as in Özdemir et al., 2021) or developing a deep learning model for voice pathology detection (as in J.-N. Lee and Lee, 2023). Thirdly, it is also possible to use different types of advanced data augmentation such as augmentation employing generative adversarial models (GANs) or diffusion models. Technique employing GAN-based augmentation was used in addressing the task of material characteristics classification with the use of imbalanced spectral data (as in Chung et al., 2024). Approach based on diffusion models was successfully employed for data augmentation for lung ultrasound images classification (as in X. Zhang et al., 2023). Finally, there is also a potential to improve the input features by using self-attention mechanisms. Thanks to such treatment, the model would be capable of focusing only on relevant the part of audio frames which may further improve the vehicle type classification accuracy. This approach was successfully used in tasks of sound event detection such as one described in Miyazaki et al., 2020.

Source code, SLR reports, and data sources used

Source code for this paper and SLR results are available at: github.com/DariuszKobiela/vehicle-type-recognition-based-on-audio-data.
Collected and tagged data sources used for this paper are available at: www.kaggle.com/datasets/brinkor/vehicle-type-sound-dataset.

References

Abdul Rahim, N., P, P. M., Adom, A. H., & Kumar, S. S. (2011). Moving vehicle noise classification using multiple classifiers. *2011 IEEE Student Conference on Research and Development*, 105–110. https://doi.org/10.1109/SCOReD.2011.6148717

Ahmed, A., Serrestou, Y., Raoof, K., & Diouris, J.-F. (2021). Sound event classification using neural networks and feature selection based methods. *2021 IEEE International Conference on Electro Information Technology (EIT)*, 1–6. https://doi.org/10.1109/EIT51626.2021.9491869

Ahmed, A. A., Nazzal, M. A., Darras, B. M., & Deiab, I. (2023). Global warming potential, water footprint, and energy demand of shared autonomous electric vehicles incorporating circular economy practices. *Sustainable Production and Consumption*, *36*, 449–462. https://doi.org/10.1016/j.spc.2023.02.001

Anwar, M. Z., Kaleem, Z., & Jamalipour, A. (2019). Machine learning inspired sound-based amateur drone detection for public safety applications. *IEEE Transactions on Vehicular Technology*, *68*(3), 2526–2534. https://doi.org/10.1109/TVT.2019.2893615

Bazilinskyy, P., van der Aa, A., Schoustra, M., Spruit, J., Staats, L., van der Vlist, K., & de Winter, J. (2018). An auditory dataset of passing vehicles recorded with a smartphone. In I. Horváth & J. Suárez (Eds.), *Proceedings of the 12th international symposium on tools and methods of competitive engineering (tmce 2018)* (pp. 417–422). https://doi.org/10.4121/uuid:bef54ab8-73ef-42f3-b6b7-54e011737e72

Becker, L., Nelus, A., Gauer, J., Rudolph, L., & Martin, R. (2020). Audio feature extraction for vehicle engine noise classification. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 711–715. https://doi.org/10.1109/ICASSP40776.2020.9053117

Ceder, A. (2021). Urban mobility and public transport: Future perspectives and review. *International Journal of Urban Sciences*, *25*(4), 455–479. https://doi.org/10.1080/12265934.2020.1799846

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *16*(1), 321–357. https://doi.org/10.5555/1622407.1622416

Chen, C., Zhang, Y., Khosravi, M. R., Pei, Q., & Wan, S. (2020). An intelligent platooning algorithm for sustainable transportation systems in smart cities. *IEEE Sensors Journal*, *21*(14), 15437–15447. https://doi.org/10.1109/JSEN.2020.3019443

Chen, H., Zhang, Z., Yin, W., Wang, M., Lifan, M., & Hao, X. (2020). Hybrid neural network based on feature fusion for vehicle type identification. *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 1–5. https://doi.org/10.1109/I2MTC43012.2020.9129183

Chen, L., Chen, Z., Zhang, Y., Liu, Y., Osman, A. I., Farghali, M., Hua, J., Al-Fatesh, A., Ihara, I., Rooney, D. W., et al. (2023). Artificial intelligence-based solutions for climate change: A review. *Environmental Chemistry Letters*, *21*(5), 2525–2557. https://doi.org/10.1007/s10311-023-01617-y

Chen, X., Jiang, J., Li, Z., Qi, H., Li, Q., Liu, J., Zheng, L., Liu, M., & Deng, Y. (2023). An online continual object detector on vhr remote sensing images with class imbalance. *Engineering Applications of Artificial Intelligence*, *117*, 105549. https://doi.org/10.1016/j.engappai.2022.105549

Chung, J., Zhang, J., Saimon, A. I., Liu, Y., Johnson, B. N., & Kong, Z. (2024). Imbalanced spectral data analysis using data augmentation based on the generative adversarial network. *Scientific Reports*, *14*(1), 13230. https://doi.org/10.1038/s41598-024-63285-4

Czyżewski, A., Kurowski, A., & Zaporowski, S. (2019). Application of autoencoder to traffic noise analysis. *Journal of the Acoustical Society of America*, *146*, 2958–2958.

Dong, Z., Wu, Y., Pei, M., & Jia, Y. (2015). Vehicle type classification using a semisupervised convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 2247–2256. https://doi.org/10.1109/TITS.2015.2402438

Huang, Y.-Q., Zheng, J.-C., Sun, S.-D., Yang, C.-F., & Liu, J. (2020). Optimized yolov3 algorithm and its application in traffic flow detections. *Applied Sciences*, *10*(9), 3079. https://doi.org/10.3390/app10093079

Jakubowski, J., & Jackowski, J. (2021). Recognition of moving tracked and wheeled vehicles based on sound analysis and machine learning algorithms. *International Journal of Automotive and Mechanical Engineering*, *18*(1), 8478–. https://doi.org/10.15282/ijame.18.1.2021.07.0642

Kehtarnavaz, N. (2008). Chapter 7 - frequency domain processing. In N. Kehtarnavaz (Ed.), *Digital signal processing system design (second edition)* (Second Edition, pp. 175–196). Academic Press. https://doi.org/10.1016/B978-0-12-374490-6.00007-6

Kobiela, D., Groth, J., Hajdasz, M., & Erezman, M. (2024). Vehicle type recognition: A case study of mobilenetv2 for an image classification task. *Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 28th International Conference KES2024*.

Kurowski, A., Czyżewski, A., & Zaporowski, S. (2019). Automatic labeling of traffic sound recordings using autoencoder-derived features. *SPA 2019 SIGNAL PROCESSING algorithms, architectures, arrangements, and applications Conference Proceedings Poznan, 18th - 20th September 2019*, 38–43. https://doi.org/10.23919/SPA.2019.8936709

Kurowski, A., Zaporowski, S., & Czyzewski, A. (2020). 1d convolutional context-aware architectures for acoustic sensing and recognition of passing vehicle type. *2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 142–145. https://doi.org/10.23919/SPA50552.2020.9241256

Lee, J.-N., & Lee, J.-Y. (2023). An efficient smote-based deep learning model for voice pathology detection. *Applied Sciences*, *13*(6). https://doi.org/10.3390/app13063571

Lee, S., Lee, J., & Lee, K. (2017). Vehiclesense: A reliable sound-based transportation mode

recognition system for smartphones. *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 1–9. https://doi.org/10.1109/WoWMoM.2017.7974318

Li, R., Yin, B., Cui, Y., Du, Z., & Li, K. (2020). Research on environmental sound classification algorithm based on multi-feature fusion. *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, *9*, 522–526. https://doi.org/10.1109/ITAIC49862.2020.9338926

Luo, Y., Chen, L., Wu, Q., & Zhang, X. (2021). Sound-convolutional recurrent neural networks for vehicle classification based on vehicle acoustic signals. *2021 International Conference on Smart City and Green Energy (ICSCGE)*, 98–102. https://doi.org/10.1109/ICSCGE53744.2021.9654357

Ma, X., & Shi, W. (2021). Aesmote: Adversarial reinforcement learning with smote for anomaly detection. *IEEE Transactions on Network Science and Engineering*, *8*(2), 943–956. https://doi.org/10.1109/TNSE.2020.3004312

Miyazaki, K., Komatsu, T., Hayashi, T., Watanabe, S., Toda, T., & Takeda, K. (2020). Weakly-supervised sound event detection with self-attention. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 66–70. https://doi.org/10.1109/ICASSP40776.2020.9053609

Montino, P., & Pau, D. (2019). Environmental intelligence for embedded real-time traffic sound classification. *2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI)*, 45–50. https://doi.org/10.1109/RTSI.2019.8895517

Mora, L., Gerli, P., Ardito, L., & Petruzzelli, A. M. (2023). Smart city governance from an innovation management perspective: Theoretical framing, review of current practices, and future research agenda. *Technovation*, *123*, 102717. https://doi.org/10.1016/j.technovation.2023.102717

Özdemir, A., Polat, K., & Alhudhaif, A. (2021). Classification of imbalanced hyperspectral images using smote-based deep learning methods. *Expert Systems with Applications*, *178*, 114986. https://doi.org/10.1016/j.eswa.2021.114986

Pala, A., Oleynik, A., Utseth, I., & Handegard, N. O. (2023). Addressing class imbalance in deep learning for acoustic target classification. *ICES Journal of Marine Science*, *80*(10), 2530–2544. https://doi.org/10.1093/icesjms/fsad165

Palomares, I., Martínez-Cámara, E., Montes, R., García-Moral, P., Chiachio, M., Chiachio, J., Alonso, S., Melero, F. J., Molina, D., Fernández, B., et al. (2021). A panoramic view and swot analysis of artificial intelligence for achieving the sustainable development goals by 2030: Progress and prospects. *Applied Intelligence*, *51*, 6497–6527. https://doi.org/10.1007/s10489-021-02264-y

Radziszewski, K., Anacka, H., & Weichbroth, P. (2021). Greencoin: A proenvironmental action-reward system.

Romero, C. A., Correa, P., Ariza Echeverri, E. A., & Vergara, D. (2024). Strategies for reducing automobile fuel consumption. *Applied Sciences*, *14*(2), 910. https://doi.org/10.3390/app14020910

Saha, S. (n.d.). *A comprehensive guide to convolutional neural networks*. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (accessed: 04.09.2024).

Scarpiniti, M., Comminiello, D., Uncini, A., & Lee, Y.-C. (2021). Deep recurrent neural networks for audio classification in construction sites. *2020 28th European Signal Processing Conference (EUSIPCO)*, 810–814. https://doi.org/10.23919/Eusipco47968.2020.9287802

Suhao, L., Jinzhao, L., Guoquan, L., Tong, B., Huiqian, W., & Yu, P. (2018). Vehicle type detection based on deep learning in traffic scene [Recent Advancement in Information and Communication Technology:]. *Procedia Computer Science*, *131*, 564–572. https://doi.org/10.1016/j.procs.2018.04.281

Sunu, J., Percus, A. G., & Hunter, B. (2018). Unsupervised vehicle recognition using incremental reseeding of acoustic signatures. In M. Ceci, N. Japkowicz, J. Liu, G. A. Papadopoulos, & Z. W. Raś (Eds.), *Foundations of intelligent systems* (pp. 151–160). Springer International Publishing. https://doi.org/10.1007/978-3-030-01851-1_15

Vij, D., & Aggarwal, N. (2020). Transportation mode detection using cumulative acoustic sensing

and analysis. *Front. Comput. Sci. 15, 151311*. https://doi.org/10.1007/s11704-019-9200-3

Wang, L., & Roggen, D. (2019). Sound-based transportation mode recognition with smartphones. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 930–934. https://doi.org/10.1109/ICASSP.2019.8682917

Wang, Q., He, Y., Chen, Z., & Luo, Y. (2021). Vehicle identification based on improved 1/3 octave and bark-scale wavelet packet methods. *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, 1256–1260. https://doi.org/10.1109/ICET51757.2021.9450949

Wieczorkowska, A., Kubera, E., Słowik, T., & Skrzypiec, K. (2018). Spectral features for audio based vehicle and engine classification. *J Intell Inf Syst 50*, 265–290. https://doi.org/10.1007/s10844-017-0459-2

Wu, J.-D., Luo, W.-J., & Yao, K.-C. (2022). Acoustic signal classification using symmetrized dot pattern and convolutional neural network. *Machines*, *10*(2). https://doi.org/10.3390/machines10020090

Wu, J.-D., Wong, Y.-H., Luo, W.-J., & Yao, K.-C. (2020). Acoustic emission signal classification using feature analysis and deep learning neural network. *Fluctuation and Noise Letters*, *20*, 2150030. https://doi.org/10.1142/S0219477521500309

Zhang, L., Long, R., Li, W., & Wei, J. (2020). Potential for reducing carbon emissions from urban traffic based on the carbon emission satisfaction: Case study in shanghai. *Journal of Transport Geography*, *85*, 102733. https://doi.org/10.1016/j.jtrangeo.2020.102733

Zhang, X., Gangopadhyay, A., Chang, H.-M., & Soni, R. (2023, October). Diffusion model-based data augmentation for lung ultrasound classification with limited data. In S. Hegselmann, A. Parziale, D. Shanmugam, S. Tang, M. N. Asiedu, S. Chang, T. Hartvigsen, & H. Singh (Eds.), *Proceedings of the 3rd machine learning for health symposium* (pp. 664–676, Vol. 225). PMLR. https://proceedings.mlr.press/v225/zhang23a.html