Scientific Research Publishing

Journal of Computer and Communications, 2023, 11, 19-27
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Comparative Analysis of Different Sampling Rates on Environmental Sound Classification Using the Urbansound8k Dataset

Ibrahim Aljubayri

Department of Computer Science and Information, Taibah University, Madinah, Saudi Arabia
Email: ijubayri@taibahu.edu.sa

How to cite this paper: Aljubayri, I. (2023) Comparative Analysis of Different Sampling Rates on Environmental Sound Classification Using the Urbansound8k Dataset. *Journal of Computer and Communications*, **11**, 19-27.
https://doi.org/10.4236/jcc.2023.116002

Received: November 20, 2022
Accepted: June 16, 2023
Published: June 19, 2023

Abstract

Environmental sound classification (ESC) has gained increasing attention in recent years. This study focuses on the evaluation of the popular public dataset Urbansound8k (Us8k) at different sampling rates using hand crafted features. The Us8k dataset contains environment sounds recorded at various sampling rates, and previous ESC works have uniformly resampled the dataset. Some previous work converted this data to different sampling rates for various reasons. Some of them chose to convert the rest of the dataset to 44,100, as the majority of the Us8k files were already at that sampling rate. On the other hand, some researchers down sampled the dataset to 8000, as it reduced computational complexity, while others resampled it to 16,000, aiming to achieve a balance between higher classification accuracy and lower computational complexity. In this research, we assessed the performance of ESC tasks using sampling rates of 8000 Hz, 16,000 Hz, and 44,100 Hz by extracting the hand crafted features Mel frequency cepstral coefficient (MFCC), gamma tone cepstral coefficients (GTCC), and Mel Spectrogram (MelSpec). The results indicated that there was no significant difference in the classification accuracy among the three tested sampling rates.

Keywords

Deep Learning, Convolutional Neural Network, Environmental Sound Classification

1. Introduction

Automatic sound recognition has gained considerable momentum recently and has been deployed in diverse fields such as audio surveillance systems [1], wild-

life area impostor detection [2], ESC [3], and noise reduction [4]. Environmental sound encompasses various non-musical noises in our daily lives, including glass breaking, door knocking, flowing water, and engine sounds. Our brain continuously processes and interprets these acoustic data to provide information about the surrounding environment, whether consciously or subconsciously. The main purpose of ESC is to identify the nature of specific sounds by classifying them into various events. ESC is a burgeoning research field with numerous practical applications. Several studies on worker safety have implemented ESC to detect noise levels and prevent hearing loss and excessive loudness. Nowadays, ESC technology is becoming increasingly popular. Multiple related works have utilized the Us8k dataset to evaluate their proposed ESC models. For instance, [5] proposed a new technique for dilated convolution and achieved 78% accuracy. [6] introduced a novel deep convolutional neural network (DCNN) model with an average accuracy of 86.7%. Similarly, [7] proposed a new convolutional network with an accuracy of 86%, and [8] proposed a 1-D CNN with an accuracy rate of 89%. The Us8k dataset comprises audio recordings with different sampling rates ranging from 8000 to 190,000. The majority of the files (8499) have sampling rates between 44,100 and 190,000. **Figure 1** illustrates the sampling rate distribution of the Us8k dataset. Many related studies employ various resampling techniques on the Us8k dataset during the pre-processing stage to standardize it to a single sampling rate. Some of these studies claim that adopting a specific sampling rate can improve the accuracy of the tested models. Some studies, such as [6] [7] [8], resampled the Us8k dataset to 8000 Hz, while others, like [9] [10] [11], standardized the sampling rates to 16,000 Hz. Similarly, [12] [13] [14] standardized the sampling rates to 44,100 Hz. The aim of this paper is to evaluate the appropriate sampling rates for the Us8k dataset to improve performance in ESC tasks.

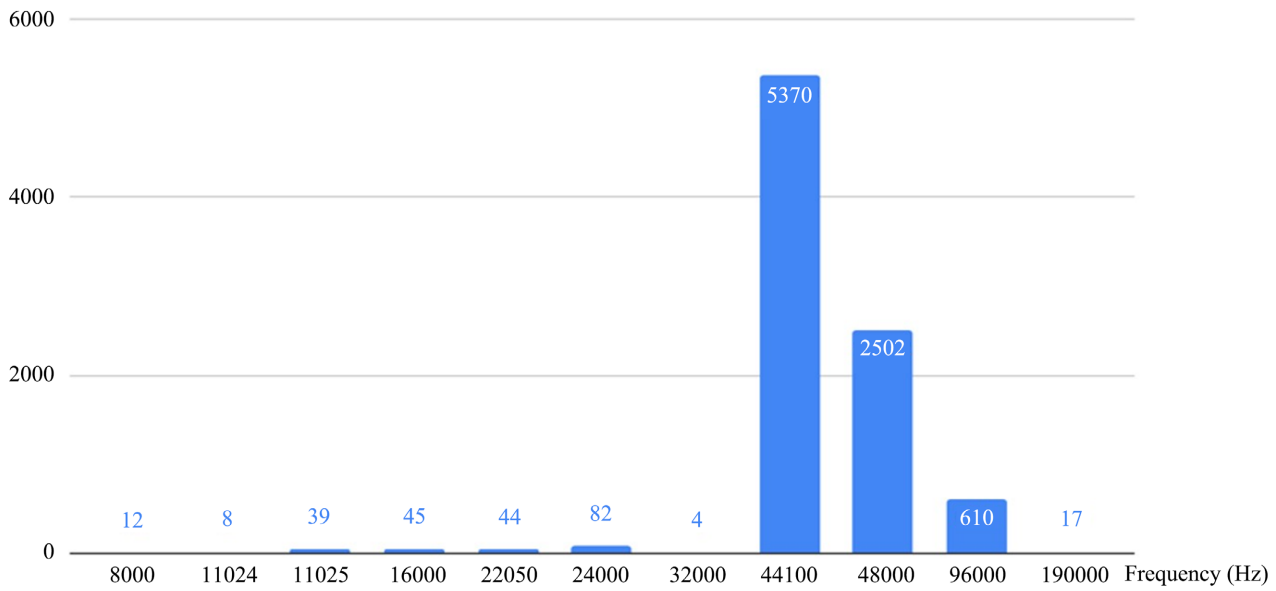

Figure 1. Sampling rate distribution for Us8k.

2. Experimental Datasets and Setup Description

The hardware platform utilized in this study consisted of an AMD Ryzen 9 3900× 12-Core Processor (3.80 GHz), NVIDIA GeForce RTX 2070 SUPER, and 64.0 GB of RAM. MATLAB 2023a was employed for model development and testing. The Us8k dataset comprises 8732 annotated audio files, each with a duration of 4 seconds or less, categorized into 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music [15]. These classes were randomly assigned to 10 folds and cross validation technique was used to evaluate this work. The total estimated duration of all audio clips is about 8.75 hours. **Figure 2** illustrates the distribution of the Us8k dataset. In this work, we resampled the Us8k dataset to three different sampling rates: 8000 Hz, 16,000 Hz, and 44,100 Hz. From each resampled version, we extracted the handcrafted features MFCC, GTCC, and MelSpec from the waveform of each audio file. The classification task employed the k-nearest neighbors algorithm (kNN).

3. Extracted Features

Mel Frequency Cepstral Coefficient (MFCC):

MFCC is a widely used feature in sound processing and speech recognition, capturing the spectral characteristics of an audio signal by representing variations in the Mel frequency scale. The computation of MFCC involves several steps. Firstly, a pre-emphasis high-pass filter is applied to enhance higher frequencies in the signal. Next, the signal is divided into frames of equal duration, typically around 20 - 40 milliseconds, through frame blocking. Each frame is windowed by multiplying it with the Hamming window function to minimize

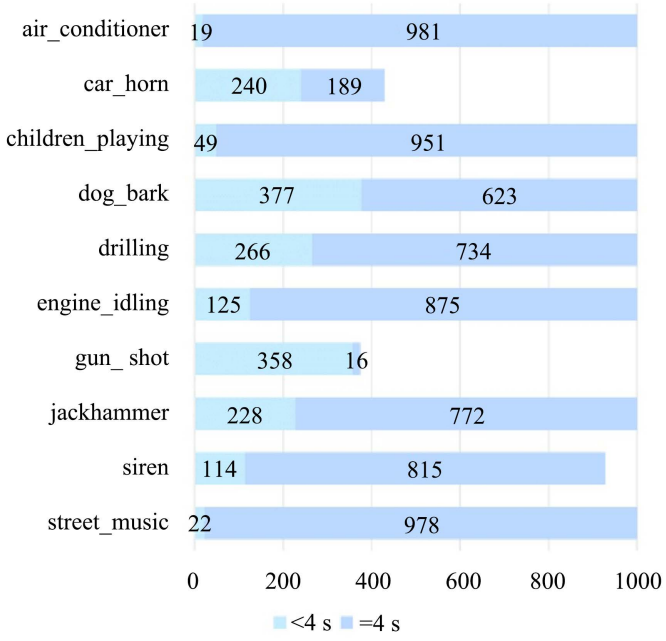

Figure 2. Us8k dataset classes and length distribution. [5]

spectral leakage. The power spectrum of each frame is obtained using the Fast Fourier Transform (FFT). Subsequently, the power spectrum is subjected to a set of triangular filters uniformly spaced on the Mel scale, known as the Mel Filterbank, and the outputs from these filters are summed within each filterbank. To compress the dynamic range, the logarithm of the filterbank outputs is calculated. Finally, the Discrete Cosine Transform (DCT) is applied to the log-filterbank energies, resulting in the extraction of compact MFCC coefficients that represent the spectral envelope. Figure 3 illustrates the Mel Filter Bank.

Gammatone Cepstral Coefficients (GTCC):

GTCC is another sound analysis feature inspired by the frequency analysis of the human auditory system. It relies on the gammatone filterbank, which emulates the filtering properties of the basilar membrane in the cochlea. Similar to MFCC, GTCC follows a computation process involving multiple steps. However, instead of using the Mel filterbank, it employs a bank of gammatone filters designed to mimic the human auditory system's response to different frequencies. Figure 4 illustrates the Gammatone Filter Bank.

Mel Spectrogram (MelSpec):

The Mel spectrogram is a visual representation of the magnitude spectrum of an audio signal in the Mel frequency domain. It is computed by dividing the audio signal into short overlapping frames and applying the FFT to each frame.

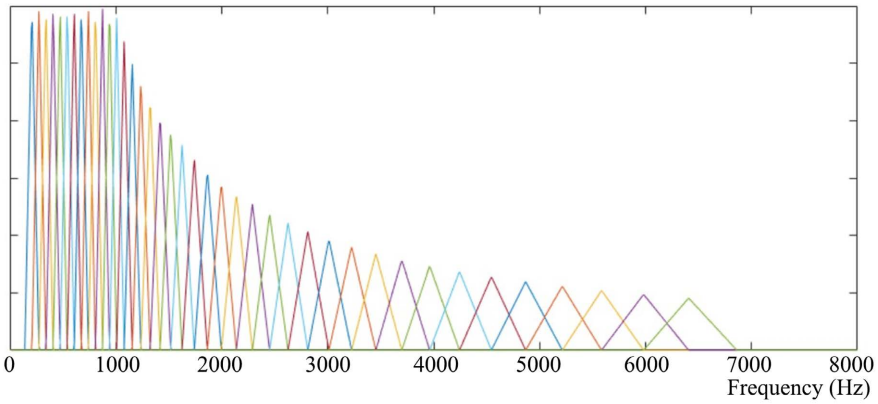

Figure 3. Mel filter bank. [16]

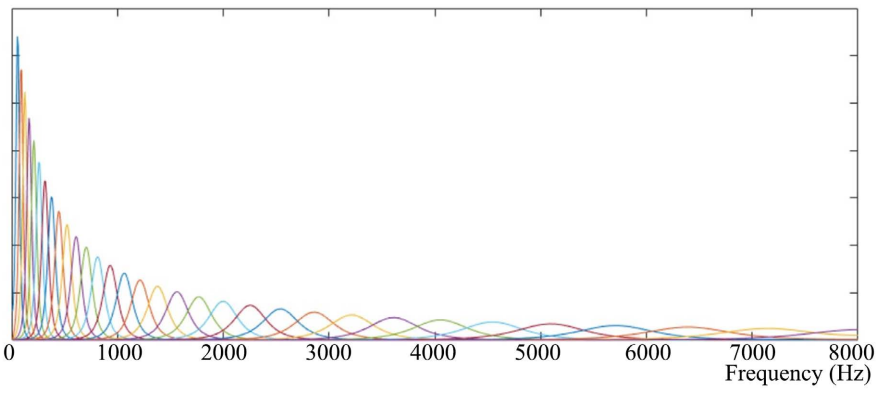

Figure 4. Gammatone filter bank. [16]

The resulting power spectrum is then transformed into the Mel scale using a Mel filter bank, similar to MFCC. The Mel spectrogram offers a detailed analysis of the audio signal's frequency content over time, enabling the extraction of frequency-based features.

4. Experimental Results

Table 1 presents a comparison between the classification accuracies for the different sampling rates (8000 Hz, 16,000 Hz, and 44,100 Hz) and various features (MFCC, GTCC, MelSpec, MFCC + GTCC, and MFCC + GTCC + MelSpec). For the 8000 Hz sampling rate, the highest accuracy is achieved with the combination of MFCC and GTCC, reaching 94.1%. The individual features MFCC, GTCC, and MelSpec achieve accuracies of 93.3%, 88.5%, and 85.6% respectively. For the 16,000 Hz sampling rate, the highest accuracy is obtained with the combination of MFCC and GTCC, reaching 94.4%. The individual features MFCC, GTCC, and MelSpec achieve accuracies of 93.6%, 90.4%, and 86.1% respectively. For the 44,100 Hz sampling rate, the highest accuracy is achieved with the combination of MFCC and GTCC + MelSpec, reaching 94.4%. The individual features MFCC, GTCC, and MelSpec achieve accuracies of 93.1%, 90.7%, and 85.5% respectively. Figure 5 illustrates the confusion matrix of the MFCC and

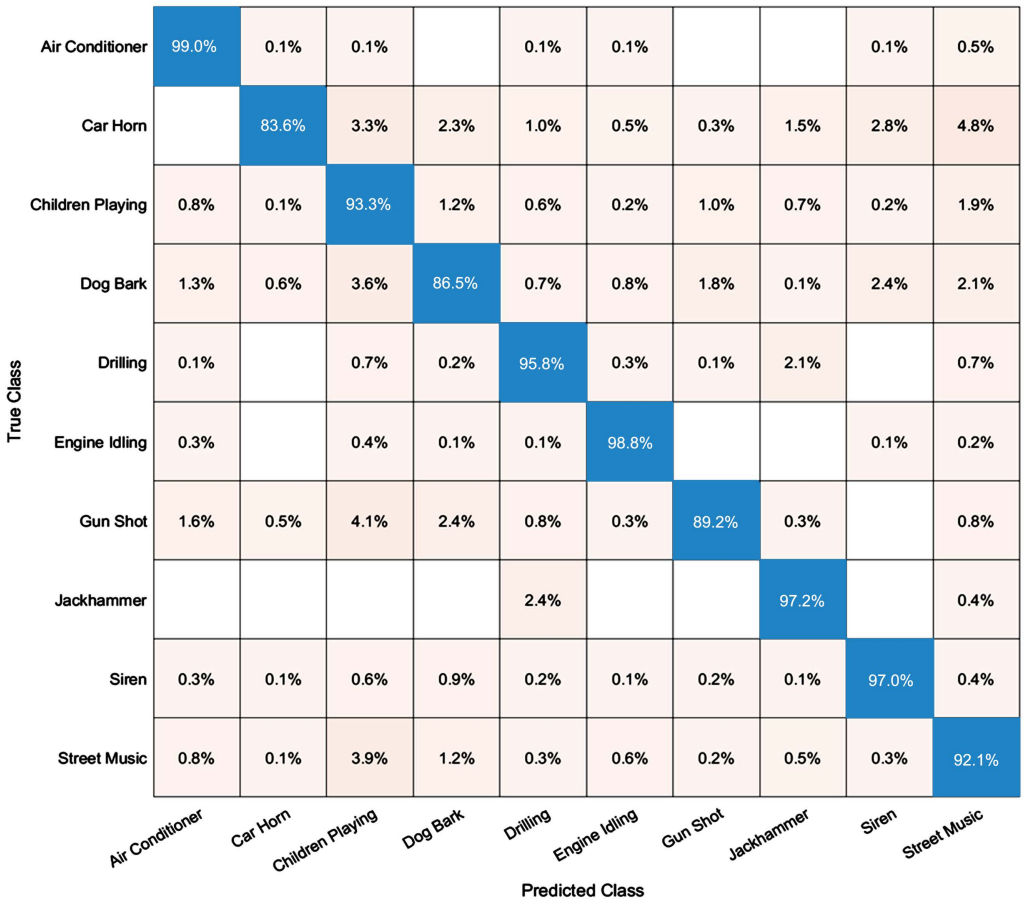

Figure 5. Confusion matrix for MFCC and GTCC with 8000 Hz sampling rate.

GTCC classification results using the 8000 Hz sampling rate. Figure 6 illustrates the confusion matrix of the MFCC and GTCC classification results using the 16,000 Hz sampling rate. Figure 7 illustrates the confusion matrix of the MFCC, GTCC, and MelSpec classification results using the 44,100 Hz sampling rate. Based on the results, there is no significant difference in classification accuracy among the three tested sampling rates. The combination of MFCC and GTCC consistently shows high accuracy across all sampling rates.

Table 1. Sampling rate result comparison.

Features \ Sampling rate	8000 Hz	16,000 Hz	44,100 Hz	# of features
MFCC	93.3	**93.6**	93.1	13
GTCC	88.5	90.4	**90.7**	13
MelSpec	85.6	**86.1**	85.5	32
MFCC + GTCC	94.1	**94.4**	94.2	26
MFCC + GTCC + MelSpec	94.0	94.2	**94.4**	58

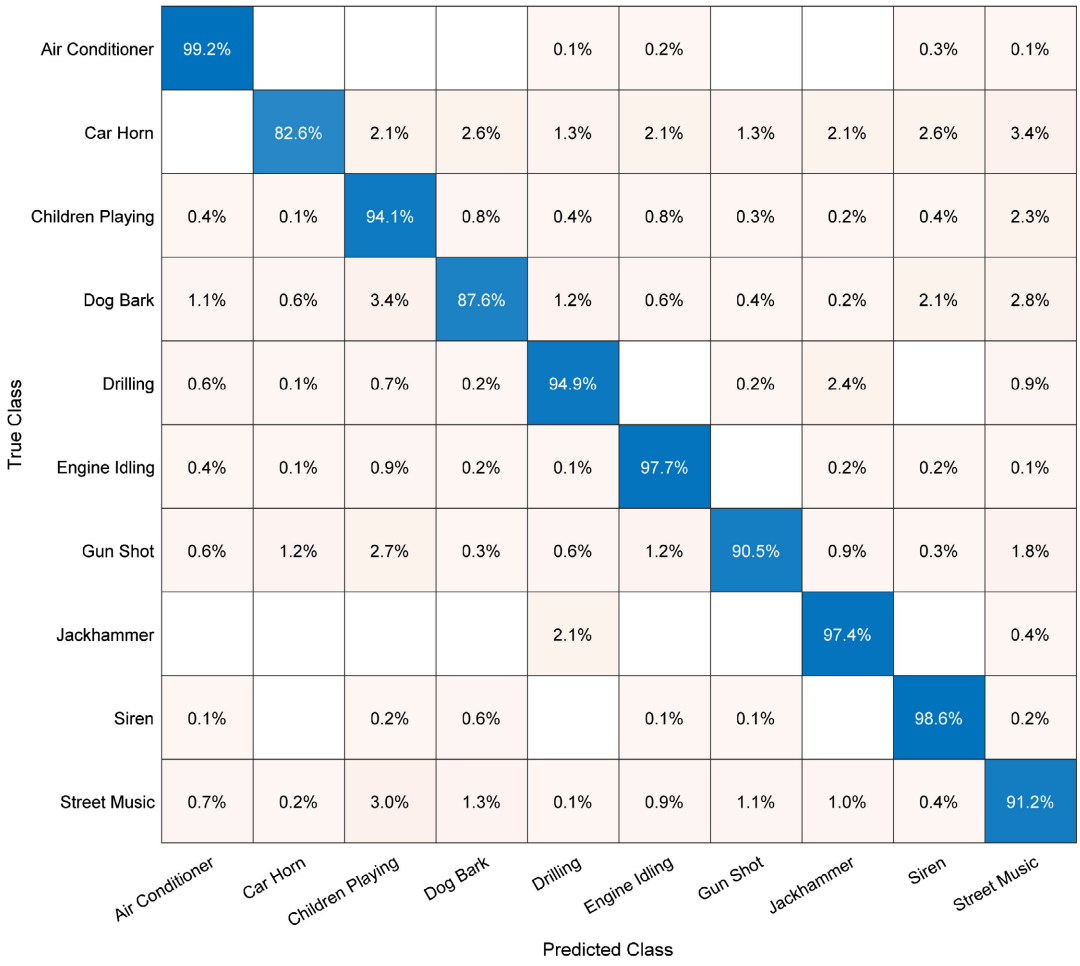

Figure 6. Confusion matrix for MFCC and GTCC with 16,000.

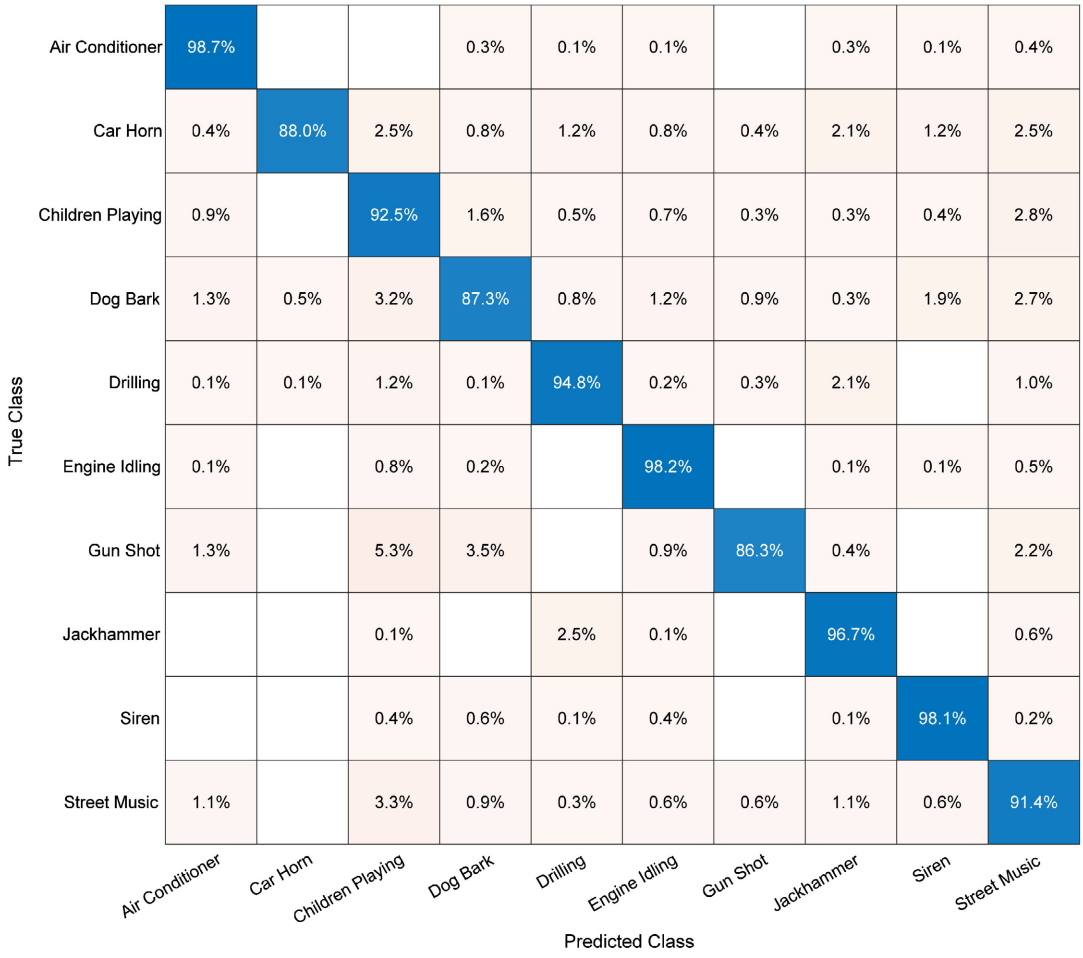

Figure 7. Confusion matrix for MFCC, GTCC, and MelSpec with 44,100 Hz.

5. Conclusion

In this work, we investigated the impact of different sampling rates on the performance of ESC tasks. We focused on the popular public dataset Us8k and evaluated its performance at three different sampling rates: 8000 Hz, 16,000 Hz, and 44,100 Hz. The following Handcrafted features, Mel frequency cepstral coefficient (MFCC), gamma tone cepstral coefficients (GTCC), and Mel spectrogram (MelSpec), were extracted from the audio files and used to train and test the model using the kNN classification algorithm. Our experimental results showed that there was no significant difference in the classification accuracy among the three tested sampling rates. The ESC performance using the 8000 Hz sampling rate experienced a slight decrease compared to the 16,000 Hz and 44,100 Hz sampling rates. However, these differences were not substantial enough to conclude a clear advantage of one sampling rate over the others. The findings indicate that the choice of sampling rate does not significantly impact the performance of ESC tasks when utilizing the Us8k dataset and the handcrafted features employed in this study. Therefore, researchers can adopt any of the tested sampling rates based on their specific requirements and computational constraints.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Rabaoui, A., Davy, M., Rossignol, S. and Ellouze, N. (2008) Using One-Class SVMs and Wavelets for Audio Surveillance. *IEEE Transactions on Information Forensics and Security*, **3**, 763-775. https://doi.org/10.1109/TIFS.2008.2008216

[2] Ghiurcau, M.V., Rusu, C., Bilcu, R.C. and Astola, J. (2012) Audio Based Solutions for Detecting Intruders in Wild Areas. *Signal Processing*, **92**, 829-840. https://doi.org/10.1016/j.sigpro.2011.10.001

[3] Wang, J.-C., Lee, H.-P., Wang, J.-F. and Lin, C.-B. (2008) Robust Environmental Sound Recognition for Home Automation. *IEEE Transactions on Automation Science and Engineering*, **5**, 25-31. https://doi.org/10.1109/TASE.2007.911680

[4] Mydlarz, C., Salamon, J. and Bello, J.P. (2017) The Implementation of Low-Cost Urban Acoustic Monitoring Devices. *Applied Acoustics*, **117**, 207-218. https://doi.org/10.1016/j.apacoust.2016.06.010

[5] Dong, X., Yin, B., Cong, Y., Du, Z. and Huang, X. (2020) Environment Sound Event Classification with a Two-Stream Convolutional Neural Network. *IEEE Access*, **8**, 125714-125721. https://doi.org/10.1109/ACCESS.2020.3007906

[6] Qu, S., Li, J., Dai, W. and Das, S. (2016) Understanding Audio Pattern Using Convolutional Neural Network from Raw Waveforms. ArXiv Preprint ArXiv: 1611.09524.

[7] Sang, J., Park, S. and Lee, J. (2018) Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms. 2018 26*th European Signal Processing Conference* (*EUSIPCO*), Rome, 3-7 September 2018, 2444-2448. https://doi.org/10.23919/EUSIPCO.2018.8553247

[8] Esmaeilpour, M., Cardinal, P. and Koerich, A.L. (2020) From Sound Representation to Model Robustness. ArXiv Preprint ArXiv: 2007.13703.

[9] Abdoli, S., Cardinal, P. and Koerich, A.L. (2019) End-to-End Environmental Sound Classification Using a 1D Convolutional Neural Network. *Expert Systems with Applications*, **136**, 252-263. https://doi.org/10.1016/j.eswa.2019.06.040

[10] Lu, X., Shen, P., Li, S., Tsao, Y. and Kawai, H. (2019) Deep Progressive Multi-Scale Attention for Acoustic Event Classification. ArXiv Preprint ArXiv: 1912.12011.

[11] Ting, P.-J., Ruan, S.-J. and Li, L.P.-H. (2021) Environmental Noise Classification with Inception-Dense Blocks for Hearing Aids. *Sensors*, **21**, Article No. 5406. https://doi.org/10.3390/s21165406

[12] Guzhov, A., Raue, F., Hees, J. and Dengel, A. (2021) Esresnet: Environmental Sound Classification Based on Visual Domain Models. 2020 25*th International Conference on Pattern Recognition* (*ICPR*), Milan, 10-15 January 2021, 4933-4940. https://doi.org/10.1109/ICPR48806.2021.9413035

[13] Madhu, A. and Suresh, K. (2019) Improved Deep CNN with Reduced Parameters for Automatic Identification of Environmental Sounds. *International Journal of Engineering Research & Technology* (*IJERT*), **7**, 1-3.

[14] Walden, F., Dasgupta, S., Rahman, M. and Islam, M. (2022) Improving the Environmental Perception of Autonomous Vehicles Using Deep Learning-Based Audio Classification. ArXiv Preprint ArXiv: 2209.04075.

[15] Salamon, J., Jacoby, C. and Bello, J.P. (2014) A Dataset and Taxonomy for Urban

Sound Research. *Proceedings of the* 22*nd ACM international conference on Multimedia*, Orlando, 3-7 November 2014, 1041-1044.
https://doi.org/10.1145/2647868.2655045

[16] MATLAB (2022) Signal Processing Toolbox User's Guide.
https://www.mathworks.com/help/pdf_doc/signal/signal.pdf