Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33353

A Survey on Lightweight Audio Classification for Edge Devices

Donghao Liu

School of Artificial Intelligence, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
125090413@link.cuhk.edu.cn

Abstract. As edge devices and the Internet of Things are widely adopted nowadays, edge devices provide limited computational resources. Thus, the application of traditional deep learning models is restricted, which could not be directly adopted. Therefore, a study of lightweight models is urgently required. This paper surveys the application of lightweight models in audio classification using edge devices, and summarizes the main methods of feature extraction, including Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCC), and Continuous Wavelet Transform (CWT), as well as the optimization ways of the models. Throughout the survey, this study found MFCC got the fastest speed, and CWT has the Best accuracy, and the knowledge distillation could maintain the performance in accuracy whilelowering the size of the model at the same time. Nevertheless, there are still several problems that exist. Current lightweight models often suffer from insufficient specialization in various audio tasks and real- world situations beyond the laboratory. Besides, multi-label classification remains a major difficulty, since the models have poor adaptability to these scenarios. In conclusion, this survey systematically summarizes the mainstream methods in terms of the lightweight models applied to edge devices. The achievements and problems that existed are all about the same accuracy-speed balance. And specialized ways for Audio classification remain the core research direction in the future.

Keywords: Audio Classification, Lightweight Model, Edge Devices, Feature Extraction, Optimization

1. Introduction

As edge devices and the Internet of Things are widely used, the classification of audio samples has become increasingly important in fields such as biological birds monitoring, smart home interaction and recognizing the daily life sounds. The application of the classification of sound samples in edge devices only can own a limited resource, thus, they usually demand lower energy cost and low latency. However, traditional deep learning models are unsuitable for the scenes, due to their high complexity in computation and large parameter sizes. Meanwhile, acoustic samples always contain multiple sound events, for example, there are various kinds of birds calling in the same sample, requiring multi-label classification ability. Therefore, how to develop the efficient Lightweight models used in Edge devices is of great importance.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33353

Most of the previous research focuses on how to improve the performance of lightweight models for audio classification. One is enhancing acoustic feature extraction, and the other is the better optimization of lightweight models. When it comes to the feature extracting methods such as Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCC), Continuous Wavelet Transform (CWT), and some other ways are widely used to transform original audio samples into time frequency representations. They extract effective features and own unique advantages. In the study made by Tsompos et al., STFT was picked as there method [1], while the research conducted by Viswanathuni et al. chose CWT [2], and another classification about bird calls tried MFCC [3]. And when it comes to model optimization, tiny machine learning (TinyML) and knowledge distillation(KD) are widely adopted. TinyML can compress the model size to the level, which is suitable for edge devices, while KD can reduce model size while keeping high accuracy. However, the previous methods still have limitations and lack special designs for different audio edge devices, no matter if it is a limitation about its accuracy or efficiency.

According to the limitations mentioned, this paper focuses on summarizing and concluding existing audio classification methods. How to extract effective features and optimize models is the top priority in existing research. Besides, this paper also pays attention to the characteristics of audio sample classification, which a important content.

2. Foundations for classification

2.1. Audio featuring

Handling the sound samples, the existing studies use different ways to generate the features. During the process of extracting the features, one typical way is to use the STFT. As presented in [4], STFT is always used to transform the raw signal into the spectrogram. And in the practical tasks, the parameters are commonly set to 22.05KHz sample rate, possessing window size of 2,048 samples, while it can be 1024. And a 75%-overlap hop size in hundreds of samples or 50% is two common choices [1]. Unlike the standard Fourier Transform, the STFT not only remains the information of frequency, but preserves temporal details, making it perform better than the standard one and suit for the non-stationary sound samples. However, STFT only provides a fixed window size, which is a bit limited when it comes to the samples with rapid temporal variations. Thus, CWT is a good choice. Mathematically defined by the following integral, CWT offers a time-frequency analysis in detail [2].

$$W_f(s,u) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt \tag{1}$$

Although CWT offers a good choice, it is challenging for the edge devices to use it due to its high-dimensional statistics. In comparison, MFCC is a better choice, for it enables real-time inference on devices using a extremely limited RAM. Applied in featuring the input of bird songs' samples often, MFCC is another widely selected way to extract the features [4], which reduces data size and saves computation. When it comes to the edge devices, the small size of statistic suits the devices better than CWT. Apart from these three, there are also MFE, mel spectrogram and other methods. All of them provide a basis for extracting the features of datasets.

DOI: 10.54254/2755-2721/2026.GU33353

2.2. Lightweight model optimization

With the rapid growth in the machine learning model, TinyML emerges. It enables intelligent processing at the edge devices, while it reduces power cost, which is a typical optimization on compact models [3]. Besides, KD is another way for optimization. As shown in the study conducted by Chen and Ye [3], while KD can improve the performance in the accuracy of classification, it can lower the RAM usage. The distillation function can formalize the goal of low energy.

$$L_{KD} = \alpha T^2 \cdot KLDiv(logSoftmax(\frac{s}{T}), Softmax(\frac{t}{T})) + (1-\alpha) \cdot CE(s, y) \tag{2}$$

3. Research data and methods

3.1. Datasets

In the research on the lightweight sound classification, datasets play a fundamental role and there are lots of different categories of sound samples used in existing studies. This section describes the acoustic datasets often used in the research, which provide the basis for using the methodologies.

A A study in the field of human sound classification utilized 3277 raw human sound samples across 9 distinct classes [2]. Providing real-world data for model training, this dataset is a typical example of edge device applications. What's more, there are many sounds made by humans, such as the mouse click, keyboard typing sounds, knock on the door [5]. The human-environment interaction sound can also help in the recognition in human activity.

InIIn addition to the human voice, biological sound datasets are also a common one. Covering birds' calls and some marine mammal vocalizations, it provides clear samples without too much noise, which suits lightweight models on edge devices. To begin with, the marine mammal samples are a valuable choice. In a conducted study, the researchers got a high-quality dataset, including 7 distinct genres of whales [6] The reason why the study chose mammals for researching is that it focus on real biological sounds, which is suitable for the case when monitoring the biological environments that can only possess edge devices. There are also classifications of birds' sounds. Among various animals' sounds, birds' is a common choice due to its high frequency being heard by humans and the variety of its sounds.

Not only are there animals' and human-make sounds, but environment ones can be used as samples. Washing machine, vacuum cleaner, alarm clock, clock ticking, and broken glass for etc., are chosen by the conductors of research [5].

3.2. Methodology

This section presents how methodologies work and why they are chosen. The overall framework of the existing studies is similar. Normally, Data Collection, Model Design and Training, followed by Model Optimization and Model Deployment, four stages are used in classification [7], as is illustrated in the following chart (Figure 1).

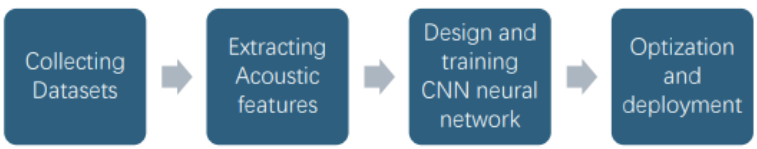

Figure 1. Flowchart for the classification methodology

DOI: 10.54254/2755-2721/2026.GU33353

3.2.1. Collecting datasets

First and foremost, why the research choose the specific genre of datasets is of importance. In a study about birds, the researchers mentioned that birds' songs have unique value for studying the classification on edge devices. It has a high complexity because the birds' songs are used for mate choice [4]. Therefore, the existing classification cannot use in a edge device because of its high energy cost. That is why studying this category is valuable.

Additionally, researchers usually do the resampling work. To ensure the similarity of the inputs, researchers split the inputs into fixed-length pieces of samples. Moreover, the difference between samples cut into different length can lead to different results, since short clips lack enough information, influencing the accuracy while longer ones may influence the speed.

Last but not least, when facing the dataset with some improper statistics, researchers used to make the data cleaning manually. In a conducted study, the researchers remove the anomalous time-frequency images by manual data cleaning. Despite the fact that it leads to a problem of imbalanced data, it ensures the data collected has a highest quality [6].

3.2.2. Extracting acoustic features

In the studies, one of the things of most importance is to choose the correct method of extracting acoustic features. The study conducted by Amarnath and Siddharth shows, due to the fact that CWT can break down the signals into wavelets, it is the best choice for them [5]. And the short ones can be used for analysis in different scales and positions. It is clear that CWT performs well in the durations of capturing detailed time-frequency information. However, it has its own limitation. Since the topic of edge devices and the requirements for the lightweight model, the high complexity and high computation cost is absolutely a big problem. In comparison, MFCC got the fastest speed, possessing only about 10 coefficients (usually 13). Thus, in some tough circumstances needing limited RAM usage, MFCC could be one of the best choices. The same situation is found on STFT. Though it is the moderate one in speed and accuracy, many research also use it for its balanced and overall performance to handle some specific situations.

Via the above comparison, no single method can handle all the situations best. This section illustrates that making decisions according to the situation is of top priority in the selection.

3.2.3. Design and training CNN neural network

Designing and training neural networks is also an irreplaceable part of the task of classification. Research always adjusts their parameters and processing the fine-tuning. Learning rate, network layers, batch sizes, convolutional kernel size and epochs are some of the parameters widely used. Changing the parameters can make a big difference in getting a higher accuracy, and it is the most efficient way to enhance the accuracy after the choice of methodology.

Besides, there is another way to fix the problem, which is to make a new model by the researchers themselves. A research team made a so-called EffiAudioLiteNet model, optimizing the MBConv module according to the need of sound data input. Meanwhile, the new neural network makes the hardware efficiency much higher [8]. That is a example of making a model by the researcher's own. The best thing about making it is the team can balance speed and accuracy by their own and lower the RAM usage as much as possible.

DOI: 10.54254/2755-2721/2026.GU33353

3.2.4. Optimization and deployment

Before the deployment of the neural network, some optimization can fix some problems on the RAM usage while lowering the energy cost. The model will not lose accuracy obviously. Thus, optimization can make the model more suitable for the edge devices and build a more lightweight model.

4. Challenge and future work

First of all, existing research has a common problem, which is it is hard to adjust the parameters when facing different situations. The environment and devices are always changing, which is not a good thing for adjusting parameters. Researchers must have enough experience to get a good adjustment [9,10].

Besides, the balance between speed and accuracy is hard. For example, MFCC owns the fastest speed, but the lowest accuracy. On the contrary, CWT has the highest accuracy with the slowest speed. The only moderate one is STFT, which is not very good at both speed and accuracy. Thus it is very hard to train and design a neural network that can make speed and accuracy as high as possible and keep them balanced.

Ultimately, there is a big problem: in real-world situations, it is hard for researchers to keep the same circumstance as they did in the laboratory. The differences between datasets and situations actually make a big difference and bring a big challenge to the researchers.

In future work, researchers can focus on the methodologies and data selections. How to choose a good dataset for training and how to get a well-trained and well-designed method for extracting acoustic features is the way for research and studies to improve.

5. Conclusions

In conclusion, this survey systematically summarizes the mainstream methods of lightweight models applied on edge devices for audio classification tasks. This survey sort out and analyze the core technologies including STFT, MFCC, CWT and other mainstream feature extraction methods, as well as TinyML, knowledge distillation and other optimization methods.

Most achievements and shortcomings in this area are centered on the accuracy-speed balance. Although existing methods can achieve an efficient balance between accuracy and speed, most models still lack specialization for audio classification tasks, and traditional deep learning methods are too large to be adopted easily. Last but not least, this survey points out the core research direction in the future.

References

[1] Tsompos C, Pavlidis V F, Siozios K. Designing a Lightweight Convolutional Neural Network for Bird Audio Detection [C/OL]//2022 Panhellenic Conference on Electronics & Telecommunications (PACET). Tripolis, Greece: IEEE, 2022: 1-5 [2026-03-15]. https: //ieeexplore.ieee.org/document/9976355/. DOI: 10.1109/PACET56979.2022.9976355.

[2] Viswanathuni V R S, Puranapand A S, PardhasaradhiB, et al. Lightweight CNN for Everyday Human Sound Classification on Edge Devices [C/OL]//2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT). Bhopal, India: IEEE, 2025: 1171-1175 [2026-03-15]. https: //ieeexplore.ieee.org/document/10967858/. DOI: 10.1109/CSNT64827.2025.10967858.

[3] Chen J, Ye T T. Lightweight MCU Implementation (Under 512KB Flash and 256KB RAM) of Neural Networks for Bird Call Classification [C/OL]//2025 IEEE International Conference on High Performance Computing and

DOI: 10.54254/2755-2721/2026.GU33353

Communications (HPCC). Exeter, United Kingdom: IEEE, 2025: 1121-1128 [2026-03-15]. https: //ieeexplore.ieee.org/document/11207442/. DOI: 10.1109/HPCC67675.2025.00161.

[4] Huang Z, Tousnakhoff A, Kozyr P, et al. TinyChirp: Bird Song Recognition Using TinyML Models on Low-power Wireless Acoustic Sensors [C/OL]//2024 IEEE 5th International Symposium on the Internet of Sounds (IS2). Erlangen, Germany: IEEE, 2024: 1-10 [2026-03-15]. https: //ieeexplore.ieee.org/document/10704131/. DOI: 10.1109/IS262782.2024.10704131.

[5] Amarnath S, Pardhasaradhi B, Kumar V S S A, et al. Indoor Sound Classification Using Lightweight CNN with Real-Time IoT Integration [C/OL]//2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT). Bhopal, India: IEEE, 2025: 1149-1153 [2026-03-15]. https: //ieeexplore.ieee.org/document/10968499/. DOI: 10.1109/CSNT64827.2025.10968499.

[6] Kumar A, Kumar V V S S A, Pardhasaradhi B, et al. Marine Mammals Sound Classification Using a Lightweight CNN for Edge Computing Devices [C/OL]//2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT). Bhopal, India: IEEE, 2025: 1182-1187 [2026-03-15]. https: //ieeexplore.ieee.org/document/10968906/. DOI: 10.1109/CSNT64827.2025.10968906.

[7] Uddin M R, Asaduzzaman A, Le K, et al. Voice Activated Edge Devices Using Tiny Machine Learning Enabled Microcontroller [C/OL]//2024 IEEE Green Technologies Conference (GreenTech). Springdale, AR, USA: IEEE, 2024: 38-42 [2026-03-15]. https: //ieeexplore.ieee.org/document/10520459/. DOI: 10.1109/GreenTech58819.2024.10520459.

[8] Han Z, Cao X, Tang C, et al. EffiAudioLiteNet: A Lightweight and Efficient Model for Audio Multi-Label Classification [C/OL]//2025 IEEE 5th International Conference on Computer Communication and Artificial Intelligence (CCAI). Haikou, China: IEEE, 2025: 18-23 [2026-03-15]. https: //ieeexplore.ieee.org/document/11189406/. DOI: 10.1109/CCAI65422.2025.11189406.

[9] Gairí P, Pallejà T, Tresanchez M. Environmental sound recognition on embedded devices using deep learning: a review [J]. Artificial Intelligence Review, 2025, 58(6): 163.

[10] Kim J, Cho J. Low-cost embedded system using convolutional neural networks-based spatiotemporal feature map for real-time human action recognition [J]. Applied Sciences, 2021, 11(11): 4940.