Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 3, 2025

Research on Urban Sound Classification based on ConvNeXt-FECA Model

Tianxiang Zhu [1, 2]

[1] School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, Jilin 132000, China
[2] School of Computer and Information, Dezhou University, Dezhou Shandong, 253000, China

Abstract: With the increasing severity of urban sound pollution, efficient and accurate urban sound classification and recognition has become an important topic in the field of urban environmental monitoring. In this paper, we propose an urban noise classification method based on the Frequency Enhanced Convolution Attention (ConvNeXt-FECA) model. By fusing the Spectrogram and MFCC features in the early stage, the method makes full use of the advantages of the two features, and introduces the Frequency Enhanced Convolutional Attention Mechanism (FECA) to adaptively pay attention to the changes in the frequency band of the audio signal, which effectively improves the classification performance. Experimental results show that the classification accuracy of the ConvNeXt-FECA model is 98.5% on the Urban Sound8K dataset, showing strong robustness and generalization ability.

Keywords: Sound Classification; ConvNeXt; Frequency-enhanced Convolution Attention; Spectrogram; MFCC.

1. Introduction

In recent years, deep learning technologies have been widely applied in urban environmental noise monitoring, gradually replacing traditional manual sampling and analysis methods. For example, audio classification methods based on convolutional neural networks (CNN) have performed well on datasets such as UrbanSound8K[1]. However, existing methods still show insufficient classification ability in complex noise environments, primarily due to limited extraction of frequency domain features.

Toaddress this issue, this paper proposes an urban noise classification method based on ConvNeXt-FECA. Compared to traditional convolutional neural networks, ConvNeXt[2] has a more efficient feature extraction capability, especially suitable for audio data processing. Innovatively, it introduces a frequency-enhanced convolutional attention mechanism (FECA)[3], which can adaptively capture changes in frequency components of audio signals, significantly improving classification accuracy and robustness.

1.1. Data Processing

Research on urban noise classification and environmental audio monitoring has made significant progress. With the continuous development of deep learning technology, more and more studies are beginning to utilize deep neural networks for automatic feature extraction and classification of audio signals. However, the noise classification task still faces challenges such as diverse noise sources, complex environmental backgrounds, and insufficient frequency feature extraction. Therefore, this paper combines early fusion feature extraction methods with the ConvNeXt-FECA model to further improve the performance of audio classification.

1.2. Dataset Processing

This article uses the UrbanSound8K dataset, which consists of 8,732 audio clips categorized into 10 different classes. Each audio clip is approximately 4 seconds long and stored in WAV format. The dataset primarily includes common noises in urban environments, such as traffic noise, mechanical noise, sirens, and dog barks. These categories represent typical urban environmental noise, and each class of audio samples comes from different urban settings, offering diversity and challenges, which can provide researchers with a rich dataset for sound classification tasks.

1.3. Data Preprocessing

In order to improve data diversity and enhance the model's generalization capability, this paper employs various data augmentation techniques in audio data preprocessing, including time stretching, gain adjustment, and noise addition, which are specifically defined as follows: (1) The time stretching technique extends the data by changing the playback speed of the audio signal, and its formula is defined as shown in (1):

$$X'(t) = X(\frac{t}{\partial}), \quad \partial \in \{0.9, 1.0, 1.1\} \tag{1}$$

In the formula: X(t) represents the original audio signal; the time scaling factor indicates that >1 means audio acceleration, and <1 means audio deceleration.

Gain adjustment achieves volume change by scaling the amplitude of the signal, as defined by the formula shown in (2):

$$X' = X \cdot 10^{\frac{Gain(dB)}{20}} \tag{2}$$

In the formula: Gain(dB) refers to the magnitude of gain, typically within the range of (-15, 15) dB, used to simulate different recording environments.

Noise addition enhances the audio signal by superimposing background noise, as defined by the formula shown in (3):

$$X' = X + N, \quad \text{SNR} = 10 \cdot \log_{10}(\frac{\| X \|^2}{\| N \|^2}) \tag{3}$$

In the equation: N represents the background noise signal, SNR is the signal-to-noise ratio, which controls the ratio of signal to noise, typically ranging from (10, 50) dB.

The audio data has been extended in the time domain and frequency domain through the above enhancement methods,

effectively improving the training efficiency and robustness of the model.

1.4. Audio Feature Extraction

Early feature fusion refers to the combination of features from different extraction methods at the initial stage of audio signal processing, thereby providing a richer audio representation. In this study, we adopt an early feature fusion method using Spectrogram and MFCC, aiming to leverage the advantages of both features to enable the model to simultaneously capture temporal-frequency information and speech feature information. Below are the relevant formulas.

The Spectrogram feature extraction method calculates the time-frequency distribution of audio signals through the Short-Time Fourier Transform (STFT), with its formula defined as shown in (4):

$$S(t,f) = \left| \sum_{n=0}^{N-1} x[n] \cdot w[n-t] \cdot e^{-j2\pi fn} \right|^2 \quad (4)$$

In the formula: x[n] represents the input audio signal; w[n] represents the window function; t and f represent the time and frequency indices, respectively.

The MFCC feature extraction method is based on the calculation of Mel frequency cepstral coefficients using a Mel filter bank, with the formula defined as shown in (5):

$$MFCC(\mathrm{m}) = \sum_{k=1}^{K} \log(S(k)) \cdot \cos\left(\frac{m(k-0.5)\pi}{K}\right) \quad (5)$$

In the formula: K represents the number of Mel filters, S(k) represents the result of the pinpu image after passing through the Mel filter bank.

The features of the Spectrogram and MFCC are dimensionally concatenated to form a comprehensive feature representation, as defined by the formula shown in (6):

$$F_{\text{fused}} = Concat(F_{Spectrogram}, F_{MFCC}) \quad (6)$$

In the formula: indicates the result of the dimension concatenation of the features of Spectrogram and MFCC.

The specific process begins by extracting the Spectrogram and MFCC from the audio signal. Figure 1 shows a diagram of the Spectrogram extracted alone, while Figure 2 shows a diagram of the MFCC extracted alone. The Spectrogram provides the frequency and time structure of the audio signal, whereas the MFCC focuses on extracting the speech features of the audio signal. After obtaining these two features, we concatenate them along the feature dimension as shown in Figure 3. This way, the information in the Spectrogram and MFCC can be fully integrated to form a comprehensive feature representation. The fused features undergo normalization to enhance the stability during model training. By removing the mean of the features and performing standardization, the amplitude differences between different audio signals can be reduced, thus helping to improve the learning efficiency of the model. The result of feature fusion is a feature map that contains composite information from both the Spectrogram and MFCC. In the time-frequency domain, this fusion can retain the broad coverage of the Spectrogram in the frequency domain as well as the detailed description of speech feature extraction provided by the MFCC. This enables the model to understand the information in the audio signal more comprehensively, making it particularly suitable for tasks such as audio classification and speech recognition.

Fig 1. Spectrogram

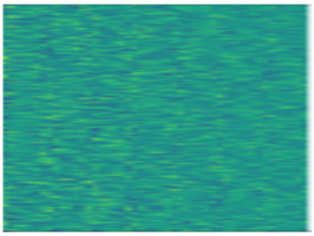

Fig 2. MFCC

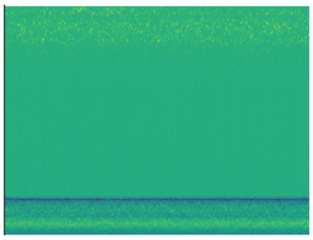

Fig 3. Feature Fusion

2. Model Design

ConvNeXt is an improved convolutional neural network proposed in recent years, characterized by its efficient feature extraction capability and lower computational overhead. Compared to traditional convolutional neural networks (such as ResNet [4]), ConvNeXt employs deeper network layers and optimized convolution operations, effectively extracting local features from audio signals, and demonstrating excellent performance particularly when processing time-frequency images (such as spectrograms).

In this study, ConvNeXt was used as the base model for noise classification. The design of ConvNeXt includes multiple convolutional layers and residual connections, with each layer's convolution kernel size, stride, and number of channels carefully designed to meet the extraction needs of different types of audio features. By performing multi-level feature extraction on audio signals, ConvNeXt can effectively capture complex frequency and temporal patterns, enhancing the model's ability to classify different noise sources.

2.1. Network Setup

This article presents a multi-scale adaptive attention mechanism based on frequency domain enhancement, aimed at optimizing feature extraction capabilities in audio analysis tasks. The model uses a ConvNeXt structure and incorporates an improved Enhanced Frequency Channel Attention Mechanism, which enhances feature capture expression through multi-scale convolution and phase information, as shown in Figure 4, which illustrates the complete model structure.

The entire model flowchart mainly consists of 4 modules, each with the following functions and submodules, (1) Initial convolution module: convolution layer Conv1, batch normalization BatchNorm, activation function. (2) Backbone network module: residual block, frequency domain enhancement ECA attention mechanism, downsampling. (3) Enhanced frequency ECA attention module: FFT, multi-scale convolution, phase enhancement, Sigmoid activation. (4) Pooling module: adaptive global pooling ASP.

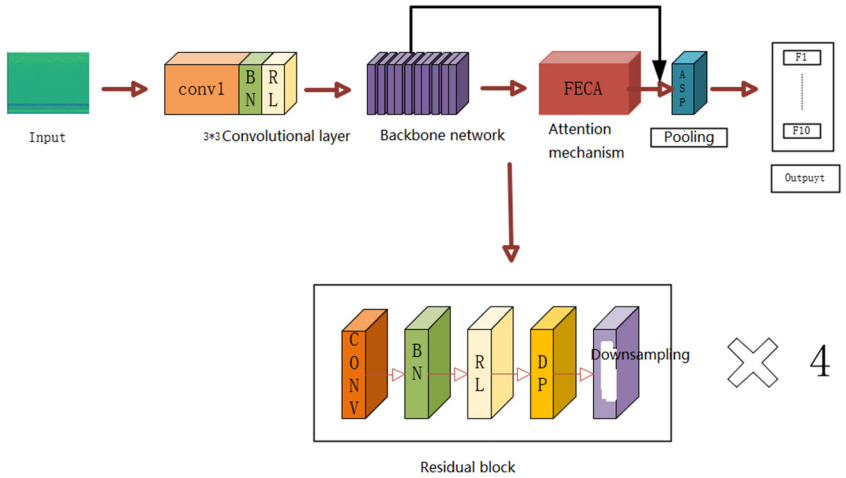

Fig 4. ConvNeXt-FECA model structure

2.2. Initial Convolution Module

The initial convolution module performs primary feature extraction on the input through a 3×3 convolution layer, and achieves normalization and nonlinear mapping of the feature space through Batch Normalization (BN) and the nonlinear activation function ReLU. The convolution operation can capture local time-frequency features, generating preliminary feature representations. Batch normalization alleviates the internal covariate shift problem through normalization, accelerating network convergence. The use of ReLU introduces nonlinearity, enhancing feature representation. This module provides high-quality initial feature representations for subsequent network layers and enhances the stability and resolution of feature distribution through normalization and activation.

2.3. Backbone Network Module

The backbone network module realizes multi-scale feature extraction and downsampling through the stacking of four levels (Layer 1 to Layer 4) as shown in Figure 5. Each level consists of multiple residual blocks to mitigate the gradient vanishing problem and enhance the effective utilization of network depth[5]. Design of residual blocks: each residual block integrates two convolutional layers, batch normalization, and ReLU activation, while introducing the frequency domain enhanced ECA attention mechanism to optimize feature weights. Downsampling gradually reduces the size of the feature maps through stride adjustments between levels, enhancing the model's time-frequency resolution capabilities. Effective gradient propagation and deep feature learning are achieved through the residual structure. The enhanced frequency ECA attention mechanism is introduced during the feature extraction process.

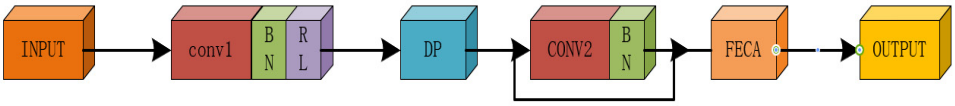

Fig 5. Main network module

2.4. Enhanced Frequency ECA Attention Module

This module is centered on frequency domain operations, integrating multi-scale convolution and phase information to enhance the precision of channel weight calculations as shown in Figure 6. The frequency domain transformation utilizes Fast Fourier Transform (FFT)[6] to map feature maps from the spatial domain to the frequency domain, separating magnitude and phase information. Multi-scale convolution processes frequency domain magnitude information using convolution kernels of different scales to capture features across multiple frequency ranges. Phase enhancement involves enabling phase processing and considering phase information during the reconstruction of feature maps to enhance feature expressive capability. Inverse transformation uses Inverse Fast Fourier Transform (IFFT) to map the processed frequency domain information back to the spatial domain. Activation and channel weight calculation generate attention weights through Sigmoid activation and dynamically adjust channel features. This module effectively combines information from both the frequency and spatial domains, introducing multi-scale perception and phase enhancement to improve the model's robustness and sensitivity to complex audio signals.

2.5. Pooling Module

The pooling module is shown in Figure 7. Through Adaptive Average Pooling (ASP)[7], it maps the feature of each channel to a single value, significantly reducing computational complexity and enhancing the global representation of features. The pooling operation: by global

aggregation, ensures that important information is preserved in each channel's features during the dimensionality reduction process. Adaptive pooling accommodates different input sizes and ensures that the network maintains consistent performance when processing audio data of various scales. The output features from Layer 4 are used as input, which then undergoes adaptive global pooling, flattening of features, and a fully connected layer (FC) to map the features into the classification target space, ultimately outputting the classification results.

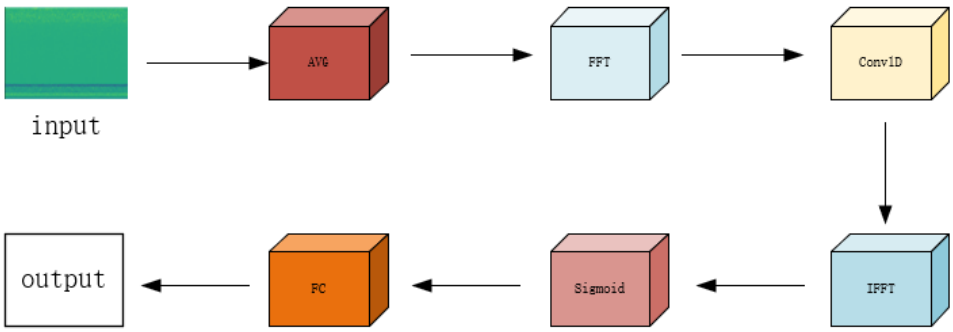

Fig 6. Enhanced Frequency ECA Attention Module

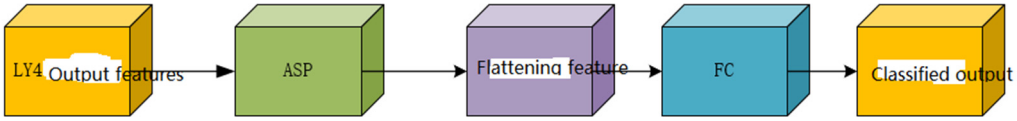

Fig 7. Pooling Module

3. Experimental Results and Analysis

3.1. Experimental Environment and Parameter Settings

The experiment was carried out using the Pytorch deep learning framework, with an NVIDIA GeForce GTX 3080Ti GPU processor, 12GB of memory, and the Ubuntu 20.04 operating system. In the UrbanSound8K dataset, 80% of the audio was designated as the training set and 20% as the test set. The adaptive moment estimation optimization algorithm (Adam) was chosen for gradient optimization, with an initial learning rate set to 0.001 and weight decay mechanisms implemented to mitigate overfitting. The training consisted of 60 epochs with a batch size of 32. This paper employs transfer learning by using the pre-trained weights of the ConvNeXt network on UrbanSound8K as the initial weights for the model in this study.

3.2. Experimental Results and Analysis

According to commonly used performance evaluation standards, the higher the accuracy of the model, the more precise its recognition ability, allowing for effective and accurate classification of sounds in urban environments; a lower loss rate indicates that the model's robustness is relatively strong. During the model training process, the comparison of training loss versus evaluation loss and the change in training accuracy versus evaluation accuracy are shown in Figures 8 and 9. The final validation set accuracy and loss for this experiment are 0.985 and 0.063, respectively, indicating good convergence of the network. During training, the training loss shows a gradual decrease, indicating that the model is continuously learning and optimizing its parameters. However, the evaluation loss (validation loss) remains at a certain level with slight fluctuations. This suggests that although the model performs well on the training data, there is some generalization error on the validation data. The specific changes in loss values are as follows: the training loss gradually decreases as training progresses and eventually stabilizes. The evaluation loss fluctuates compared to the training loss but has an overall declining trend, confirming the model's learning capability and effectiveness. Regarding accuracy, the comparison between training accuracy and evaluation accuracy reveals the model's learning status. The training accuracy continues to rise and approaches 100% in the final stages, indicating that the model can effectively fit the training data. However, the evaluation accuracy is slightly lower than the training accuracy, which indicates that the model's performance on the validation set does not fully match that on the training set. The specific accuracy data is as follows: training accuracy steadily increases as training progresses, ultimately approaching 100%, while the evaluation accuracy, although slightly lower than the training accuracy, shows minimal fluctuation, indicating that the model possesses a certain degree of generalization ability.

In order to further analyze the model's performance across different categories, we present the confusion matrix of the model evaluation as shown in Figure 10. The confusion matrix provides a comparison between the true labels and predicted labels for each category, which can help us identify the model's recognition capability in specific categories, especially in terms of misclassification. Based on the analysis results of the confusion matrix, the model performs well in most categories, but there are some misclassification phenomena in certain categories, mainly manifested as confusion between categories. Through the analysis of the confusion matrix, the model demonstrates excellent classification ability for most categories, with an accuracy

rate exceeding 98%. For instance, the classification accuracy for the gun_shot and jackhammer categories reached 100%, indicating that the model shows extremely high sensitivity to sounds with significantly high-frequency features.

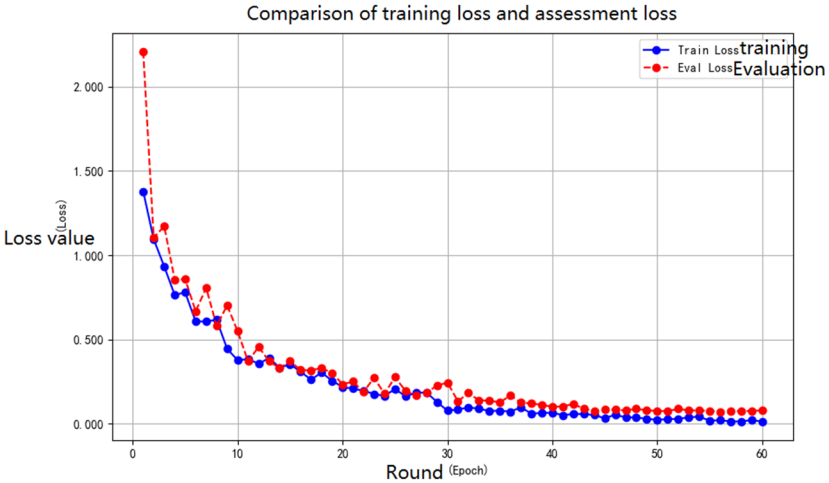

Fig 8. Comparison of training loss and evaluation loss

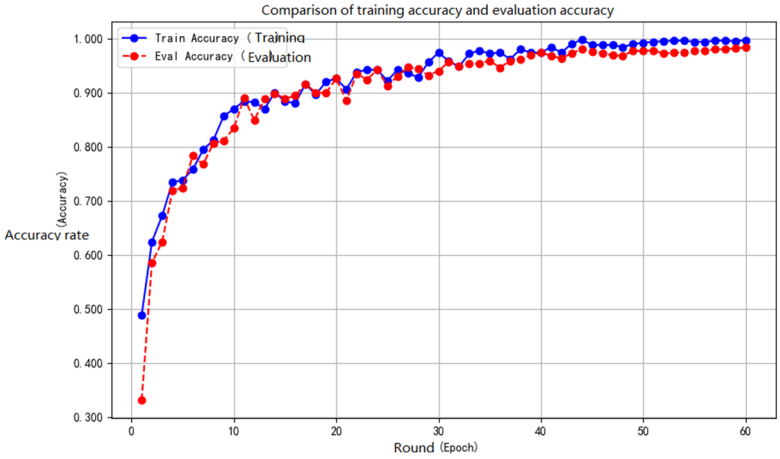

Fig 9. Comparison of training accuracy and evaluation accuracy

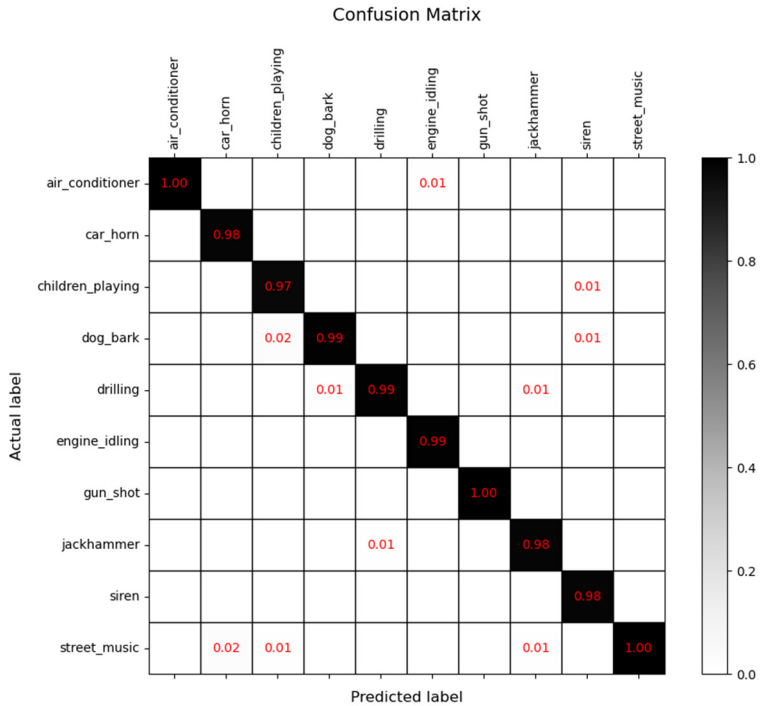

Fig 10. Confusion matrix of the experimental model on the test set

3.3. Experimental Comparative Analysis

In order to comprehensively evaluate the performance of the ConvNeXt-FECA model proposed in this paper for audio classification tasks, we conducted comparative experiments with several current mainstream audio analysis models, including ERes2NetV2[8], ResNetSE, ERes2Net, CAMPPlus [9], PANNS (CNN10)[10], and EcapaTdnn[11]. The experiments were uniformly based on the UrbanSound8K dataset, using a 10-class audio classification task as the experimental object, and consistently employing Spectrogram as the feature extraction method. Table 1 shows the classification accuracy of different models, where the accuracy of the ConvNeXt-FECA model reaches 0.978, outperforming all comparison models. Compared to the standard ConvNeXt model (accuracy 0.976), the FECA mechanism enhances the ability to extract frequency domain features and combines multi-scale convolution operations, giving the model an advantage in feature representation for complex environmental audio. Moreover, compared with classic models such as ResNetSE and CAMPPlus, the classification performance of ConvNeXt-FECA is significantly improved, further proving its superiority on the UrbanSound8K dataset. These experimental results indicate that the ConvNeXt-FECA model proposed in this paper has strong generalization ability and robustness in audio classification tasks.

Table 1. Performance Indicators of Mainstream Audio Analysis Models

model	Preprocessing method	Dataset	Number of categories	Accuracy
ConvNeXt-FECA	Spectrogram	UrbanSound8K	10	0.978
ConvNeXt	Spectrogram	UrbanSound8K	10	0.969
ERes2NetV2	Spectrogram	UrbanSound8K	10	0.956
ResNetSE	Spectrogram	UrbanSound8K	10	0.955
ERes2Net	Spectrogram	UrbanSound8K	10	0.948
CAMPPlus	Spectrogram	UrbanSound8K	10	0.954
PANNS (CNN10)	Spectrogram	UrbanSound8K	10	0.938
EcapaTdnn	Spectrogram	UrbanSound8K	10	0.935

In order to further explore the impact of different feature extraction methods on model performance, we based our experiments on the ConvNeXt-FECA model, using Spectrogram, MFCC, MelSpectrogram, and various feature fusion methods (such as Spectrogram MFCC and MFCC MelSpectrogram, etc.) for experimental analysis. The results are shown in Table 2. When using Spectrogram for feature extraction, the model's accuracy reached 0.978, while the accuracies for MFCC and MelSpectrogram were 0.976 and 0.964, respectively. This indicates that Spectrogram has certain advantages in capturing the time-frequency domain information of audio. However, when we fused different feature extraction methods, the model's classification performance was significantly improved, especially when using the fused features of Spectrogram and MFCC, achieving a classification accuracy of 0.985. This result suggests that Spectrogram and MFCC have complementary characteristics in terms of time-frequency domain information and their fused features can represent the complexity of audio signals more comprehensively. Additionally, the accuracies when using the features of Spectrogram MelSpectrogram and MFCC MelSpectrogram also reached 0.972 and 0.970, respectively, both of which outperformed the single feature extraction method.

Table 2. Performance metrics of the ConvNeXt-FECA model with different feature extraction methods.

model	Preprocessing method	Accuracy
ConvNeXt-FECA	Spectrogram	0.978
ConvNeXt-FECA	MFCC	0.976
ConvNeXt-FECA	MelSpectrogram	0.964
ConvNeXt-FECA	Spectrogram+MFCC	0.985
ConvNeXt-FECA	Spectrogram+MelSpectrogram	0.972
ConvNeXt-FECA	MFCC+MelSpectrogram	0.970

Based on the experimental results, it can be seen that the ConvNeXt-FECA model, by introducing the frequency domain enhanced ECA attention mechanism, effectively improves the multi-scale feature extraction capability and has significant advantages in the complexity modeling of audio signals. Meanwhile, the choice of feature extraction method has a critical impact on model performance; among the single-feature methods, Spectrogram performs the best, while the combination of fused features, especially Spectrogram and MFCC, further enhances the model's classification ability. Future research can further optimize the model structure and parameters to achieve a higher level of classification capability.

4. Conclusion

This article focuses on the application of audio analysis technology in urban environmental monitoring. To address the limitations of current audio classification tasks in complex acoustic scenes, an improved method based on the ConvNeXt-FECA model is proposed. The main contributions of this paper are reflected in several aspects: First, by combining the powerful feature extraction capability of ConvNeXt with the frequency-domain enhanced attention mechanism (FECA), an efficient model for environmental audio classification, ConvNeXt-FECA, is designed. Second, through experiments on the UrbanSound8K dataset, we systematically evaluated the performance of the model, and the results showed that it outperformed various mainstream comparative models in terms of classification accuracy, robustness, and generalization ability. Additionally, this paper further explores the impact of different feature extraction methods and their fusion approaches on model performance, with experimental results indicating that the Spectrogram MFCC feature fusion method has significant advantages in capturing time-frequency information of audio signals, providing a new perspective for feature modeling of complex audio data.

However, this study still has certain limitations. On one hand, the current experiments are mainly focused on the UrbanSound8K dataset, which, although it covers a variety of common urban environmental sounds, needs further validation for applicability on larger scale and more complex

scene datasets; on the other hand, feature fusion methods may have certain burdens in terms of computational costs, and how to optimize model computational efficiency while improving performance is a direction that requires in-depth research in the future. Additionally, audio analysis technology also needs to address various non-stationary noises and dynamic environmental changes in practical environmental monitoring applications, which raises higher requirements for the model's real-time performance and robustness.

References

[1] Chen X,Wang M,Kan R, et al. Improved Patch-Mix Transformer Sound Classification in Noisy Environments using Contrastive Learning methods [J]. Applied Sciences, 2024, 14（21）: 9711-9711.

[2] Talukder A M, Khalid M, Kazi M, et al. A hybrid cardiovascular arrhythmia disease detection using ConvNeXt-X models on electrocardiogram signals. [J]. Scientific reports, 2024, 14(1):30366.

[3] Huang Wenbo, Huang Yuxiang, Yao Yuan, et al. Automatic Grading of ConvNeXt Retinopathy Fused with Attention [J]. Optics and Precision Engineering, 2022,30(17):2147-2154.

[4] Wang J, Zhang B, Yin D, et al. Distribution network fault identification method based on multimodal ResNet with recorded waveform-driven feature extraction[J]. Energy Reports, 2025, 1390-104.

[5] Lozoya L S R ,Domínguez O J D H ,Azuela S H J , et al. Residual shallow convolutional neural network to classify microcalcifications clusters in digital mammograms[J]. Biomedical Signal Processing and Control,2025,102107209-107209.

[6] Schmidt A ,Gierden C, Heinen F R , et al.Efficient thermo-mechanically coupled and geometrically nonlinear two-scale FE-FFT-based modeling of elasto-viscoplastic polycrystalline materials[J].Computer Methods in Applied Mechanics and Engineering,2025,435117648-117648.

[7] Hantao Q, Xin G, Hualei X , et al.Comprehensive receptive field adaptive graph convolutional networks for action recognition[J].Journal of Visual Communication and Image Representation,2023,97.

[8] Chen Y , Zheng S , Wang H ,et al.ERes2NetV2: Boosting Short-Duration Speaker Verification Performance with Computational Efficiency[J]. arXiv preprint arXiv: 2406. 02167 (2024).

[9] Curtis J. KEEGAN ISSUES WARNING TO PLAYERS; FOOTBALL: Latest from the England Camp Plus Local Round-Up[J].[2024-12-25].

[10] Ning L, Weina J ,Xia L , et al.A high mechanical strength, deformable, fatigue-resistant polyacrylonitrile nanosphere-reinforced gel electrolyte for supercapacitors[J].Chemical Engineering Journal,2023,474.

[11] Yang Junjie, Ding Jiahui, Weng Shilong, et al. An automatic classification method for indoor environmental sounds based on lightweight ECAPA-TDNN neural network: CN2022117 15093.7[P].CN116013276A[2024-12-25].