Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9937383, 11 pages
https://doi.org/10.1155/2021/9937383

Research Article

Sound Classification Based on Multihead Attention and Support Vector Machine

Lei Yang and Hongdong Zhao

School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China

Correspondence should be addressed to Hongdong Zhao; zhaohd@hebut.edu.cn

Received 18 March 2021; Revised 9 April 2021; Accepted 12 April 2021; Published 8 May 2021

Academic Editor: Ali Ahmadian

Sound classification is a broad area of research that has gained much attention in recent years. The sound classification systems based on recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have undergone significant enhancements in the recognition capability of models. However, their computational complexity and inadequate exploration of global dependencies for long sequences restrict improvements in their classification results. In this paper, we show that there are still opportunities to improve the performance of sound classification by substituting the recurrent architecture with the parallel processing structure in the feature extraction. In light of the small-scale and high-dimension sound datasets, we propose the use of the multihead attention and support vector machine (SVM) for sound taxonomy. The multihead attention is taken as the feature extractor to obtain salient features, and SVM is taken as the classifier to recognize all categories. Extensive experiments are conducted across three acoustically characterized public datasets, UrbanSound8K, GTZAN, and IEMOCAP, by using two commonly used audio spectrograms as inputs, respectively, and we fully evaluate the impact of parameters and feature types on classification accuracy. Our results suggest that the proposed model can reach comparable performance with existing methods and reveal its strong generalization ability of sound taxonomy.

1. Introduction

As an essential medium in human-machine interactions, automatic sound classification (ASC) is a wide area of study that has been investigated for years, mostly focusing on its subareas such as environment sound classification (ESC), music genre classification (MGC), and speech emotion recognition (SER). Few works attend to the uniform model for these specific tasks, which can increase the expenditure on the hardware devices in the application scenarios. Hence, how to explore a general-purpose approach to improve sound classification performance remains a challenge for researchers.

In recent years, deep learning approaches are drawing more attention in sound classification tasks. Due to its strong learning ability and excellent generalizability to extract task-specific hierarchical feature representations from large quantities of training data, the deep neural network (DNN) has shown impressive performance in the research of automatic speech recognition and music information retrieval [1, 2]. Compared with traditional machine learning classifiers [3], the DNN could capture features from the raw data and obtain impressive results. Han et al. [4] propose to train a DNN model using segments with the highest energy in one utterance as inputs to discover valid segment-level feature information. However, the DNN's in-depth fully connected architecture is not robust for feature conversion. Several latest studies have revealed that the CNN is powerful in investigating potential relationships between adjacent frames. CNN involves a group of interleaved convolutional layers and pooling layers followed by a certain number of fully connected layers. It shares weights by using locally connected filters, which makes inputs translation invariant, and these convolution filters have interpretable time and frequency significance for the audio spectrogram. Mao et al. [5] utilize the CNN to learn affect-salient feature representations from local invariant features that are preprocessed by a sparse autoencoder for the speech emotion

recognition task and achieve stable performance on several benchmark datasets. Zhang et al. [6] improved the model effectiveness by using structural adjustment of the CNN and newly generated training data with the mixup method. Medhat et al. [7] have found that the enforced systematic sparseness of embedded filter banks within the model could facilitate exploring the nature of audio. Audio signals could convey contextual information in the temporal domain, which means the audio information at the current time step is relevant to that at the previous time steps. Hence, it is beneficial to apply the RNN and LSTM to capture time-dependent feature representation in sound classification tasks [8]. As an advanced version of RNN architecture, LSTM has a built-in storage gate to retain temporal context-related information. It is suitable for LSTM to extract chronological characteristics from input sequences. Sainath et al. [9] have demonstrated that LSTM outperforms the conventional RNN techniques in handling large-scale training data. Nevertheless, LSTM's sequential nature means increasing the training time since larger sequence lengths require more training steps. Furthermore, the structures of RNN and LSTM have to suffer from a big problem of long-range dependency disappearance, which can affect the performance stability of models in real-world application scenarios.

The attention mechanism has become a hot topic in the sequence-to-sequence field in recent times. As a type of attention mechanism, self-attention could learn inherent spatial connections between each frame in a sequence to capture the hierarchical structure of the whole sequence. Transformer [10] is a sequence transduction model based entirely on the self-attention mechanism and has shown excellent performance in the field of natural language processing (NLP). Compared with traditional RNN networks, transformer replaces the recurrent structure with multihead attention to process all frames of the sequence in parallel efficiently, which is able to solve the problem of long-range dependency disappearing and greatly outperforms LSTM in NLP, such as pretraining language models [11, 12] and the end-to-end speech recognition approach [13]. Moreover, the high training efficiency of the transformer also facilitates the development of models for various real-world scenarios. Since audio signals express their semantic content relying on the implicit relations between their components in sparse positions, the multihead attention mechanism is more suitable for sound classification tasks. As a classifier, the support vector machine (SVM) has shown great advantages in dealing with the small sample, nonlinear, and high-dimensional pattern recognition. It transforms the original samples by nonlinear mapping algorithm from low-dimensional feature space to high-dimensional feature space or even infinite-dimensional feature space (Hilbert space) and uses a radial kernel function to construct the optimal hyperplane in high-dimensional space to make these samples linearly separable. With respect to the small-size and high-dimension sound datasets, this paper proposes an architecture based on two stages: MhaNN, a variant of transformer, where the salient feature representations are extracted, and a second stage based on SVM where we obtain the final classification results on the basis of the extracted feature representations. In order to show that SVM is more suitable for our task, we also report the experimental results of two other classifiers, K-Nearest Neighbor (KNN) and Logistic Regression (LR). The idea of KNN is to assign the most frequent category of the K-nearest samples to that sample. LR is to create regression equations for decision boundaries based on training data and then map the regression equations to classification functions to realize the classification. Experiments showed that the proposed MhaNN-SVM achieved the strongest results across three acoustically characterized datasets and demonstrates its generalization capability.

The summarized contributions of the proposed framework are demonstrated as follows:

(1) Leveraging theories from the multihead attention and SVM, we proposed a sound classification model based on the multihead attention and SVM, which was able to extract the spatiotemporal hierarchical feature representations from inputs through the multihead attention and map these representations into high-dimensional feature space by SVM to further improve classification results. To demonstrate the suitability of SVM for our task, we also tested the classification performances of KNN and LR as final classifiers separately.

(2) To enhance the comparability of our model, we chose two types of features as inputs: mel-spectrograms and the 68-dimensional feature set, both of which are most commonly used in sound classification fields, and also investigate the impacts of the different numbers of layers or heads to the model.

(3) We compared the proposed model with other top-performing methods on the UrbanSound8K, GTZAN, and IEMOCAP dataset individually to access its generalization capability. Experiment results showed the simplicity and the robustness of our model and demonstrated its applicability for the actual application.

The remainder of the paper is organized as follows: Section 2 reviews some related work in the field of sound classification. Section 3 describes the details of the proposed model. The evaluation of the model is provided in Sections 4, and Section 5 presents our conclusions.

2. Related Works

Most works on the abovementioned architectures are commonly restricted to a single audio classification task, and a few studies have focused on a uniform framework to solve different audio classification problems. To compare with our experiments on the same datasets, we review some task-independent models.

Medhat et al. [14] suggest a binary-masked CNN model that adopts a controlled systematic sparseness such as embedding a filterbank-like behavior within the network to preserve the spatial locality of features in the process of

training weights. The experiments on GTZAN and UrbanSound8K have achieved the accuracies of 85.1% and 74.2%, respectively. In their model, they introduce a set of hidden layers in which each neuron establishes a connection with the input by the activation function through the influence of distinct active regions in the feature vector, so the spatial information of the learned feature is saved as these active weights' locations which are fixed in position. In order to exploit the interframe relations in a sequential signal, the network allows a concurrent exploration of a range of feature combinations to find the optimum combination of features through an exhaustive manual search. This causes the increased complexity and training difficulties, especially when the number of features increased substantially.

Palanisamy et al. [15] present an ensembled Dense CNN architecture with five identical dense blocks. They pretrain the model with a larger size of image dataset ImageNet for the problems of MRC and ESC on the basis of transfer learning knowledge and then initialize the network with these pretrained weights instead of random settings, which brings 90.50% accuracy on the GTZAN dataset and 87.42% accuracy on the UrbanSound8K dataset. This model can solve the problem of poor prediction results caused by the insufficient number of samples. However, its performance is based on consistent data distribution between the pretraining dataset and the target dataset, which limits the broad exploitation of the model for real application scenarios.

Choi et al. [16] provide a transfer learning approach. They design a CNN network with five convolutional layers for a source task on a training dataset and then use the network with trained weights as a feature extractor for target tasks. The network achieves the accuracies of 89.8% on GTZAN and 69.1% on UrbanSound8K. Since the authors choose the Million Song dataset to pretrain their model, the experimental results of GTZAN outperform UrbanSound8K, which illustrates the importance of data distribution consistency between the pretraining dataset and the target dataset.

Transformer was first presented in 2017, and its related studies mainly concentrate on the area of text processing. Relevant to the sound classification on the same datasets used in our experiments is the multimodal transformer [17]; Delbrouck et al. induce a multimodality fusion method based on LSTM and transformer for emotion recognition and sentiment analysis. The method takes mel-spectrograms as input to feed into LSTM, where the linguistic features and acoustic features are extracted independently. Then, these two types of features are fused into a new multimodel representation in the transformer encoder through multihead attention modulation and linear modulation. The experiment on IEMOCAP yields a 74% precision rate that outperforms the single-modal approach, which shows the multihead attention is adept at handling sequential data with different characteristics. Because the multimodal model only has strong effects for some specific types of classification tasks, when encountering large-scale datasets, its internal circular structure will reduce the efficiency of model operations, which is not conducive to the terminal deployment of the model.

3. Proposed Model

In the literature of sound classification, mel-spectrograms and mel-spectrogram-related feature sets have been broadly applied as acoustic features in many deep learning models and shown their powerful performance. In this paper, two types of spectrograms were used as features to be fed into the model, respectively. We applied Librosa [18] to extract the mel-spectrograms, named Feature 1, with 128 mel-filters banks ranging from 0 to 22050 Hz. The PyAudioAnalysis library by Giannakopoulos [19] was used to produce a 68-dimensional feature vector, namely, Feature 2, listed in detail in the following:

(i) 3 time domains: zero-crossing rate, energy, and entropy of energy

(ii) 5 spectral domains: spectral centroid, spectral entropy, spectral flux, spectral roll-off, and spectral spread

(iii) 13 MFCCs

(iv) 13 chroma: 12-dimensional chroma vector and standard deviation of chroma vector

(v) 34 first-order differential features: first-order differential of the abovementioned 34 features

The proposed feature-extracted model MhaNN was composed of a stack of L identical blocks with their own set of training parameters. Each of these contained a multihead self-attention layer and a feed-forward layer. Around each of these two layers, the residual connection [20] was used and followed by layer normalization [21]. To facilitate residual connections, all layers within the model produced the output with dimension $d_{\text{model}} = 256$. The structure of MhaNN is described in Figure 1.

In MhaNN, the attention value is the outcome of a Key K, Query Q, and Value V that interact together, as shown in the following:

$$\text{attention} = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Query Q, Key K, and Value V are input matrices in case of self-attention. QK^T represents similarities between Q and K. $\sqrt{d_k}$ is a scaling factor. In the multihead attention mechanism, a feature space needs to be evenly segmented into m slices, each of which corresponds to one subspace. The multihead attention is to conduct m self-attention functions simultaneously for learning the information from these different subspaces at different positions. Its function is listed as follows:

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_m\right)W^O,\\ \text{head}_i &= \text{Attention}\left(\text{QW}_i^Q, \text{KW}_i^K, \text{VW}_i^V\right),\end{aligned} \tag{2}$$

where $W_i^Q \in R^{d_{\text{model}} \times d_k}$, $W_i^K \in R^{d_{\text{model}} \times d_k}$, $W_i^V \in R^{d_{\text{model}} \times d_v}$, and $W^O \in R^{md_v \times d_{\text{model}}}$. In the feed-forward layer, we adopted a

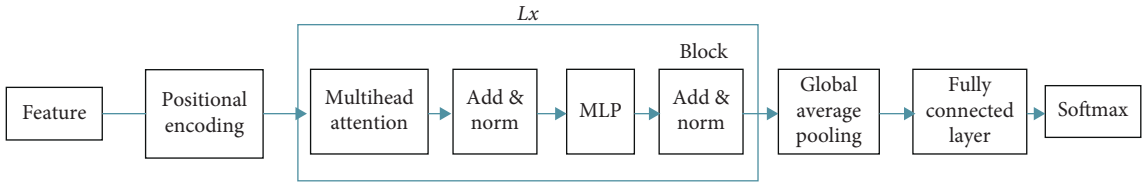


Figure 1: The structure of MhaNN.

multilayer perceptron that consists of two linear projections and a ReLU activation. The hyperparameters in the fully connected layer were set as follows: dropout was 0.5, and the number of neurons was 64. The cross entropy was taken as the loss function. In SVM, we employed the one-versus-one method and choose the radial basis function, in which the penalty coefficient of C is set to 1 and the parameter γ is set to 0.01.

The schematic diagram of our model is shown in Figure 2. Firstly, we entered the target training dataset into MhaNN and extracted features in its global average pooling layer to input into the SVM for training. Then, we applied the trained MhaNN extractor to capture features from the target testing dataset and fed these features into the trained SVM classifier for final results.

4. Experiment Evaluation

4.1. Datasets. In this paper, we evaluated the proposed MhaNN-SVM on three publicly available datasets, respectively, including UrbanSound8K, GTZAN, and IEMOCAP. Their classes and the number of samples contained in each class are given in Table 1.

UrbanSound8K [22]: this dataset is widely used as a benchmark in environmental sound classification. It includes 8732 short (less than 4 seconds) audio clips derived from the live sound recordings of the FreeSound online archive. These clips are pregrouped into ten folders, each containing one of the ten possible urban sound sources.

GTZAN [23]: it is a very popular dataset in the research of music genre classification. It includes 1000 audio excerpts that are collected from radio compact disks and MP3 compressed audio files. Each excerpt is 30 seconds long and annotated with one of the ten music genres.

IEMOCAP [24]: this dataset is designed for the task of multimodal speech emotion recognition. It is composed of twelve hours of audio-visual recordings made by ten professional actors in the form of conversations between two actors of different genders performing scripts or improvising. These collected recordings are divided into short utterances of length between 3 to 15 seconds, each utterance labeled as one of ten emotions (neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited, and other). In this paper, we only use audio data. For the consistent comparison with previous works [25–27], all utterances annotated "excited" are grouped into the happiness category with that annotated "happiness," and only four emotion categories (neutral, happiness, sadness, and anger) are considered. As shown in Table 1, the imbalanced distribution of the IEMOCAP makes a great challenge to the classification task.

4.2. Training and Other Details. To minimize the cross entropy, we employed the ADAM optimizer to train the model on three public datasets for 400 epochs with a batch size of 64, respectively. The 10-fold cross validation was implemented to access the proposed MhaNN-SVM. In all experiments, the training set, test set, and validation set were randomly divided at the ratio of 8/1/1, and we calculated the classification accuracy as the average of 10-fold cross validation. Our experiments were carried out on the Tensorflow2.0 framework, running on NVIDIA TITAN Xp GPU with 16 GB memory. In the following, we analysed the performance of our model on each dataset in more detail.

For optimizing the parameters of the network, we first tested the MhaNN-SVM with the different number of layers, namely, L, and heads on these two types of features individually to evaluate the impacts of parameters and feature type on the classification results. We also attempted two other classifiers LR and KNN to verify the suitability of SVM for processing high-dimensional data. Last, we chose the best accuracy to compare with other competitive methods.

4.3. Experiment and Discussion on UrbanSound8K. Environment sound classification (ESC) has gained considerable attention in recent times. ESC aims at capturing a variety of natural and artificial sounds, rather than speech and music, presented in an acoustic scene by audio sensors, and classifying these sounds into predefined categories.

4.3.1. Preprocessing. As introduced in Section 3, the preprocessing was implemented with the frame size of 50 ms at a rate of 25 ms to obtain Feature 1 with the size of 128×173 and Feature 2 with 68×173. These two spectrograms are depicted in Figure 3.

4.3.2. Results and Analysis. We fed the two types of spectrograms into the model separately to yield sixty classification accuracies. As listed in Table 2, we can observe that SVM was significantly more effective in improving MhaNN classification results than KNN and LR. Feature 1 made a larger contribution to the improvement of accuracy than Feature 2, no matter how many L and heads were set, which indicated that mel-spectrograms can enhance the predictive capability of the model effectively in ESC tasks. In particular, when we set 4 to head and 3 to L, the accuracy can reach 94.6%, the highest in all of the results. Figure 4 showed the

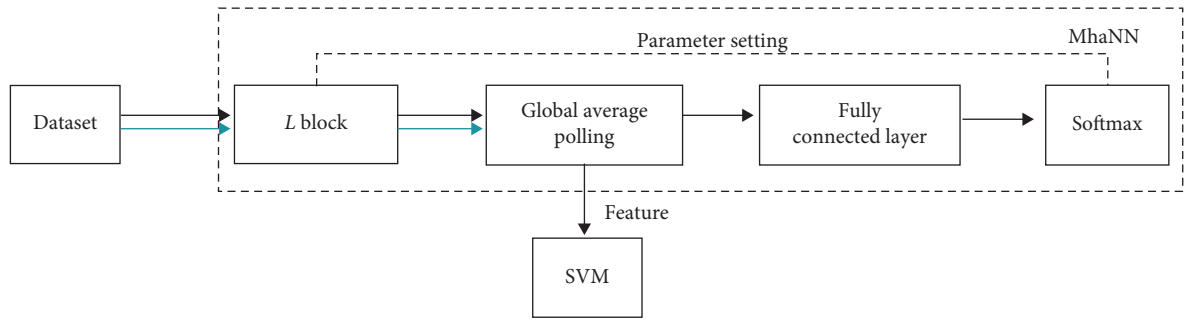


Figure 2: The schematic diagram of MhaNN-SVM.

Table 1: Dataset description.

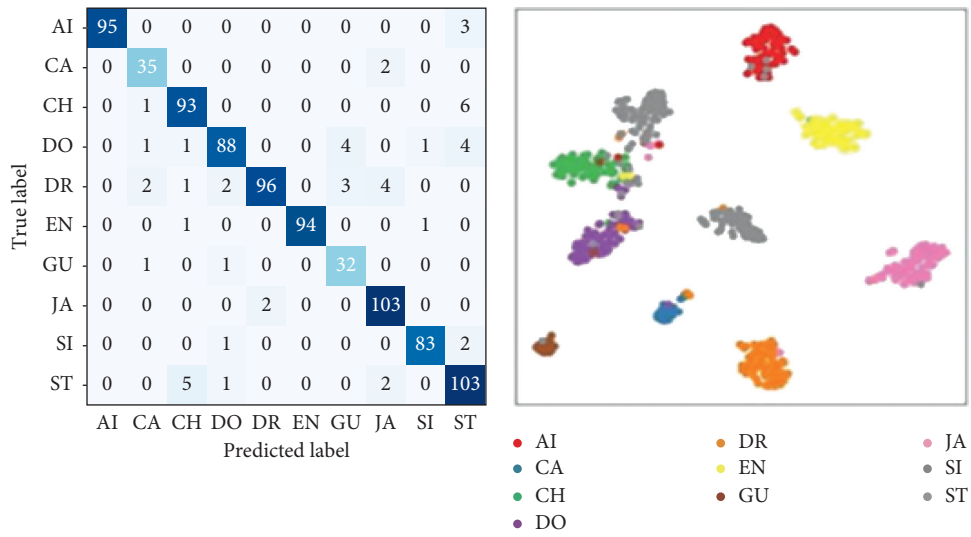
	UrbanSound8K		GTZAN		IEMOCAP	
	Genre	Samples (#)	Genre	Samples (#)	Genre	Samples (#)
0	Air condition (AI)	1000	Classic	100	Happy	1636
1	Car horn (CA)	429	Jazz	100	Sad	1084
2	Children playing (CH)	1000	Blues	100	Angry	1103
3	Dog dark (DO)	1000	Metal	100	Neutral	1708
4	Drilling (DR)	1000	Pop	100		
5	Engine idling (EN)	1000	Rock	100		
6	Gun shot (GU)	374	Country	100		
7	Jackhammer (JA)	1000	Disco	100		
8	Siren (SI)	929	Hip-hop	100		
9	Street music (ST)	1000	Reggae	100		
	Total	8732		1000		5531

(a) (b)

Figure 3: The spectrograms of Feature 1 and Feature 2 on UrbanSound8K. (a) Feature 1 and (b) Feature 2.

Table 2: Classification accuracy on UrbanSound8K compared across different numbers of heads and layers with Feature 1 and Feature 2 individually.

Feature	Head (#)	L (#)	MhaNN accu. (%)	MhaNN-SVM accu. (%)	MhaNN-LR accu. (%)	MhaNN-KNN accu. (%)
Feature 1	2	1	91.6	92.1	92.3	91.5
		2	92.2	93.3	93.0	92.9
		3	91.6	93.3	91.7	92.2
	4	1	91.8	92.7	91.6	92.1
		2	92.1	93.6	92.8	93.2
		3	92.1	94.6	92.3	93.0
	8	1	91.4	93.2	91.0	92.9
		2	90.9	92.1	91.7	91.0
		3	90.5	91.0	90.8	91.2
Feature 2	2	1	83.7	84.8	86.1	85.2
		2	89.1	90.3	87.8	88.1
		3	86.2	87.4	86.1	86.8
	4	1	85.5	86.7	85.9	85.1
		2	87.1	89.7	87.2	88.4
		3	83.0	84.1	82.7	83.0



(a) (b)

Figure 4: Experiment results on UrbanSound8K. (a) Confusion matrix. (b) t-SNE plot.

confusion matrix and *t*-SNE plot. In this confusion matrix, the rows stood for true labels, the columns for predicted labels, and values along the diagonal represent the number of samples classified correctly for each category, while those nondiagonal terms denoted incorrect classification. We can find that the proposed model distinguished most environmental sound categories effectively, but it was difficult to separate drilling from street music. One explanation was that they shared a lot of similar frequency information, which made it more difficult to discriminate them in nature.

Table 3 illustrated the detailed information about precision, recall, and F-score for each category. Overall, most of the categories had been correctly classified, and the F-scores of AI, EN, and JA were able to reach more than 96%. Even though the number of CA and GU training samples were both small, their recognition accuracies were considerably high, based on which we can assume that the proposed model had outstanding classification capability on the UrbanSound8K dataset.

To evaluate the performance of the presented model, we compared the MhaNN-SVM with five other top-performing methods on the UrbanSound8K in Table 4. DenseNet [28] exploited a dense connectivity pattern between layers in a feed-forward pattern to realize feature reuse, which led to high convergence efficiency. MelNet [29] and RawNet [29] both used five convolutional layers with batch normalization and a fully connected layer, but took mel-spectrograms and raw audio waves as input, respectively. GoogleNet [30] applied a deep convolutional neural network that is designed for image recognition to the ESC field, achieving a 93% accuracy rate with a large number of parameters up to 6.7 M. 1D CNN [31] was an end-to-end environmental sound recognition model initialized with Gammatone filter banks at the first convolutional layer. It utilized convolutional layers to capture temporal information from raw waveforms. From these results, it can be found that our model could yield 94.6% accuracy on the UrbanSound8K dataset, surpassing the compared models.

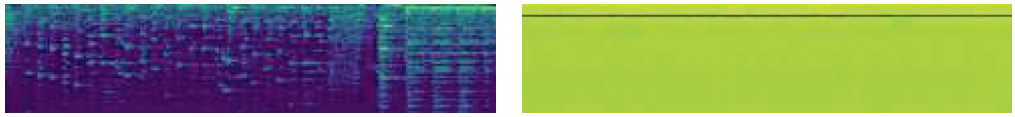
Table 3: Precision, recall rate, and F-score of each category obtained with UrbanSound8K.

Genre	Precision (%)	Recall rate (%)	F-score (%)
AI	97	97	97
CA	92	97	94
CH	84	98	90
DO	95	91	93
DR	99	94	96
EN	98	98	98
GU	100	91	95
JA	99	100	99
SI	95	97	96
ST	97	91	94

Table 4: Comparison of the classification accuracy with six methods on UrbanSound8K.

Methods	Preprocessing	Accuracy (%)
DenseNet [28]	MFCC + GFCC	85.1
MelNet [29]	Mel-spectrogram	92.2
GoogleNet [30]	Mel-spectrogram	93.0
RawNet [29]	Wav	87.0
1D CNN [31]	Wav	89.1
MhaNN-SVM	Mel-spectrogram	94.6

4.4. Experiments and Discussion on GTZAN. Automatic music classification could provide convenience for music storage and improve the efficiency of music information retrieval systems. The challenges involved in the MGC are typically related to a variety of attributes, such as instruments, melody, timbre, the mixing of two or more genres in the same piece of music, and even human voices appearing



(a) (b)

Figure 5: The spectrograms of Feature 1 and Feature 2 on GTZAN. (a) Feature 1 and (b) Feature 2.

in the music, all of which need to be considered for decision making.

4.4.1. Preprocessing. First, we segmented each audio excerpt in the dataset evenly into four parts and then used two different tool libraries to transform each part with 50 ms frame size at a rate of 25 ms into two types of spectrograms separately, Feature 1 and Feature 2. Their own sizes were 128×300 and 68×300, respectively, visualized as Figure 5.

4.4.2. Results and Analysis. The spectrograms extracted from the preceding step were input into the model individually, and all experimental results are listed in Table 5 by adjusting the different numbers of layers and heads. From these results, we can find that MhaNN-SVM had the highest classification result, while LR had no boosting effect on MhaNN. Meanwhile, the performance of Feature 1 was overall better than that of Feature 2, which indicated that mel-spectrogram can improve the accuracy of music genre classification effectively, especially when the head was set to 4 and L to 2, and the model can obtain the best result of 88.4% whose experimental description is shown in Figure 6.

Table 6 lists the precision, recall rate, and F-score for each class. The proposed model recognized most classes, but it was difficult to discriminate rock from country and disco. One reason was that they may share more similar frequency information, which made it more challenging to categorize them in nature. In general, most genres can be correctly recognized, with blues, classic, metal, and hip-hop all achieving over 90% precision rates.

In Table 7, we made comparisons between the MhaNN-SVM with existing models. Hybrid [32] and CVAF [33] investigated different feature fusion strategies to improve classification accuracy, with results of 88.3% and 90.9%, respectively. Choi et al. [16] applied a transfer learning framework to the MGC task and obtained an accuracy of 89.8%. Freitag et al. [34] proposed an AuDeep approach based on the recurrent sequence-to-sequence autoencoder that can extract the characteristics of time series data from their temporal dynamics, resulting in an accuracy of 85.4%. Liu et al. [35] employed a CNN-based BBNN method to extract the low-level information from spectrograms of audio signals for the long context involved in recognition decisions and achieved 93.3% accuracy.

4.5. Experiment and Discussion on IEMOCAP. Human emotions have been playing an essential role in human communication. Speech emotion recognition refers to the analysis of changes of emotion hidden in human conversations, by extracting the relevant features from the speech to feed into the neural network for classification, to identify the possible emotional changes of the speaker.

4.5.1. Preprocessing. All files in the dataset went through the preprocessing step stated in Section 3, with settings of 50 ms frame length and 25 ms frame shift size. As a result, Feature 1 had a size of 128×310 and Feature 2 had 68×310, as portrayed in Figure 7.

4.5.2. Results and Analysis. Two types of spectrograms produced from the preprocessing were input into the network separately and yielded a variety of classification accuracies under different quantities of head and layers. Table 8 showed all the experimental results. The performance of Feature2 outperformed that of Feature1 overall. In terms of impact on MhaNN classification results, SVM was able to improve, LR had no effect, and KNN even decreased, especially in the Feature2 group. One reason is that LR is a linear classification model, and KNN is not able to estimate the prediction error statistically, which leads to great fluctuations in the results. This indicated that compared to environmental events and music genres, we need more kinds of features to make accurate judgments on the emotion. The high dimension of the data and the uneven distribution of data onto categories are not suitable for LR and KNN classifiers. When the model parameters were set as head = 4 and $L = 1$, the accuracy can reach 62.8%, the highest of all the results. We visualized the experiment results of 62.8% in Figure 8.

Table 9 shows the precision, recall, and F-score for each category. Due to the few training samples of the sadness category in the training set, the model can only learn limited discriminative information, so it could misclassify some sad samples into the neutral or happy label. Lower precision and recall rate on the sad label could weaken the ability of the model to identify the sad emotion. On the whole, the precision rates of other labels were all over 60%, including 68% for the angry category, which indicated that MhaNN-SVM had better recognition performance on an unbalanced dataset.

In Table 10, the accuracy of MhaNN-SVM was close to the best available results on the IEMOCAP dataset. Haytham et al. [36] provided an RNN-based SER technique that used mel-spectrograms to train the model at a large computing cost. Gu et al. [37] introduced a multimodal architecture based on hierarchical attention to fuse text and audio at word level for emotion classification. The HSF-CRNN [38] was a two-channel SER system that combined handcrafted

Table 5: Classification accuracy on GTZAN compared across different numbers of heads and layers with Feature 1 and Feature 2 individually.

Feature	Head (#)	L (#)	MhaNN accu. (%)	MhaNN-SVM accu. (%)	MhaNN-LR accu. (%)	MhaNN-KNN accu. (%)
Feature 1	2	1	81.8	82.9	81.6	82.2
		2	82.9	84.0	81.3	83.4
		3	81.2	81.7	79.2	82.0
	4	1	82.3	83.1	82.0	83.3
		2	85.4	88.4	84.7	86.7
		3	84.2	86.1	85.5	84.8
	8	1	82.7	84.8	82.7	83.1
		2	83.6	85.1	83.3	84.1
		3	81.2	83.2	78.7	80.1
Feature 2	2	1	70.1	72.2	70.3	72.0
		2	76.5	78.7	77.0	74.8
		3	72.5	74.6	70.8	72.2
	4	1	71.0	73.3	72.7	70.9
		2	75.1	78.0	76.2	75.8
		3	73.7	75.3	73.6	72.1

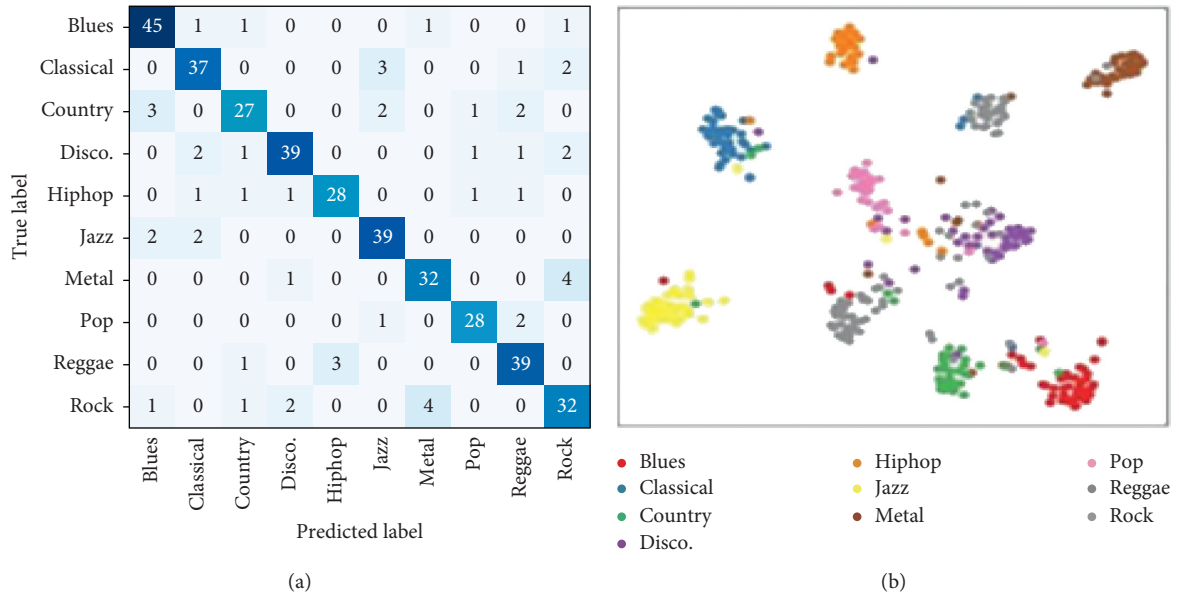


Figure 6: Experiment results on GTZAN. (a) Confusion matrix. (b) t-SNE plot.

Table 6: Precision, recall rate, and F-score of each category obtained on GTZAN.

Genre	Precision (%)	Recall rate (%)	F-score (%)
Blues	98	92	95
Classic	93	86	89
Country	81	86	83
Disco	81	85	83
Hip-hop	100	94	97
Jazz	89	95	92
Metal	92	97	95
Pop	89	100	94
Reggae	90	86	88
Rock	84	78	81

Table 7: Comparison of the classification accuracy with six methods on GTZAN.

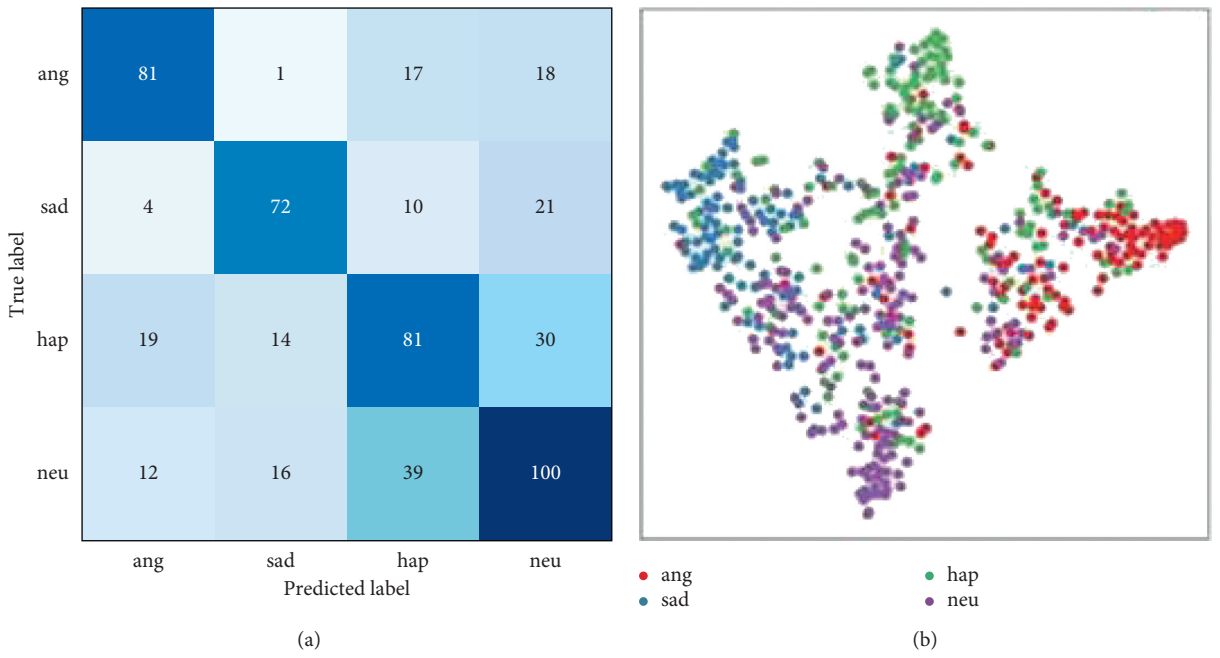
Methods	Preprocessing	Accuracy (%)
Hybrid model [32]	MFCC, SSD, etc.	88.3
CVAF [33]	Mel-spectrogram, SSD, etc	90.9
Transfer learning [16]	MFCC	89.8
AuDeep [34]	Mel-spectrogram	85.4
BBNN [35]	Mel-spectrogram	93.3
MhaNN-SVM	Mel-spectrogram	88.4

(a) (b)

Figure 7: The spectrograms of Feature 1 and Feature 2 on IEMOCAP. (a) Feature 1 and (b) Feature 2.

Table 8: Classification accuracy on IEMOCAP compared across different numbers of heads and layers with Feature1 and Feature2 individually.

Feature	Head (#)	L (#)	MhaNN accu. (%)	MhaNN-SVM accu. (%)	MhaNN-LR accu. (%)	MhaNN-KNN accu. (%)
Feature 1	2	1	53.3	55.3	54.0	52.2
		2	54.8	56.8	55.0	55.5
		3	51.9	53.3	52.3	49.8
	4	1	56.5	58.1	55.7	51.5
		2	54.9	56.8	55.1	54.6
		3	52.7	54.9	52.0	53.3
	8	1	55.6	57.3	56.0	54.9
		2	56.1	58.2	57.4	56.7
		3	52.1	54.2	55.0	53.7
Feature 2	2	1	56.1	58.8	56.1	53.9
		2	58.5	61.1	58.7	54.8
		3	57.8	60.1	57.7	55.5
	4	1	60.5	62.8	58.0	56.2
		2	58.3	60.4	58.2	56.9
		3	57.7	59.5	58.1	55.6



(a) (b)

Figure 8: Experiment results on IEMOCAP. (a) Confusion matrix. (b) t-SNE plot.

Table 9: Precision, recall rate, and F-score of each category obtained on IEMOCAP.

Genre	Precision (%)	Recall rate (%)	F-score (%)
Angry	68	70	69
Sad	54	61	57
Happy	64	62	63
Neutral	64	56	60

Table 10: Comparison of the classification accuracy with six methods on IEMOCAP.

Methods	Preprocessing	Accuracy (%)
LSTM-RNN [36]	Mel-spectrogram	64.8
FAF [37]	Mel-spectrogram	61.4
HSF-CRNN [38]	Mel-spectrogram, LLDs, etc.	60.4
CNN-LSTM-DNN [39]	Raw speech	60.2
Progressive net [40]	Mel-spectrogram, MFCC, etc.	58.1
MhaNN-SVM	Mel-spectrogram, MFCC, etc.	62.8

HSFs and CRNN-learning spectrograms to jointly extract emotion-related feature representations for SER. Reference [39] proposed an architecture of a parallel multilayer CNN stacked on the LSTM to capture multiple temporal-contextual interactions. Lakomkin et al. [40] presented a progressively trained neural network based on transfer knowledge of automatic speech recognition with SER.

5. Conclusions

In this paper, we presented a method of combining the multihead attention mechanism with SVM to replace the recurrent framework most commonly used in sound recognition models. Three benchmarked datasets (UrbanSound8K, GTZAN, and IEMOCAP) were employed to validate its applicability to different acoustically characterized fields. In order to highlight the advantages of SVM as a classifier, we also compare SVM with two other classifiers, LR and KNN. Experiment results demonstrated that the proposed method could bring outstanding improvements in classification accuracy. However, we mainly focused on the generality of the model and did not perform feature fusion and model parameter setting according to the characteristics of each dataset. In the future, we will fully explore the characteristics of different audio datasets and design the task-specific feature representation and model parameters to further evaluate the performance of the transformer.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] M. Schedl, E. Gómez, and J. Urbano, *"Music Information Retrieval: Recent Developments and Applications, Foundations and Trends in Information Retrieval*, Now Publishers Inc., Hanover, MA, USA, 2014.

[3] I. Mcloughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of the Interspeech 2014 Conference*, Singapore, September 2014.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[6] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Proceedings of the 2018 Chinese Conference on Pattern Recognition and Computer Vision*, Guangzhou, China, November 2018.

[7] F. Medhat, D. Chesmore, and J. Robinson, "Masked conditional neural networks for environmental sound classification," in *Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics*, Tokyo, Japan, October 2017.

[8] I. Lezhenmin, N. Bogach, and E. Pyshkin, "Urban sound classification using long short-term memory neural network," in *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*, Leipzig, Germany, September 2019.

[9] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, April 2015.

[10] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Computer Science*, Article ID 49313245, 2018.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the NAACL–HLT 2019 Conference*, Minneapolis, MN, USA, June 2019.

[13] S. Karita, N. E. Y. Soplin, S. Watanabe et al., "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *Proceedings of the Interspeech 2019 Conference*, Graz, Austria, September 2019.

[14] F. Medhat, D. Chesmore, and J. Robinson, "Masked conditional neural networks for sound classification," *Applied Soft Computing*, vol. 90, Article ID 106073, 2020.

[15] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking CNN models for audio classification," November 2020, https://arxiv.org/pdf/2007.11154.pdf.

[16] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proceedings of 18th International Society of Music Information Retrieval Conference*, Suzhou, China, October 2017.

[17] J. Delbrouck, N. Tits, and S. Dupond, "Modulated fusion using transformer for linguistic-acoustic emotion recognition," in *Proceedings of the First International Workshop on Natural Language Processing beyond Text*, Stroudsburg, PA, USA, November 2020.

[18] M. McVicar, C. Raffel, D. Liang et al., "Librosa," 2015, https://github.com/librosa/librosa.

[19] T. Giannakopoulos, "pyAudioAnalysis: an open-source python library for audio signal analysis," *PloS One*, vol. 10, no. 12, Article ID e0144610, 2015.

[20] M. Wan and J. McAuley, "Item recommendation on monotonic behavior chains," in *Proceedings of the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, October 2018.

[21] Q. Xia, P. Jiang, F. Sun et al., "Modelling consumer buying decision for recommendation based on multi-task deep learning," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management CIKM 2018*, Turin, Italy, October 2018.

[22] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM Inter-national Conference on Multimedia*, Orlando, FL, USA, November 2014.

[23] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[24] C. Busso, M. Bulut, C. C. Le et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[25] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop*, Athens, Greece, December 2018.

[26] J. Cho, R. Pappagari, P. Kulkarni et al., "Deep neural networks for emotion recognition combining audio and transcripts," in *Proceedings of the Interspeech 2018 Conference*, Hyderabad, India, September 2018.

[27] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, March 2017.

[28] Z. Huang, C. Liu, H. Fei et al., "Urban sound classification based on 2-order dense convolutional network using dual features," *Applied Acoustics*, vol. 164, Article ID 107243, 2020.

[29] S. Li, Y. Yao, J. Hu et al., "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Applied Science*, vol. 8, no. 7, pp. 1–20, 2018.

[30] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Computer Science*, vol. 112, pp. 2048–2056, 2017.

[31] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[32] N. Karunakaran and A. Arya, "A scalable hybrid classifier for music genre classification using machine learning concepts and spark," in *Proceedings of the 2018 International Conference on Intelligent Autonomous Systems*, Singapore, March 2018.

[33] L. Nanni, Y. M. G. Costa, D. R. Lucio, C. N. Silla, and S. Brahnam, "Combining visual and acoustic features for audio classification tasks," *Pattern Recognition Letters*, vol. 88, pp. 49–56, 2017.

[34] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[35] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2021.

[36] H. M. Favek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[37] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, July 2018.

[38] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proceedings of the Interspeech 2018 Conference*, Hyderabad, India, September 2018.

[39] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," July 2020, https://arxiv.org/pdf/1904.03833.pdf.

[40] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing neural speech representations for auditory emotion recognition," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Taipei, Taiwan, November 2017.