applied sciences

MDPI

Article

An East Asian Traditional Instrument Dataset with Cross-Scale Spectro-Temporal Interaction for Instrument Classification

Chan-Woo Kang [1], Bo Dong [1], Yu Shen [2] and Sungmin Kim [1,*]

[1] School of Mechanical Engineering, University of Ulsan, Ulsan 44610, Republic of Korea; clmogo01@gmail.com (C.-W.K.); sebastian@mail.ulsan.ac.kr (B.D.)
[2] School of Automotive and Traffic Engineering, Liaoning University of Technology, Jinzhou 121001, China; shen12345@mail.ulsan.ac.kr
* Correspondence: sungminkim@ulsan.ac.kr

Abstract

Automatic musical instrument classification is important for music information retrieval and computational music analysis, yet East Asian traditional instruments remain underexplored and dedicated datasets are scarce. Moreover, the acoustically similar timbral and spectral characteristics of related instruments make fine-grained classification challenging. To address these challenges, this study constructs the East Asian Traditional Musical Instrument dataset (EATMI-20) and proposes a Cross-Scale Spectro-Temporal Interaction Network (CSTI-Net). EATMI-20 comprises 20 instrument categories from China, Japan, Korea, and Mongolia, providing a dedicated benchmark for East Asian traditional instrument recognition. CSTI-Net integrates cross-scale acoustic representation learning with spectro-temporal dual-path modeling and bidirectional spectro-temporal interaction to capture complementary information across acoustic scales, spectral structures, and temporal dynamics. The resulting joint representation enables more effective discrimination of acoustically similar instruments. Experimental results show that CSTI-Net outperforms several existing approaches on EATMI-20 and maintains strong performance on additional musical instrument datasets, demonstrating its generalization capability across different data distributions. The proposed dataset and network provide a unified foundation for automatic recognition and computational analysis of East Asian traditional musical instruments.

Keywords: musical instrument classification; East Asian traditional musical instruments; EATMI-20; cross-scale representation learning; spectro-temporal interaction

Check for updates

Academic Editor: Edoardo Piana

Received: 9 October 2026
Revised: 16 November 2026
Accepted: 16 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the rapid development of digital audio technology, music information retrieval (MIR), and intelligent audio analysis, automatic musical instrument classification has become an important research topic in audio signal processing and computational music analysis. Musical instruments exhibit distinctive acoustic characteristics associated with their physical structures, materials, and playing techniques [1]. These characteristics provide important cues for distinguishing instruments and make automatic classification valuable for music content analysis, audio retrieval, automatic annotation, and the digital preservation of musical heritage. East Asia possesses a particularly rich musical heritage, encompassing traditional instruments from China, Japan, Korea, Mongolia, and other regions. Their diverse construction, playing techniques, and performance practices result in complex timbral characteristics, making automatic classification of East Asian traditional musical instruments both valuable and challenging.

Early studies on automatic musical instrument recognition mainly relied on handcrafted acoustic features and conventional classifiers. Kaminskyj and Materka [2] investigated acoustic signal features for monophonic instrument identification, while Brown [3] employed cepstral coefficients and pattern recognition for computer-based instrument identification. Eronen and Klapuri [4] further incorporated temporal features with cepstral coefficients, demonstrating the complementary role of temporal information. Agostini et al. [5] investigated spectral features for musical instrument timbre classification. Subsequently, Kaminskyj and Czaszejko [6] explored k-nearest-neighbor and Gaussian mixture model classifiers for isolated instrument recognition. Although these approaches established the foundation for automatic instrument recognition, their performance depended heavily on manually designed acoustic representations, limiting their ability to capture complex and variable timbral characteristics.

The emergence of deep learning introduced representation learning into musical instrument classification. Park and Lee [7] investigated deep CNN-based instrument sound classification, while Li et al. [8] applied CNNs to instrument recognition in polyphonic music. Compared with conventional methods, CNNs enabled feature extraction and classification to be optimized jointly in an end-to-end framework. However, convolutional operations primarily focus on local patterns through fixed receptive fields, which may limit the modeling of long-range temporal dependencies and relationships between distant acoustic regions.

The Transformer architecture subsequently introduced self-attention for modeling long-range dependencies and global relationships [9]. Attention-based approaches were then increasingly applied to musical instrument analysis. Gururani et al. [10] investigated deep neural networks for instrument activity detection in polyphonic music, and Gururani et al. [11] introduced attention mechanisms to selectively emphasize informative temporal regions. Wang and Yu [12] further investigated an improved CNN using auditory-spectrum representations for instrument identification. These studies demonstrated the value of selectively modeling informative acoustic regions. Nevertheless, many approaches still primarily emphasize either spectral patterns or temporal information, leaving the interaction between frequency structures and temporal dynamics insufficiently explored.

The development of large-scale pretrained audio models further expanded representation learning for audio analysis. Kong et al. [13] proposed PANNs, a large-scale pretrained audio neural network capable of learning general-purpose audio representations from extensive audio data. Such models demonstrated the potential of transferring knowledge from large audio corpora to downstream recognition tasks. However, general-purpose pretrained models are not specifically designed for traditional musical instruments, whose acoustic characteristics may differ substantially from those represented in general audio datasets.

Research specifically targeting traditional musical instruments subsequently became more active. Xu [14] proposed a deep-learning-based approach for recognizing and classifying Chinese traditional musical instruments, while Li and Zhang [15] employed Mel-spectral representations and an eight-layer CNN for Chinese traditional instrument recognition. Reghunath and Rajan [16] further introduced a Transformer-based ensemble approach for predominant instrument recognition in polyphonic music. More recently, Chen et al. [17] investigated hierarchical residual attention with multi-spectrogram representations, Zheng et al. [18] introduced ICKAN based on Kolmogorov–Arnold Networks, and Zhao et al. [19] explored Mamba-based long-range sequence modeling for musical instrument classification. Zheng et al. [20] further developed a two-stage framework for lead-instrument recognition in polyphonic music containing Chinese instruments. These studies demonstrate the continuing evolution toward more powerful feature representations

and sequence modeling, but systematic classification across traditional instruments from multiple East Asian countries remains insufficiently investigated.

Despite these advances, three major challenges remain. First, existing musical instrument recognition research has been largely supported by datasets and benchmarks centered on Western musical traditions, whereas datasets specifically designed for underrepresented and traditional musical cultures remain comparatively limited [21]. This limitation is particularly evident for systematic datasets covering traditional instruments from multiple East Asian countries. Second, musical instrument timbre is characterized by complex spectro-temporal patterns, in which spectral structures and temporal evolution jointly contribute to instrument identity [4,17]. Modeling these dimensions independently may therefore limit the ability to capture their complementary relationships. Third, although recent studies have explored multi-spectrogram representations, attention mechanisms, and temporal sequence modeling for musical instrument recognition [17–20], unified architectures that explicitly integrate multi-scale spectral representations with interactive frequency–temporal modeling remain relatively underexplored, particularly for East Asian traditional musical instruments.

To address the limited availability of dedicated datasets and insufficient research on East Asian traditional musical instruments, this study constructs the East Asian Traditional Musical Instrument Dataset (EATMI-20), comprising 20 instrument categories from China, Japan, Korea, and Mongolia. Each category contains 15 original recordings, with each recording divided into five non-overlapping 10-s clips, resulting in 1500 clips. Four clips from each recording are assigned to the training set, while the remaining clip is assigned to the test set, yielding 60 training clips and 15 testing clips per category. To improve the recognition of acoustically similar instruments, a Cross-Scale Spectro-Temporal Interaction Network (CSTI-Net) is proposed. The model combines multi-resolution acoustic representations with frequency–temporal cross-attention to jointly exploit spectral and temporal characteristics and capture discriminative acoustic cues under diverse performance conditions. The main contributions of this study are summarized as follows:

1. A dedicated dataset is constructed. EATMI-20 covers 20 traditional musical instrument categories from China, Japan, Korea, and Mongolia, providing a benchmark for East Asian traditional musical instrument classification.
2. A Cross-Scale Spectro-Temporal Interaction Network is proposed. CSTI-Net integrates multi-resolution representation learning and bidirectional frequency–temporal cross-attention to enhance the discrimination of acoustically similar instruments.
3. The effectiveness and applicability of CSTI-Net are evaluated through experiments on EATMI-20 and ICKAN, with independent training and evaluation on each dataset. Ablation studies and comparisons of different front-end time-frequency representations are also conducted to assess their contributions to classification performance.

The remainder of this paper is organized as follows. Section 2 introduces the EATMI-20 dataset. Section 3 discusses the challenges associated with traditional musical instrument classification. Section 4 presents the proposed CSTI-Net architecture. Section 5 reports and discusses the experimental results, including comparative experiments, ablation studies, comparisons of different front-end time-frequency representations, and an independent evaluation of the applicability of CSTI-Net on the ICKAN dataset. Finally, Section 6 concludes the paper.

2. Dataset Description

2.1. Existing Musical Instrument Databases

Publicly available musical instrument databases have provided important benchmarks for the development and evaluation of automatic musical instrument recognition systems. However, existing datasets differ substantially in their instrument coverage and research objectives. Widely used datasets, such as IRMAS [22] and Medley-solos-DB [23], primarily focus on Western musical instruments, while the more recent ICKAN dataset [18] provides a 20-class benchmark mainly comprising Western musical instruments. In contrast, datasets involving traditional musical instruments in East Asia are relatively fragmented. For example, the Chinese Traditional Instrument Sound Database (CTIS) [24] provides acoustic recordings of Chinese traditional instruments, while the Sinawi dataset [25] focuses on Korean traditional music and includes multiple traditional instrument parts. Such resources are valuable for instrument documentation and acoustic research, but they are generally limited to a single country, region, or specific musical tradition rather than providing a unified classification benchmark.

Consequently, existing musical instrument databases either predominantly cover Western instruments or focus on traditional instruments from individual East Asian countries or specific musical contexts. A dedicated and standardized dataset covering traditional musical instruments across multiple East Asian countries remains limited. This lack of a unified cross-country resource restricts the systematic evaluation of automatic recognition methods for East Asian traditional musical instruments and motivates the construction of a dedicated dataset in this study.

2.2. EATMI-20

To address the lack of a unified dataset for traditional musical instruments across multiple East Asian countries, an East Asian Traditional Musical Instrument Dataset, termed EATMI-20, was constructed in this study. The dataset comprises 20 traditional musical instrument categories from China, Japan, Mongolia, and the Republic of Korea, covering diverse sound-production mechanisms, playing techniques, and acoustic characteristics. The database construction mainly involved two stages: data selection and collection, and data processing. An overview of the data collection and processing workflow is illustrated in Figure 1.

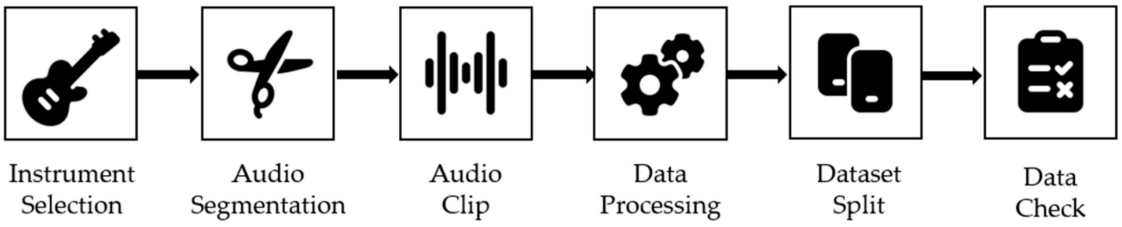


Figure 1. Overall workflow for establishing the EATMI-20 database.

2.2.1. Data Selection and Collection

The audio data were collected from publicly accessible musical recordings and other legally available audio resources. Solo performances were preferentially selected to ensure that the target instrument could be clearly identified and its characteristic timbral properties adequately represented. Recordings with prominent accompaniment, speech, or other interfering sound sources were excluded whenever possible. For each instrument category, 15 complete recordings were selected, resulting in 300 original recordings across the 20 categories.

The selected recordings covered different musical pieces and performance instances, introducing variations in pitch, playing technique, articulation, dynamics, and performance style. Such diversity provides broader acoustic coverage and reduces dependence on individual recordings or specific performance conditions. During the selection process, recordings with ambiguous instrument identities, substantial interference, obvious quality issues, or inappropriate content were excluded or replaced when necessary. In addition, data containing privacy-sensitive information, politically sensitive content, or other material unsuitable for inclusion in the database were excluded to ensure the appropriateness and reliability of the dataset.

2.2.2. Data Processing

Each original recording was divided into five non-overlapping 10-s excerpts, yielding 75 audio clips per instrument category and 1500 clips in total. The excerpts were obtained from different temporal positions within each recording to increase sample diversity while avoiding temporal overlap. All clips were standardized as mono-channel WAV files with a sampling rate of 44.1 kHz and 16-bit PCM resolution. Recordings that did not contain a valid 10-s segment were excluded, and the resulting clips were further checked for abnormal signals, excessive silence, clipping, and other apparent recording artifacts.

For dataset evaluation, five non-overlapping excerpts were extracted from different temporal positions within each original recording, with appropriate time intervals between excerpts. Each excerpt was manually inspected to ensure sufficient temporal and content diversity and to avoid excessive similarity among the selected segments. Subsequently, four excerpts from each original recording were randomly assigned to the training set, while the remaining excerpt was assigned to the test set, resulting in 60 training clips and 15 test clips per instrument category. This corresponds to an 80:20 training-to-test ratio while maintaining an equal number of samples across all 20 categories. The instrument label of each excerpt was inherited from its corresponding original recording, ensuring consistent labeling among segments derived from the same recording.

The final EATMI-20 database contains 1500 labeled audio clips covering 20 traditional musical instrument categories from four East Asian countries. The 20 instrument categories are illustrated in Table 1, while Table 2 summarizes the comparison between EATMI-20 and other representative musical instrument datasets.

Table 1. EATMI-20 categories.

China			Japan	Korea	Mongolia
Bianzhong Erhu	Daruan Guqin	Dizi Guzheng	Koto Nohkan	Gayageum	Aman Khuur
Pipa Xiao	Sheng Yangqin	Suona Zhongruan	Shakuhachi Shamisen	Geomungo	Morin Khuur

Table 2. Comparison of representative musical instrument datasets.

Dataset	Categories	Length	Sampling Rate	Region	Purpose
IRMAS	11	3 s	44.1 kHz	Western	Classification
Medley-solos-DB	8	2.972 s	44.1 kHz	Western	Classification
ICKAN	20	10 s	44.1 kHz	Western	Classification
CTIS	287	Variable	44.1 kHz	CHN	Analysis & Preservation
Sinawi	8	Variable	44.1 kHz	KOR	Analysis & Modeling
EATMI-20	20	10 s	44.1 kHz	CHN, JPN, KOR, MNG	Classification

As shown in Table 2, EATMI-20 differs from existing representative musical instrument datasets in both geographical coverage and research orientation. Widely used classification datasets have predominantly focused on Western musical instruments, whereas many existing resources for East Asian traditional instruments are limited to individual countries, regions, or specific musical traditions. In contrast, EATMI-20 is specifically designed as a unified, classification-oriented dataset covering traditional instruments from China, Japan, Mongolia, and the Republic of Korea, providing a dedicated resource for the development and evaluation of automatic recognition methods for East Asian traditional musical instruments.

3. Acoustic Characteristics and Classification Challenges

3.1. Acoustic Characteristics Analysis of East Asian Traditional Musical Instruments

The acoustic characteristics of musical instruments are closely related to their sound production mechanisms, resonant structures, performance techniques, and material properties, which jointly determine their spectral composition and time-varying acoustic responses. Instrument timbre is reflected in multiple interrelated acoustic dimensions, including spectral energy distribution, harmonic structure, and temporal dynamics. Therefore, musical instrument sounds exhibit characteristic patterns in both the frequency and temporal domains.

Figure 2 presents the log-Mel spectrograms of four representative instruments from the EATMI-20 dataset, including the Dizi as a wind instrument, the Bianzhong as a percussion instrument, and the Guzheng and Gayageum as plucked-string instruments. Instruments with different sound production mechanisms exhibit distinct time–frequency patterns in terms of energy distribution, harmonic structure, and temporal evolution. For example, the Dizi and Bianzhong show markedly different spectral and temporal characteristics, while the Guzheng and Gayageum exhibit relatively similar time–frequency patterns despite belonging to different musical traditions.

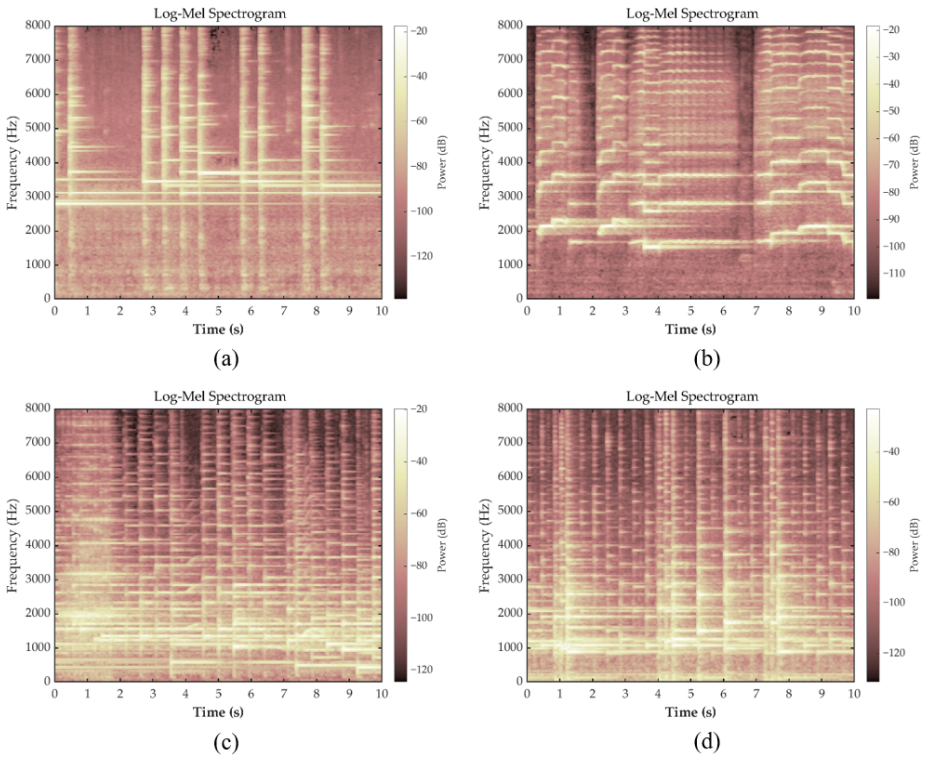


Figure 2. log-Mel spectrograms of instruments from the EATMI-20. (**a**) Bianzhong, (**b**) Dizi, (**c**) Guzheng, (**d**) Gayageum.

These observations demonstrate that traditional East Asian musical instruments exhibit diverse acoustic characteristics, while some instruments may also share considerable similarities in their time–frequency representations. Such characteristics reflect the complex relationships among spectral structure, temporal evolution, and sound production mechanisms, providing an important acoustic basis for automatic instrument classification.

3.2. Classification Challenges and Methodological Limitations

The acoustic characteristics observed in the EATMI-20 dataset introduce several challenges for automatic musical instrument classification. First, the discriminative information is distributed across different time and frequency scales. Fine-grained structures, such as transient components and localized harmonic patterns, may coexist with broader spectral and temporal characteristics. Consequently, representations based on a single receptive-field scale may not adequately capture both local acoustic details and broader contextual information [17,26].

Second, the acoustic characteristics of traditional instruments exhibit distinct frequency-dependent and time-varying properties. Spectral structures provide important information related to harmonic organization and timbre, whereas temporal evolution reflects attack, decay, resonance, articulation, and other performance-related characteristics. Modeling only one of these dimensions may therefore result in incomplete acoustic representations [4,27].

More importantly, frequency and temporal characteristics are not completely independent. The spectral composition of an instrumental sound can change over time, particularly during note onset, decay, resonance, and articulation. Thus, the discriminative information may be jointly distributed across the frequency and temporal dimensions. Recent studies have therefore explored joint time–frequency modeling and attention-based mechanisms to capture complementary information across these dimensions [26,28]. Conventional feature aggregation or independent frequency–temporal modeling may not explicitly characterize such cross-dimensional dependencies, which can limit the discrimination of acoustically similar instruments.

These challenges suggest that an effective classification framework should satisfy three requirements: it should capture acoustic patterns at multiple resolutions, model frequency-dependent and temporal characteristics in a complementary manner, and explicitly characterize their interactions. These considerations provide the basis for the design of the proposed CSTI-Net.

4. Proposed CSTI-Net

4.1. Overall Architecture of CSTI-Net

The overall architecture of the proposed Cross-Scale Spectro-Temporal Interaction Network (CSTI-Net) is illustrated in Figure 3. The network takes a log-Mel spectrogram as input and progressively transforms it into a joint frequency–temporal representation through multi-resolution feature extraction, dual-path modeling, and cross-attentive interaction, ultimately predicting 20 musical instrument classes.

First, the input spectrogram is fed into the Multi-Resolution Stem (MRS), where parallel convolutions with different receptive fields extract time–frequency patterns at multiple resolutions. The resulting features are further integrated through a Feature Pyramid to form a unified multi-resolution acoustic representation that preserves both local details and broader contextual information.

The resulting representation is then processed by two complementary paths: the Frequency Transformer (FT) and the Time ConvRNN (TC). The FT models spectral structures and long-range frequency dependencies, while the TC captures temporal dynamics and

onset-related variations. Together, the two paths provide complementary frequency- and temporal-oriented representations of instrumental sounds.

Finally, the Frequency–Temporal Cross-Attention Fusion (CAF) module establishes bidirectional interactions between the two representations and adaptively integrates their complementary information into a joint frequency–temporal embedding. The resulting representation is then fed into a lightweight classification head to produce the final predictions over the 20 instrument categories.

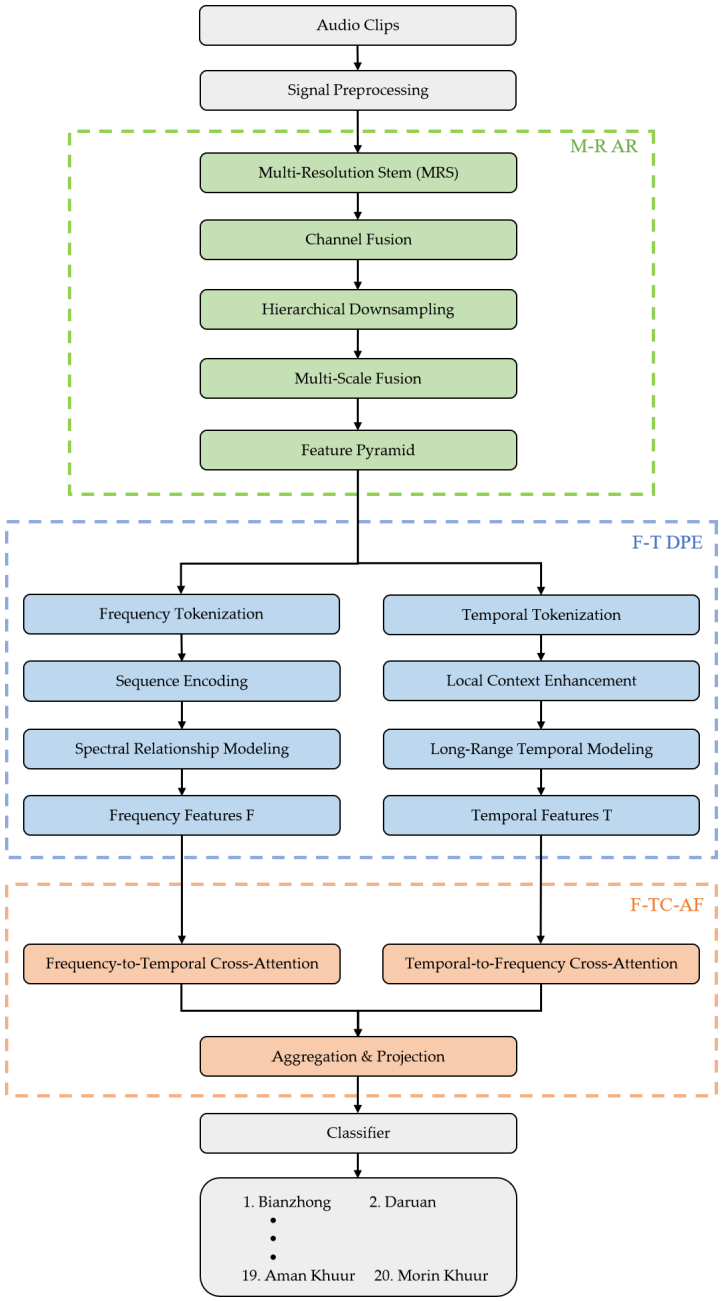


Figure 3. The overall scheme of the CSTI-Net.

4.2. Multi-Resolution Acoustic Representation

Before multi-resolution feature extraction, each audio clip is converted into a normalized log-Mel spectrogram to provide a compact time–frequency representation for CSTI-Net. The audio signals are resampled to 22.05 kHz and converted to mono when multi-channel recordings are available. Each clip is standardized to 10 s, corresponding to 220,500 samples.

The short-time Fourier transform (STFT) is then computed using a 2048-point Hamming window with a hop length of 512 samples, and the resulting spectral representation is mapped onto 128 Mel-frequency bands. Logarithmic compression is subsequently applied to reduce the dynamic range, followed by global z-score normalization on each sample. Under these settings, each audio clip is represented as a single-channel time–frequency feature with a fixed dimension of $1 \times 128 \times 431$, where the three dimensions correspond to the channel, Mel-frequency, and temporal axes, respectively.

To adequately characterize time–frequency information at different acoustic scales, CSTI-Net first constructs a multi-resolution acoustic representation. Instrumental timbre typically contains fine-grained harmonic structures, local spectral textures, and broader energy distributions, which may exhibit different characteristics across the temporal and frequency dimensions. Therefore, multi-resolution convolution and hierarchical feature fusion are employed to extract complementary acoustic information across different receptive fields and subsequently establish a unified multi-scale representation.

Given a normalized log-Mel spectrogram $X \in \mathbb{R}^{1\times F\times T}$, F and T denote the numbers of Mel-frequency bands and temporal frames, respectively, the Multi-Resolution Stem (MRS) first processes the input spectrogram through parallel convolutional branches with different receptive fields. Each branch applies a convolution at scale s, followed by batch normalization and a GELU activation, to produce the scale-specific feature X_s. The different convolutional scales provide distinct receptive fields, enabling the extraction of complementary features ranging from fine-grained local patterns to broader time–frequency contexts. The outputs of the different branches are then concatenated along the channel dimension and initially fused using a 1×1 convolution, yielding the multi-resolution feature X_{mr}. This operation integrates local time–frequency patterns extracted with different receptive fields into a common feature space, providing a shared representation for subsequent hierarchical multi-scale modeling.

The fused representation is subsequently processed by hierarchical convolutional blocks with progressive down sampling to construct multi-scale features with different spatial resolutions and contextual ranges, as formulated in Equation (1):

$$(P_1, P_2, \ldots, P_l) = Pyramid(F_{stem}) \tag{1}$$

where P_l denotes the feature at the $l - th$ pyramid level. As the network proceeds toward deeper levels, the spatial resolution gradually decreases while the contextual receptive range increases, thereby producing hierarchical acoustic representations ranging from local details to broader contextual information.

Since the feature maps at different pyramid levels have different spatial dimensions and channel sizes, each scale is first projected into a common feature space and adaptively pooled to a shared feature grid. The aligned multi-scale features are then concatenated and fused to obtain a unified shared acoustic representation, as formulated in Equation (2):

$$F_{MS} = Conv_{1\times1}(Concat[Pool(Proj(P_1), \ldots, Pool(Proj(P_l)))]) \tag{2}$$

here, $Proj(\cdot)$ denotes feature projection, $Pool(\cdot)$ represents adaptive pooling, and F_{MS} denotes the resulting multi-scale shared acoustic representation.

Through this process, CSTI-Net progressively integrates acoustic information from different convolutional receptive fields and hierarchical feature levels, preserving both local time–frequency details and broader acoustic context within a unified feature space. The resulting representation F_{MS} serves as the common input to the subsequent frequency–temporal dual-path encoding and provides a shared feature basis for further interaction between the two dimensions.

4.3. Frequency–Temporal Dual-Path Representation

Instrument identity is characterized by complementary spectral and temporal cues. Harmonic organization and spectral-envelope patterns provide discriminative frequency information, whereas attacks, decays, and modulation patterns are primarily reflected in temporal dynamics. Building upon the multi-resolution representation obtained in Section 4.2, CSTI-Net employs a dual-path representation framework to explicitly capture the complementary frequency and temporal characteristics of instrumental sounds. Specifically, the multi-resolution features are projected into two complementary token sequences and processed by dedicated frequency and temporal modeling branches, respectively. The resulting frequency and temporal representations are then integrated through the Frequency–Temporal Cross-Attention Fusion module. As illustrated in Figure 4, this module employs bidirectional cross-attention to enable information exchange between the two representations through cross-axis similarity scoring, weighted aggregation, and residual refinement. Finally, the refined representations are subjected to sequence-level mean pooling, concatenated, and linearly projected to obtain the final joint embedding.

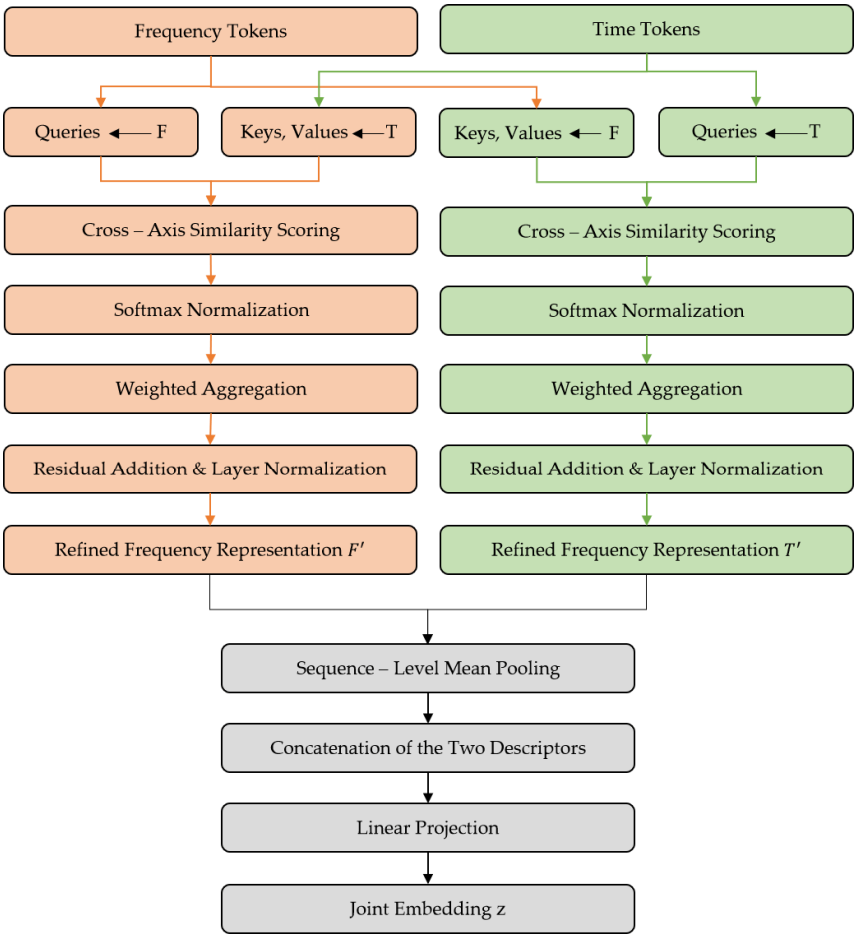


Figure 4. Architecture of the Frequency–Temporal Cross-Attention Fusion module.

Let the multi-resolution feature representation obtained from Section 4.2 be denoted as $F_{MS} \in \mathbb{R}^{C \times F \times T}$, C, F, and T denote the channel, frequency, and temporal dimensions, respectively. To construct complementary frequency- and temporal-oriented representations from the shared acoustic feature, the feature is aggregated along the temporal and frequency dimensions, respectively.

The frequency-oriented representation is obtained by temporal aggregation, as formulated in Equation (3):

$$X_f(d,f) = \frac{1}{T}\sum_{t=1}^{T} F_{MS}(d,f,t) \tag{3}$$

where d, f, and t represent the channel, frequency, and temporal indices, respectively. Each frequency token therefore summarizes the acoustic characteristics associated with a specific frequency position across the entire temporal dimension.

Conversely, the temporal-oriented representation is obtained by frequency aggregation, as formulated in Equation (4):

$$X_t(d,t) = \frac{1}{F}\sum_{f=1}^{F} F_{MS}(d,f,t) \tag{4}$$

Each temporal token consequently characterizes the acoustic state at a specific temporal position while integrating information across the frequency dimension.

This dual-view aggregation strategy decomposes the two-dimensional time-frequency representation into two complementary one-dimensional marginal views. Temporal averaging preserves the global energy distribution across frequency bins, highlighting the harmonic structure associated with instrument timbre, whereas frequency averaging summarizes the overall energy variation across time frames, capturing temporal envelope characteristics such as onset, sustain, and decay. This decomposition enables the two branches to focus on complementary acoustic properties while reducing the computational burden compared with processing the full two-dimensional representation.

The two representations are then projected into a common D-dimensional embedding space. The frequency-oriented sequence is processed by a Transformer encoder to capture global dependencies among spectral components, while the temporal-oriented sequence is first enhanced using a depthwise one-dimensional convolution and subsequently modeled by a bidirectional GRU to characterize local temporal variations and bidirectional temporal dependencies [29,30].

4.3.1. Frequency Aggregation

The frequency-oriented representation is designed to capture global dependencies among spectral components. The fused feature map is first transposed along the temporal dimension and projected into a common D-dimensional embedding space using a learnable projection matrix. A learnable classification token is then prepended to the projected sequence, followed by the addition of learnable positional embeddings. The resulting sequence is processed by a Transformer encoder to model long-range dependencies across frequency tokens.

For the l-th Transformer encoder layer, a standard pre-normalized Transformer structure with multi-head self-attention (MSA), feed-forward network (FFN), and residual connections is employed, as formulated in Equation (5):

$$\tilde{H}_f^{l-1} = H_f^{l-1} + MSA_f^l\left(LN\left(H_f^{l-1}\right)\right) \quad H_f^l = \tilde{H}_f^l + FFN_f^l\left(LN\left(\tilde{H}_f^l\right)\right) \tag{5}$$

where H_f^{l-1} denotes the input representation of the l-th encoder layer, and LN, MSA, and FFN denote layer normalization, multi-head self-attention, and feed-forward network, respectively. The classification token is removed after the final Transformer layer, yielding the frequency-aware representation $H_f \in \mathbb{R}^{F\times D}$.

4.3.2. Temporal Aggregation

In parallel, the temporal-oriented representation is constructed to characterize local temporal variations and longer-range temporal dependencies. The fused feature map is

transposed along the frequency dimension and projected into the same D-dimensional embedding space. Before recurrent modeling, a depthwise one-dimensional convolution followed by a GELU activation is applied along the temporal dimension, as formulated in Equation (6):

$$\tilde{Z}_t = GELU\left(DWConv_{1D}\left(X_t W_t^E\right)\right),\ W_t^E \in \mathbb{R}^{C\times D} \tag{6}$$

The resulting sequence is then processed by a bidirectional GRU to capture temporal dependencies from both forward and backward directions. The hidden states from the two directions are concatenated and normalized to obtain the temporal representation, as formulated in Equation (7):

$$H_t = LN\left(\vec{H}_t; \overleftarrow{H}_t\right) \in \mathbb{R}^{T\times D} \tag{7}$$

Here, $\vec{H}_t$ and $\overleftarrow{H}_t$ denote the forward and backward hidden states, respectively. The hidden dimension of each direction is set to D/2, ensuring that their concatenation remains in the common D-dimensional embedding space.

The frequency and temporal pathways therefore encode complementary characteristics within a unified representation space. Specifically, the frequency path emphasizes global spectral dependencies, while the temporal path focuses on local temporal variations and bidirectional temporal evolution. The resulting representations H_f and H_t are subsequently supplied to the proposed cross-attention module for adaptive information exchange.

4.4. Frequency–Temporal Cross-Attention Fusion

Although the frequency and temporal pathways provide complementary acoustic representations, they are initially encoded along different dimensions. Direct concatenation would combine these representations without explicitly modeling their interdependencies. To address this limitation, a Frequency–Temporal Cross-Attention Fusion (CAF) module is introduced to establish adaptive interactions between the two representations.

Given the frequency and temporal representations $H_f \in \mathbb{R}^{F\times D}$ and $H_t \in \mathbb{R}^{T\times D}$, CAF employs two bidirectional cross-attention branches. In the frequency-to-temporal branch, the frequency representation serves as the query, while the temporal representation provides the key and value, as formulated in Equation (8):

$$C_{f\to t} = softmax\left(\frac{\left(H_f W_f^Q\right)\left(H_t W_t^K\right)^T}{\sqrt{d_h}}\right) H_t W_t^V \tag{8}$$

where W_f^Q, W_t^K, and W_t^V are learnable projection matrices, and d_h denotes the attention-head dimension. The attended temporal information is projected back to the common embedding dimension and integrated with the original frequency representation through a residual connection, as formulated in Equation (9):

$$H_f^{out} = LN\left(H_f + C_{f\to t} W_f^o\right) \tag{9}$$

Conversely, the temporal-to-frequency branch reverses the query, key, and value roles, as formulated in Equation (10):

$$H_t^{out} = LN\left[H_t + softmax\left(\frac{\left(H_t W_t^o\right)\left(H_f W_f^K\right)^T}{\sqrt{d_h}}\right) H_f W_f^V W_t^o\right] \tag{10}$$

Here, $W_f^o, W_t^o \in \mathbb{R}^{d_h\times D}$ project the attended features back to the common D-dimensional embedding space. Through the two symmetric branches, each pathway can selectively incorporate information from the other according to feature correlations. Consequently,

the frequency representation is enhanced by temporally relevant cues, while the temporal representation is simultaneously refined by frequency-dependent information.

To obtain a compact joint representation, the two enhanced representations are first aggregated along their respective sequence dimensions using average pooling, as formulated in Equation (11):

$$h_f = \frac{1}{F}\sum_{i=1}^{F} H_f^{out}[i] \;\; h_t = \frac{1}{T}\sum_{j=1}^{T} H_t^{out}[j] \tag{11}$$

The pooled representations are then concatenated and projected into the final D-dimensional joint embedding, as formulated in Equation (12):

$$z = \left[h_f; h_t\right] W_z + b_z \;\; W_z \in \mathbb{R}^{2D \times D} \tag{12}$$

The resulting embedding z integrates spectral structure and temporal dynamics through correlation-aware bidirectional interaction and is subsequently passed to the classification head for musical instrument prediction.

4.5. Classification and Training

The joint representation (z) generated by the Frequency–Temporal Cross-Attention Fusion module is subsequently fed into the classification head. The classification head maps the joint representation to class logits corresponding to the (M) instrument categories, followed by a Softmax operation to obtain the predicted class probabilities. The category with the highest predicted probability is selected as the final classification result.

CSTI-Net is trained in an end-to-end manner using a classification-oriented optimization objective. The training objective combines cross-entropy loss with label smoothing and supervised contrastive loss, enabling the network to simultaneously optimize classification accuracy and the discriminative structure of the learned feature representations.

5. Experimental and Result Analysis

To comprehensively evaluate the effectiveness of the proposed CSTI-Net for traditional musical instrument classification, experiments were conducted from several perspectives using two datasets. First, CSTI-Net was evaluated on the EATMI-20 dataset and compared with representative methods under consistent experimental conditions. Second, CSTI-Net was independently trained and evaluated on the ICKAN dataset to investigate its applicability to different musical instrument datasets with distinct instrument categories and acoustic characteristics. Furthermore, ablation experiments were conducted to examine the contributions of individual network components, while comparisons of different front-end time-frequency representations were performed to determine their effects on classification performance. Based on these experiments, the effectiveness and applicability of the proposed method are analyzed and discussed.

5.1. Experimental Setup

All experiments were conducted in a Python 3.10 environment using PyTorch 2.6.0, torchaudio, scikit-learn, and SciPy. Model training and evaluation were performed on a single NVIDIA GeForce GTX 1660 Ti GPU with 6 GB of VRAM under CUDA 12.4. The input features were generated according to the audio preprocessing and log-Mel spectrogram extraction procedure described in Section 4.2. Unless otherwise specified, CSTI-Net and all comparison methods were implemented and evaluated using the same dataset partitions, input representations, and evaluation protocols to ensure a fair comparison.

CSTI-Net was trained for 120 epochs with a batch size of 8 using the AdamW optimizer. The initial learning rate was set to 1.5×10^{-3}, and the weight decay was set to 5×10^{-4}. A

linear warm-up strategy was employed during the first 10 epochs, followed by cosine annealing for the remaining training epochs. The training objective consisted of cross-entropy loss with a label-smoothing factor of 0.1 and supervised contrastive loss with a weight of 0.2. The supervised contrastive loss was activated after the warm-up stage. In addition, gradient clipping with a maximum norm of 1.0 was applied to improve optimization stability.

For the comparative experiments, the baseline methods were trained and evaluated under the same data partitioning and preprocessing conditions as CSTI-Net. Unless a method-specific configuration was required by the original architecture, the training procedure and evaluation protocol were kept consistent across different models. The final classification performance was evaluated on the test sets using the selected evaluation metrics.

5.2. Evaluation Metrics

To evaluate the classification performance of CSTI-Net, six metrics are adopted: accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, micro-averaged F1-score, and average Pearson correlation coefficient [31,32]. For the $M = 20$ instrument categories, TP_m, FP_m, and FN_m denote the true-positive, false-positive, and false-negative counts of category m, respectively. N and $N_{correct}$ denote the total number of test samples and correctly classified samples, respectively. The evaluation metrics are calculated as follows:

1. Accuracy measures the proportion of correctly classified samples among all test samples, providing an overall measure of classification performance, as formulated in Equation (13):

$$A = \frac{N_{correct}}{N} \times 100\% \tag{13}$$

2. Precision measures the proportion of correctly predicted samples among all samples assigned to a specific class, reflecting the reliability of the model's predictions, as formulated in Equation (14):

$$P_{macro} = \frac{1}{M}\sum_{m=1}^{M} \frac{TP_m}{TP_m + FP_m} \times 100\% \tag{14}$$

3. Recall measures the proportion of correctly identified samples among all actual samples of a specific class, indicating the model's ability to detect samples belonging to each instrument category, as formulated in Equation (15):

$$R_{macro} = \frac{1}{M}\sum_{m=1}^{M} \frac{TP_m}{TP_m + FN_m} \times 100\% \tag{15}$$

4. Macro-F1 is obtained by calculating the F1-score for each class and averaging the results with equal weights. It provides a balanced evaluation by assigning equal importance to all instrument categories, as formulated in Equation (16):

$$F1_{macro} = \frac{1}{M}\sum_{m=1}^{M} \frac{2P_m R_m}{P_m + R_m} \times 100\% \tag{16}$$

5. Micro-F1 is calculated by aggregating the true positives, false positives, and false negatives across all classes before computing the F1-score. It reflects the overall classification performance across all samples, as formulated in Equation (17):

$$F1_{micro} = \frac{2\sum_{m=1}^{M} TP_m}{2\sum_{m=1}^{M} TP_m + \sum_{m=1}^{M} FP_m + \sum_{m=1}^{M} FN_m} \times 100\% \tag{17}$$

6. Pearson (r) measures the linear correlation between the overall predicted results and the corresponding ground-truth results. A higher value indicates stronger agreement between the predictions and the ground truth, as formulated in Equation (18):

$$r_{avg} = \frac{1}{M}\sum_{m=1}^{M} Pearson(P_m, y_m) \quad (18)$$

where P_m and y_m denote the predicted probability vector and one-hot encoded ground-truth label vector for category m, respectively.

To ensure reliable and reproducible evaluation, all experiments are independently conducted using three random seeds (42, 1234, and 2024). The reported results are presented as the mean ± standard deviation (SD) across the three runs. All methods and configurations are evaluated under the same experimental settings to ensure fair comparisons.

In addition, a confusion matrix is employed to visualize the class-wise prediction distribution and to identify instrument pairs that are frequently misclassified.

5.3. Results on EATMI-20

To comprehensively assess the classification performance of CSTI-Net, experiments are conducted on the EATMI-20, which comprises 20 instrument categories. The performance is quantitatively evaluated using six metrics: accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, micro-averaged F1-score, and the average Pearson correlation coefficient.

For comparative evaluation, CSTI-Net is compared with six representative musical instrument classification methods, namely CNN-Small, CNN-Medium, CNN-Large, ICKAN [18], MIR-CRNN [33], and InstruNET [34]. The CNN-Small, CNN-Medium, CNN-Large, and ICKAN models are implemented according to the configurations reported in the original study [18], while the MIR-CRNN model is implemented based on the publicly available source code [33]. To ensure a fair comparison, all methods are evaluated using the same dataset partitioning, preprocessing procedures, and experimental settings. The comparative results are reported in Table 3.

Table 3. Comparison of Classification Performance on the EATMI-20.

Methods	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Micro-F1 (%)	Pearson r
CNN-Small	80.44 ± 2.99	80.92 ± 2.66	80.44 ± 2.99	80.19 ± 3.19	80.44 ± 2.99	0.8255 ± 0.0232
CNN-Middle	84.33 ± 0.94	84.84 ± 0.88	84.33 ± 0.94	84.23 ± 0.94	84.33 ± 0.94	0.8632 ± 0.0060
CNN-Large	84.11 ± 1.37	85.47 ± 1.34	84.11 ± 1.37	84.15 ± 1.37	84.11 ± 1.37	0.8644 ± 0.0069
ICKAN	87.22 ± 1.10	87.52 ± 1.16	87.22 ± 1.10	87.11 ± 1.15	87.22 ± 1.10	0.8889 ± 0.0134
MIR-CRNN	86.44 ± 2.20	87.30 ± 2.04	86.44 ± 2.20	86.24 ± 2.20	86.44 ± 2.20	0.8861 ± 0.0084
InstruNET	86.22 ± 1.40	86.62 ± 1.32	86.22 ± 1.40	84.81 ± 1.51	86.22 ± 1.40	0.8711 ± 0.0120
CSTI-Net	**91.67 ± 1.09**	**92.41 ± 0.75**	**91.67 ± 1.09**	**91.66 ± 0.94**	**91.67 ± 1.09**	**0.9267 ± 0.0055**

Note: The best performance in each evaluation metric is highlighted in **bold**.

As shown in Table 3, CSTI-Net achieves the highest mean Accuracy, macro-averaged Precision, macro-averaged Recall, Macro-F1, Micro-F1, and Pearson correlation coefficient among all seven methods, reaching 91.67%, 92.41%, 91.67%, 91.66%, 91.67%, and 0.9267, respectively. Compared with ICKAN, which achieves the best overall performance among the baseline methods, CSTI-Net improves the mean Accuracy by 4.45 percentage points, Precision by 4.89 percentage points, Recall by 4.45 percentage points, Macro-F1 by 4.55 percentage points, and Micro-F1 by 4.45 percentage points. Furthermore, CSTI-Net achieves a higher mean Pearson correlation coefficient than ICKAN (0.9267 vs. 0.8889), representing an absolute improvement of 0.0378. Among the three CNN baselines, CNN-Middle

achieves the highest mean Accuracy of 84.33%, slightly outperforming CNN-Large (84.11%), indicating that increasing the CNN model scale does not necessarily lead to improved classification performance. In addition, InstruNET achieves a mean Accuracy of 86.22%, while MIR-CRNN obtains 86.44%. These results demonstrate that CSTI-Net achieves superior average classification performance compared with the conventional CNN-based and existing musical instrument recognition methods considered in this study.

Overall, CSTI-Net consistently outperforms all six baseline methods across the reported evaluation metrics. The improvements in both classification accuracy and F1-based metrics demonstrate its effectiveness in distinguishing different East Asian traditional musical instruments. These results further support the effectiveness of the proposed multi-resolution frequency–temporal modeling strategy in capturing discriminative acoustic characteristics for musical instrument classification.

To further examine the class-level classification behavior of CSTI-Net, the normalized confusion matrix obtained from the experiment using random seed 42 is presented in Figure 5. The predictions are predominantly concentrated along the diagonal, demonstrating that the proposed method can effectively distinguish the majority of instrument categories. Specifically, 14 of the 20 instrument categories achieve a recall of 93.3%, another four categories reach 86.7%, one category obtains 80.0%, and only one category has a recall of 73.3%. Given that each category contains 15 test samples, these results correspond to an overall classification accuracy of 90.33% (271 out of 300 correctly classified samples) for the seed 42 experiment. The limited off-diagonal values further demonstrate that misclassifications are generally sparse and concentrated among certain instrument categories. For example, the most prominent misclassifications occur from Suona to Morin Khuur, from Koto to Zhongruan, and from Geomungo to Gayageum, each with a misclassification rate of 0.13. These relatively localized confusions do not substantially affect the overall classification performance. Instead, the strong diagonal concentration and high recall across most categories demonstrate that CSTI-Net can effectively capture complementary frequency–temporal characteristics and maintain discrimination among diverse East Asian traditional musical instruments. Overall, the confusion matrix provides further evidence that the proposed method achieves high classification accuracy and consistent class-level recognition performance on the EATMI-20 dataset.

5.4. Ablation Study

To further investigate the contributions of different components to the classification performance of CSTI-Net, a series of ablation experiments is conducted on the EATMI-20 dataset. Specifically, the effects of different time-frequency representations and key architectural components are examined by modifying or removing the corresponding elements while keeping the remaining experimental settings unchanged. These experiments aim to assess the influence of individual design choices on classification performance and to provide a more comprehensive understanding of the effectiveness of the proposed CSTI-Net architecture.

5.4.1. Effect of Different Time-Frequency Representations

To investigate the influence of different time-frequency representations on the classification performance of CSTI-Net, four input representations are evaluated: Mel spectrogram, multi-window short-time Fourier transform (STFT), constant-Q transform (CQT), and continuous wavelet transform (CWT). In each experiment, the original Mel spectrogram input is replaced with one of the alternative representations, while the network architecture, dataset partitioning, training strategy, and other experimental settings remain unchanged. This experimental design enables a controlled comparison of the effects of different time-frequency representations on classification performance. The comparative results are presented in Table 4.

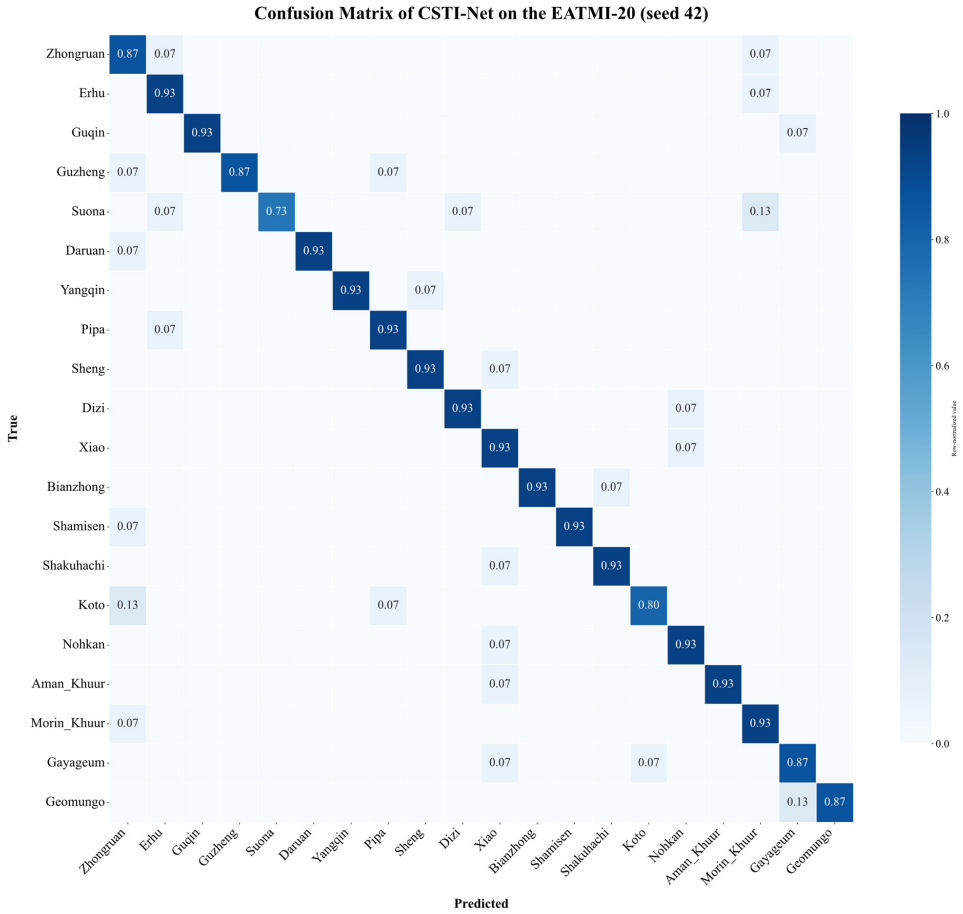


Figure 5. The Confusion Matrix of CSTI-Net on the EATMI-20 Dataset.

Table 4. Performance comparison of different time-frequency representations using CSTI-Net.

Representation	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Micro-F1 (%)	Pearson r
Multi-window STFT	90.55 ± 1.10	**92.64 ± 0.63**	90.55 ± 1.10	90.63 ± 1.06	90.55 ± 1.10	0.9222 ± 0.0087
CQT	86.44 ± 1.64	89.19 ± 1.90	86.44 ± 1.64	86.17 ± 1.83	86.44 ± 1.64	0.8802 ± 0.0176
CWT	84.22 ± 1.10	85.32 ± 1.25	84.22 ± 1.10	84.01 ± 1.40	84.22 ± 1.10	0.8600 ± 0.0096
Mel spectrogram (Proposed)	**91.67 ± 1.09**	92.41 ± 0.75	**91.67 ± 1.09**	**91.66 ± 0.94**	**91.67 ± 1.09**	**0.9267 ± 0.0055**

Note: The best performance in each evaluation metric is highlighted in **bold**.

As shown in Table 4, the Mel spectrogram achieves the highest mean performance in five of the six reported evaluation metrics, including Accuracy (91.67%), Recall (91.67%), Macro-F1 (91.66%), Micro-F1 (91.67%), and Pearson correlation coefficient (0.9267). In contrast, multi-window STFT achieves the highest mean Precision of 92.64%, slightly exceeding that of the Mel spectrogram (92.41%). Nevertheless, the Mel spectrogram achieves higher mean Accuracy (91.67% vs. 90.55%), Macro-F1 (91.66% vs. 90.63%), and Pearson correlation coefficient (0.9267 vs. 0.9222) than multi-window STFT. CQT achieves a mean Accuracy of 86.44% and a Pearson correlation coefficient of 0.8802, whereas CWT exhibits the lowest mean performance, with an Accuracy of 84.22% and a Pearson correlation coefficient of 0.8600. Overall, these results indicate that the Mel spectrogram provides the most favorable overall classification performance among the four evaluated representations, although multi-window STFT achieves a slightly higher mean Precision.

Overall, the results demonstrate that the choice of time-frequency representation influences the classification performance of CSTI-Net. Although multi-window STFT achieves a

slightly higher mean Precision and a Pearson correlation coefficient close to that of the Mel spectrogram, the Mel spectrogram achieves the highest mean values in five of the six evaluation metrics. Considering its overall performance across the evaluated classification metrics, the Mel spectrogram is retained as the input representation in the proposed CSTI-Net.

5.4.2. Ablation of Network Components

To further investigate the contributions of individual components to CSTI-Net, an ablation study is conducted using six configurations: the full model (Full) and five variants, each with one component removed (No MultiRes, No Time Path, No Freq Path, No SupCon, and No Cross Attn). Specifically, the five variants are obtained by removing the MultiResStem module, the time-domain pathway, the frequency-domain pathway, the supervised contrastive learning objective, and the cross-attention module, respectively. In each configuration, only the corresponding component is removed, while all other architectural components and experimental settings remain unchanged. This controlled design allows the contribution of each component to be assessed individually. The comparative results are presented in Table 5.

Table 5. Ablation results of different network components in CSTI-Net.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Micro-F1 (%)	Pearson r
No MultiRes	80.89 ± 1.91	83.87 ± 1.73	80.89 ± 1.91	81.19 ± 2.17	80.89 ± 1.91	0.8392 ± 0.0203
No Time Path	84.89 ± 1.81	86.54 ± 1.10	84.89 ± 1.81	84.63 ± 1.73	84.89 ± 1.81	0.8588 ± 0.0128
No Freq Path	89.22 ± 1.57	90.39 ± 1.03	89.22 ± 1.57	89.26 ± 1.34	89.22 ± 1.57	0.9110 ± 0.0133
No SupCon	84.00 ± 3.92	87.08 ± 2.92	84.00 ± 3.92	84.35 ± 3.95	84.00 ± 3.92	0.8607 ± 0.0315
No Cross Attn	87.11 ± 2.47	89.32 ± 2.10	87.11 ± 2.47	87.39 ± 2.46	87.11 ± 2.47	0.8927 ± 0.0315
Full	**91.67 ± 1.09**	**92.41 ± 0.75**	**91.67 ± 1.09**	**91.66 ± 0.94**	**91.67 ± 1.09**	**0.9267 ± 0.0055**

Note: The best performance in each evaluation metric is highlighted in **bold**.

As shown in Table 5, the full CSTI-Net achieves the highest mean performance across all six evaluation metrics, with an Accuracy of 91.67% and a Macro-F1 score of 91.66%. Among the ablated configurations, removing the MultiResStem module causes the largest performance degradation, reducing Accuracy and Macro-F1 to 80.89% and 81.19%, respectively. This highlights the importance of multi-scale feature extraction for capturing discriminative acoustic characteristics.

Removing supervised contrastive learning and the time-domain pathway also leads to considerable performance decreases, with Accuracy dropping to 84.00% and 84.89%, respectively. Similarly, removing the cross-attention module reduces Accuracy to 87.11%, indicating its contribution to feature interaction and integration. In comparison, removing the frequency-domain pathway results in a relatively smaller decrease, with Accuracy reaching 89.22%.

Overall, the ablation results demonstrate that all five components contribute to the classification performance of CSTI-Net, although their contributions vary in magnitude. Their integration enables the full model to achieve the best overall performance among the evaluated configurations.

5.5. Results on ICKAN Dataset

To further evaluate the applicability of CSTI-Net to different musical instrument datasets, additional experiments are conducted on the existing musical instrument dataset adopted in the ICKAN study. In contrast to EATMI-20, this dataset primarily consists of Western musical instruments, exhibiting substantial differences in instrument categories, timbral characteristics, and acoustic properties. It therefore provides an independent eval-

uation setting with a different instrument domain from that of EATMI-20, allowing the applicability of the proposed network architecture to a different musical instrument classification task to be investigated.

The dataset is selected for this evaluation primarily because of its relatively large sample size and representative coverage of commonly used Western musical instruments. Moreover, it has been adopted in previous musical instrument classification studies, making it a suitable benchmark for comparative assessment. More importantly, the substantial differences between the Western instruments covered by this dataset and the East Asian traditional instruments in EATMI-20 provide an opportunity to examine the applicability of the proposed network architecture to different musical instrument classification tasks.

To ensure a fair comparison, CSTI-Net was independently trained from scratch and evaluated on the ICKAN dataset using the same instrument categories, training and test data, and corresponding experimental settings as those adopted in the original ICKAN study. Furthermore, all baseline models, including ICKAN, CNN-Medium, CNN-Small, CNN-Large, and MLP, are independently trained and evaluated under the same experimental conditions rather than directly adopting the performance values reported in their original publications. The resulting performance comparison is presented in Table 6.

Table 6. Overall Classification Results on the ICKAN Dataset.

Methods	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Micro-F1 (%)
ICKAN	95.82 ± 0.35	96.51 ± 0.53	95.88 ± 0.34	96.02 ± 0.40	95.82 ± 0.35
CNN Medium	92.89 ± 0.45	96.34 ± 0.40	92.98 ± 0.44	93.93 ± 0.21	92.89 ± 0.45
CNN Small	93.31 ± 0.40	95.63 ± 0.76	93.39 ± 0.39	93.99 ± 0.45	93.31 ± 0.40
CNN Large	93.83 ± 0.50	96.07 ± 0.17	93.91 ± 0.49	94.50 ± 0.32	93.83 ± 0.50
MLP [35]	90.46 ± 0.02	93.82 ± 0.60	90.54 ± 0.03	91.41 ± 0.26	90.46 ± 0.02
CSTI-Net	**96.66 ± 0.30**	**96.82 ± 0.24**	**96.74 ± 0.30**	**96.68 ± 0.34**	**96.66 ± 0.30**

Note: The best performance in each evaluation metric is highlighted in **bold**.

As shown in Table 6, CSTI-Net achieves the best performance across all five evaluation metrics, with an accuracy of 96.66%, precision of 96.82%, recall of 96.66%, Macro-F1 score of 96.68%, and Micro-F1 score of 96.66%. Compared with ICKAN, which achieves the second-best overall classification performance among the evaluated methods, CSTI-Net improves accuracy, precision, recall, Macro-F1, and Micro-F1 by 0.84, 0.31, 0.78, 0.66, and 0.84 percentage points, respectively. Compared with the CNN-based baselines, CSTI-Net achieves accuracy improvements ranging from 2.83 to 3.77 percentage points, while the improvement over the MLP baseline reaches 6.20 percentage points. These results demonstrate that CSTI-Net maintains strong classification performance when independently trained and evaluated on a Western musical instrument dataset, indicating that the proposed architecture is also applicable to a certain extent to other musical instrument datasets.

It is also notable that the performance differences for several individual instruments are relatively small. For example, Harmonica, Piano, Trombone, and Xylophone exhibit comparable classification performance between CSTI-Net and ICKAN. These results indicate that CSTI-Net maintains competitive performance on instrument categories that are already well recognized by ICKAN while achieving more substantial improvements on challenging categories. In particular, noticeable performance gains over ICKAN are observed for Trumpet, Vibraphone, Celesta, Oboe, Marimba, and Guitar. This suggests that the proposed multi-resolution frequency–temporal representation is effective in capturing discriminative acoustic characteristics, particularly for instruments with complex or overlapping spectral structures.

The confusion matrix of CSTI-Net on the ICKAN dataset is further presented in Figure 6 to provide a class-level view of its classification performance on this independent dataset. The predictions are highly concentrated along the diagonal, with 12 of the 20 instrument categories achieving recall values of at least 97%. Even the lowest-performing category, Double Bass, achieves a recall of 87.3%, while most categories maintain recall above 90%. The strongest classification performance is observed for instruments such as Accordion, Celesta, Guitar, Harmonica, Piano, Trombone, and Xylophone, which achieve approximately 99.7% recall. The most evident confusion occurs between Vibraphone and Oboe, with a misclassification rate of 0.055, reflecting the difficulty of distinguishing certain instruments with overlapping acoustic characteristics. Double Bass shows more dispersed misclassifications across several categories, but its F1-score remains 0.9292, indicating that its overall predictive performance is still relatively high. Overall, the strong diagonal concentration and high recall across most categories are consistent with the results in Table 6 and further demonstrate the applicability of the proposed architecture to a substantially different musical instrument dataset.

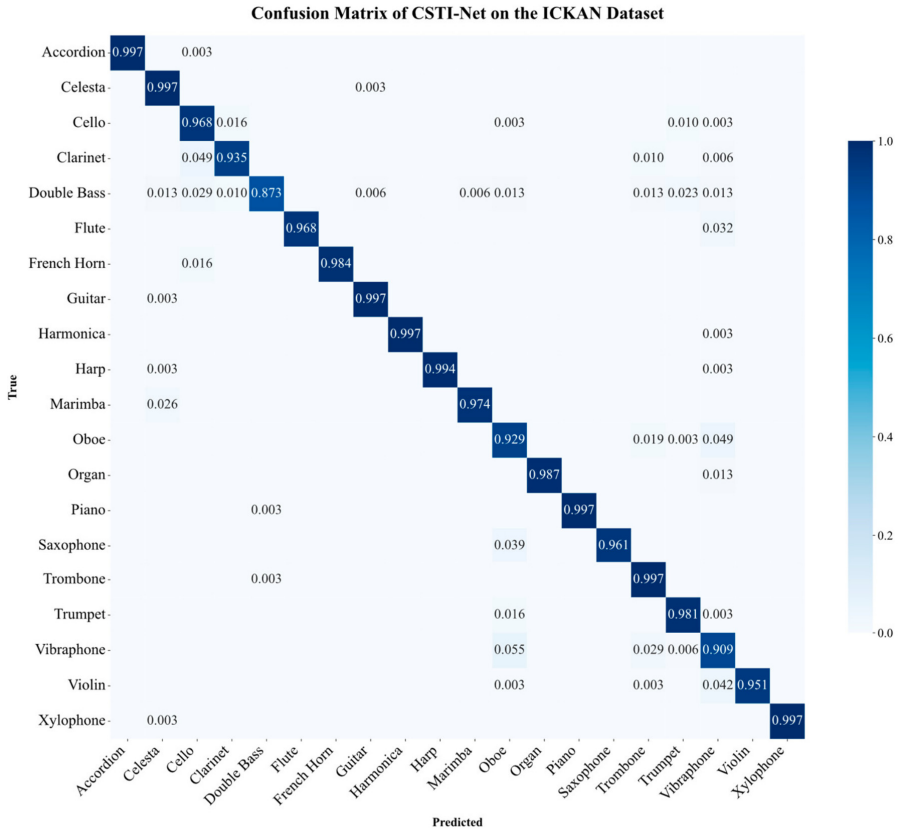


Figure 6. The Confusion Matrix of CSTI-Net on the ICKAN Dataset.

5.6. Discussion

The experimental results demonstrate the effectiveness of CSTI-Net in musical instrument classification. On the EATMI-20 dataset, CSTI-Net achieves the best overall performance among the compared methods, with balanced precision, recall, and F1-scores. The class-level results show that most instrument categories are reliably recognized, while the remaining misclassifications mainly occur between instruments with similar timbral and acoustic characteristics.

The additional evaluation on the independent ICKAN dataset further demonstrates the applicability of CSTI-Net to different musical instrument datasets. Despite the differences in instrument categories and acoustic characteristics between East Asian traditional

and Western musical instruments, CSTI-Net maintains the highest overall classification performance when independently trained and evaluated on the ICKAN dataset. These results suggest that the proposed network architecture is also applicable to a certain extent to other musical instrument classification tasks. Furthermore, the ablation study verifies the effectiveness of the individual network components and their complementary roles in the overall architecture. The results indicate that multi-scale feature extraction, supervised contrastive learning, and the time-domain pathway contribute to the learning of discriminative representations, while the cross-attention module facilitates the integration of frequency- and temporal-domain features. These components work together to enhance the model's classification performance.

The comparison of different time-frequency representations shows that the Mel spectrogram achieves the best performance across all evaluation metrics, outperforming multi-window STFT, CQT, and CWT. Therefore, it is retained as the input representation of CSTI-Net. Overall, the results support the effectiveness of combining multi-scale feature extraction, complementary frequency–temporal modeling, supervised contrastive learning, and cross-attention for musical instrument classification. The consistent performance advantages on EATMI-20 and the independently trained ICKAN evaluation further support the effectiveness of CSTI-Net and its applicability, to a certain extent, to different musical instrument datasets.

6. Conclusions

Automatic musical instrument classification is an important task in music information retrieval and computational music analysis. To address the limited research on East Asian traditional musical instruments, this study constructed a 20-class East Asian Traditional Musical Instrument Dataset (EATMI-20) and proposed a Cross-Scale Spectro-Temporal Interaction Network (CSTI-Net) for automatic instrument classification.

CSTI-Net integrates multi-resolution feature extraction with complementary frequency and temporal representation learning. The proposed frequency–temporal cross-attention mechanism enables bidirectional interaction between frequency- and temporal-domain representations, allowing complementary acoustic information to be jointly exploited. This design provides a unified representation framework for capturing the complex spectral and temporal characteristics of East Asian traditional musical instruments.

Experimental results on EATMI-20 demonstrate the effectiveness of the proposed approach. CSTI-Net achieves an accuracy of 91.67% and obtains the best overall performance among the compared methods in terms of accuracy, precision, recall, and F1-score. Furthermore, independent training and evaluation on a Western musical instrument dataset achieve an accuracy of 96.66%, indicating that CSTI-Net is also applicable to a certain extent to different musical instrument datasets. Additional ablation experiments confirm the contributions of the individual network components, while comparisons of different time-frequency representations demonstrate the influence of input feature selection on classification performance and support the use of Mel spectrograms in CSTI-Net.

Overall, this study provides both a dedicated dataset and an effective deep learning framework for the automatic classification of East Asian traditional musical instruments. The experimental findings support the effectiveness of multi-resolution feature extraction, joint frequency–temporal representation learning, and component integration in modeling diverse instrumental acoustics. Future work will focus on expanding the scale and diversity of traditional musical instrument datasets, incorporating richer acoustic representations, and investigating more general cross-domain and cross-cultural musical instrument classification methods.

Author Contributions: S.K.: supervision, conceptualization—review & editing. C.-W.K.: methodology, resources, writing—original draft. B.D.: validation, formal analysis, investigation, software, Y.S.: data curation, visualization. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Technology Innovation Program (RS-2024-00507664, Development of collaborative robot-based AI autonomous manufacturing system for application to ship assembly process) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The EATMI-20 dataset proposed in this study was specifically constructed for the classification of traditional musical instruments from East Asia. To comply with privacy regulations, copyright restrictions, and ethical considerations, the dataset is not publicly available for direct download. However, restricted access may be granted upon request for academic research purposes. Researchers interested in accessing the dataset may contact Bo Dong via email at sebastian@mail.ulsan.ac.kr to request access and obtain the download link.

Conflicts of Interest: The authors declare that there is no conflict of interest or personal relationship.

Abbreviations

The following abbreviations are used in this manuscript:

MIR	Music Information Retrieval
CNNs	Convolutional Neural Networks
PANNs	Pretrained Audio Neural Networks
EATMI-20	East Asian Traditional Musical Instrument Dataset (20 instrument categories)
CSTI-Net	Cross-Scale Spectro-Temporal Interaction Network
CTIS	Chinese Traditional Instrument Sound Database
MRS	Multi-Resolution Stem
FT	Frequency Transformer
ConvRNN	Convolutional Recurrent Neural Network
TC	Time ConvRNN
CAF	Cross-Attention Fusion
STFT	Short-Time Fourier Transform
GELU	Gaussian Error Linear Unit
GRU	Gated Recurrent Unit
MSA	Multi-Head Self-Attention
FFN	Feed-Forward Network
LN	Layer Normalization
MLP	Multi-Layered Perceptron

References

1. Ando, Y. Several Peculiar Structural Devices of Japanese Musical Instruments and Some Aspects of Timbre Taste in Japanese Music Common with Western Ones. *J. Acoust. Soc. Jpn.* **1993**, *14*, 377–382. [CrossRef]
2. Kaminskyj, I.; Materka, A. Automatic Source Identification of Monophonic Musical Instrument Sounds. In *Proceedings of the IEEE International Conference on Neural Networks*; IEEE: New York, NY, USA, 1995; pp. 189–194. [CrossRef]
3. Brown, J.C. Computer Identification of Musical Instruments Using Pattern Recognition with Cepstral Coefficients as Features. *J. Acoust. Soc. Am.* **1999**, *105*, 1933–1941. [CrossRef] [PubMed]
4. Eronen, A.; Klapuri, A. Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In *Proceedings of the ICASSP*; IEEE: New York, NY, USA; pp. II-753–II-756. [CrossRef]
5. Agostini, G.; Longari, M.; Pollastri, E. Musical Instrument Timbres Classification with Spectral Features. In *Proceedings of the 4th IEEE Workshop on Multimedia Signal Processing*; IEEE: New York, NY, USA, 2001. [CrossRef]
6. Kaminskyj, I.; Czaszejko, T. Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using kNNC. *J. Intell. Inf. Syst.* **2005**, *24*, 199–221. [CrossRef]
7. Park, T.; Lee, T. Musical Instrument Sound Classification with Deep Convolutional Neural Network Using Feature Fusion Approach. *arXiv* **2015**, arXiv:1512.07370.

8. Li, P.; Qian, J.; Wang, T. Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks. *arXiv* **2015**, arXiv:1511.05520.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
10. Gururani, S.; Summers, C.; Lerch, A. Instrument Activity Detection in Polyphonic Music Using Deep Neural Networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*; International Society for Music Information Retrieval: Delft, The Netherlands, 2019; pp. 557–564.
11. Gururani, S.; Sharma, M.; Lerch, A. An Attention Mechanism for Musical Instrument Recognition. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Delft, The Netherlands, 4–8 November 2019; pp. 83–90.
12. Wang, F.; Yu, F. Musical Instrument Identification Based on Improved Convolutional Neural Network and Auditory Spectrum. *Comput. Eng.* **2019**, *45*, 199–205. [CrossRef]
13. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]
14. Xu, K. Recognition and Classification Model of Music Genres and Chinese Traditional Musical Instruments Based on Deep Neural Networks. *Sci. Program.* **2021**, *2021*, 2348494. [CrossRef]
15. Li, R.; Zhang, Q. Audio Recognition of Chinese Traditional Instruments Based on Machine Learning. *Cogn. Comput. Syst.* **2022**, *4*, 108–115. [CrossRef]
16. Reghunath, L.C.; Rajan, R. Transformer-Based Ensemble Method for Multiple Predominant Instruments Recognition in Polyphonic Music. *EURASIP J. Audio Speech Music Process.* **2022**, *2022*, 11. [CrossRef]
17. Chen, R.; Ghobakhlou, A.; Narayanan, A. Hierarchical Residual Attention Network for Musical Instrument Recognition Using Scaled Multi-Spectrogram. *Appl. Sci.* **2024**, *14*, 10837. [CrossRef]
18. Zheng, J.; Cao, M.; Zhang, C. ICKAN: A Deep Musical Instrument Classification Model Incorporating Kolmogorov–Arnold Network. *Sci. Rep.* **2025**, *15*, 21573. [CrossRef] [PubMed]
19. Zhao, E.; Zheng, J.; Cao, M.; Zhang, C. MMIC: A Forward-Bidirectional Mamba Architecture for Musical Instrument Classification. *Signal Image Video Process.* **2026**, *20*, 364. [CrossRef]
20. Zheng, J.; Cao, M.; Zhang, C. A Two-Stage Deep Learning Framework for Lead Instrument Recognition in Polyphonic Music Featuring Chinese Instruments. *PLoS ONE* **2026**, *21*, e0327442. [CrossRef] [PubMed]
21. Mohseni, P.; BabaAli, B.; Asadi, H. HamNava: A Dataset for Multi-Label Instrument Classification. *Trans. Int. Soc. Music Inf. Retr.* **2025**, *8*, 236–247. [CrossRef]
22. Bosch, J.J.; Janer, J.; Fuhrmann, F.; Herrera, P. IRMAS: A Dataset for Instrument Recognition in Musical Audio Signals. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*; International Society for Music Information Retrieval: Porto, Portugal, 2012; pp. 559–564.
23. Lostanlen, V.; Cella, C.-E.; Bittner, R.; Essid, S. Medley-solos-DB: A Cross-Collection Dataset for Musical Instrument Recognition. In Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR), Delft, The Netherlands, 4–8 November 2019.
24. Liang, X.; Li, Z.; Liu, J.; Li, W.; Zhu, J.; Han, B. Constructing a Multimedia Chinese Musical Instrument Database. In *Proceedings of the 6th Conference on Sound and Music Technology (CSMT)*; Springer: Singapore, 2019; pp. 53–60. [CrossRef]
25. Park, H.; Han, D.-B.; Oh, C.; Jung, D. Fantastic AI Sinawi: Composing Korean Traditional Music Using Deep Neural Networks. *J. Digit. Contents Soc.* **2023**, *24*, 785–796. [CrossRef]
26. Li, Q.; Hu, B. Joint Time and Frequency Transformer for Chinese Opera Classification. In *Proceedings of INTERSPEECH 2023*; International Speech Communication Association: Baixas, France, 2023; pp. 3919–3923. [CrossRef]
27. Mohammadi, M.; Maarefvand, M.; Pourbakht, A. Contribution of Spectral and Temporal Cues in Timbre Recognition among Normal-Hearing and Hearing-Impaired People. *Audit. Vestib. Res.* **2024**, *33*, 322–329. [CrossRef]
28. Bai, N.; Wu, Z.; Zhang, J. Multi-Instance Multi-Scale Graph Attention Neural Net with Label Semantic Embeddings for Instrument Recognition. *Signals* **2025**, *6*, 30. [CrossRef]
29. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the EMNLP*; Association for Computational Linguistics: Kerrville, TX, USA, 2014; pp. 1724–1734.
30. Schuster, M.; Paliwal, K.K. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef] [PubMed]
31. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
32. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4. [CrossRef]

33. kryptologyst. *Musical-Instrument-Recognition*; Computer software; GitHub: San Francisco, CA, USA, 2025. Available online: https://github.com/kryptologyst/Musical-Instrument-Recognition (accessed on 1 October 2026).
34. Baron, D.; Gilani, A.; Lee, H. InstruNET: Convolutional-Based Model for Instrument Identification in Raw Audio Signals. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024.
35. Loughran, R.; Walker, J.; O'Neill, M.; O'Farrell, M. Musical Instrument Identification Using Principal Component Analysis and Multi-Layered Perceptrons. *Proc. ICALIP* **2008**, 643–648. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.