Sensors & Transducers, Vol. 263, Issue 4, December 2023, pp. 119-130

Sensors & Transducers

Published by IFSA Publishing, S. L., 2023
http://www.sensorsportal.com

Classifying Musical Instrument Using Spatiotemporal Features with Deep Neural Networks

Chai Xiu CHIAH, * Lee Choo TAY and Weng Kin LAI
Tunku Abdul Rahman University of Management and Technology, Jalan Genting Kelang, Setapak, 53300 Kuala Lumpur, Malaysia
Tel.: 6034145123, fax: 60341423166
E-mail: taylc@tarc.edu.my

Received: 2 October 2023 / Accepted: 8 December 2023 / Published: 21 December 2023

Abstract: Similar to trained human ears, a machine could be trained to recognize musical instruments presence in audio tracks for the purpose of musical information retrieval. The audio signal produced by an instrument has unique temporal and spectral features, which could be extracted for machine learning purpose. This research investigates the use of spatiotemporal features, which were converted from temporal and spectral features of monophonic music for this application. The performance of a recurrent neural network model called bidirectional Long Short Term Memory (Bi-LSTM) network and two convolutional neural networks (CNNs) in musical instruments classification with log Mel-spectrogram were analyzed and evaluated. With a dataset of 14 musical instruments, the Bi-LSTM and 1-dimensional CNN models obtained a macro F1 score of 0.976 and 0.977 respectively, while 2-dimensional CNN model achieved 0.985.

Keywords: Long Short-Term Memory Network (LSTM), Convolutional Neural Network (CNN), Musical instruments classification, Mel-Spectrogram, Spectral features, Spatiotemporal features.

1. Introduction

This paper is an extended version of the ASPAI ’2023 conference paper published in June 2023 [1]. Musical instrument recognition is an active research area in the music information retrieval (MIR) field and it has contributed significantly in providing essential and useful information for various audio or music applications. Some examples are playlist generation, music recommendation system, music browsing, music transcription, and genre classification. Compared to text search with its input queries and data are both in text format, music search is a more complex task as the music files are in digital format. Hence, it is imperative to convert music information into text format, i.e., music tags, to facilitate music search. Musical instrument recognition plays an important role to provide instrumental information in music search with the increase of digital music files. Automatic music transcription and playlist generation in music streaming platforms rely heavily on MIR tasks to provide better users experience.

Audio information extraction is a challenging task, especially from polyphonic music that includes human voice. The task of musical instruments identification in polyphonic music with multiple musical instruments requires a complicated system, as the sound generated by different instruments may varies in timbre, instrument quality and playing style, not to mention the overlapping of sound from multiple instruments. Various forms of convolutional neural network (CNN) approach using spectral and cepstral features have been widely used and have shown a better performance over the traditional machine

learning approach. However, MIR is intrinsically a temporal task and some past literatures have indicated that temporal features provide important information as well in musical instrument classification [2]. It is necessary to develop an effective model on processing temporal information and learning from time series data, i.e., music. In this paper, a temporal model is developed and its performance compared to two commonly employed neural networks using cepstral features.

2. Prior Work

Recent developments in musical instrument classification involve the use of machine learning and deep learning using neural networks. As music is a form of audio signals that varies with time, conversion from its temporal to matrix-like format in order to match the inputs requirements to any neural network is necessary. This conversion process involves the selection of important features in the audio signals. Generally, the audio features are divided into physical features and perceptual features [3]. The former includes measurable temporal and spectral features, e.g., the zero crossing rate, the energy function, the spectrum, while the latter refers to subjective terms such as relative loudness, brightness, pitch and timbre.

Based on the literature reviewed, the performance of the classification systems depends on the features selected, the classifier used, and the number of and the family of musical instruments trained. This section discusses the first two aspects, while keeping the third aspect in consideration, followed by some review on instrument identification in monophonic and polyphonic music pieces.

2.1. Feature Selection

In order to recognize a particular instrument, the unique characteristics of the musical notes played by the instrument ought to be properly selected.

Among the many features extraction methods, wavelet transform coefficients, Mel-spectrogram, Linear Frequency Cepstral Coefficients (LFCCs), Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP), are commonly used in musical instrument recognition.

Both continuous and discrete wavelet transforms convert the audio signal into time-frequency domain for the extraction of features such as instantaneous frequency, pitch and sub-band energies from wavelet coefficients. One of the significant work using wavelet is [4]. Evaluation on 13 instruments using 3 significant features of log Scale Distribution Width (SDW), Time Variance of Dominant Scale (TVDS) and Wavelet Mean of Inverse Scale (WMIS) achieved an accuracy of 85.5 %. A simple classifier called linear discrimination analysis (LDA) classifiers with leave-one-out cross-validation method was employed in their work. Subsequent work done by Kothe et al [5] in comparing the performance of wavelet features with temporal features, spectral features, cepstral features and a combination of all features showed that wavelet features performed better than others using SVM classifier. Their work also demonstrated that best performance could be obtained by using a smaller number of features with proper weight.

Various forms of spectrogram have been adopted by researchers in training intelligent systems. Huaysrijan and Pongpinigpinyo [6] compared the performance of Mel-spectrogram and MFCCs in recognizing 13 Thai classical instruments with CNN. MFCCs marginally outperformed Mel-spectrogram by 1 % in both accuracy and F1 score at 0.99. They also compared the performance of 3 deep neural network, namely CNN, LSTM (Long Short-Term Memory) and CRNN (Convolutional Recurrent Neural Network), in which CNN ranked top at 99.44 %, LSTM at 81.42 % while CRNN at 93.35 % accuracy. In another work, log Mel-spectrogram was used as acoustic features by Yun and Bi [2] in comparing the performance of LeNet (a form of CNN) and LSTM in recognizing 14 instruments. Again, LeNet was slightly better than LSTM with an accuracy of 0.83 versus 0.82. Szeliga et al [7] also used Mel-spectrograms in their work on staged training a CNN in musical instrument recognition. Both monophonic and polyphonic data of 7 instruments were used. The accuracy is lower at about 75 %. This is expected as the identifying the predominant instrument in polyphonic music is a big challenge.

Many have used MFCCs, either alone or together with other features in musical instrument classification [6, 8-15]. Mahanta et al. [8] used the MFCCs obtained from the audio data of 20 instruments belonging to 4 families viz. woodwinds, brass, percussion, and strings to train an Artificial Neural Network (ANN) with Rectified Linear Unit (ReLU) activation function and ADAM optimizer. A high accuracy and F1 score of 0.97 was reported. In their work on comparing 4 feature extraction methods, Rodin et al [9] discovered that MFCCs did extremely well in classifying 15 classes of instruments using two types of CNN. ResNet-34 models that were trained using Mel-Spectrograms, MFCCs and LFCCs achieved a perfect accuracy of 100 %, while VGG-16 models obtained 100 % accuracy using MFCCs while Mel-Spectrograms 99.72 % and LFCCs 99.44 %. Masood and Shadab Khan [10] used MFCCs with 7 timbre features to train an ANN on classifying 4 instruments, i.e., violin, piano, flute, drum and guitar, yielded an average accuracy of 89.17 %. A long training time of 450 epochs was reported.

2.2. Classifier

The types of classifiers used in musical instrument recognition basically fall into two categories: machine learning classifiers, e.g., K-Nearest Neighbour (KNN) [5, 12], Support Vector Machine (SVM) [5, 16-18],

Hidden Markov Model (HMM) [11, 18]; and neural network classifiers, e.g., ANN [8, 13, 19, 20], CNN [2, 6, 7, 9, 15, 16, 21-23] and LSTM [2, 6].

Bhalke et al [12] used 39 MFCC features to recognize 6 Indian musical instruments namely piano, guitar, violin, synthesizer, harmonica and accordion using KNN obtained an average accuracy of 91.83 %. Joder, *et al* [18] employed a combination of 13 MFCC features and other specially selected temporal and spectral features to classify 8 instruments: piano, guitar, oboe, clarinet, French horn, trumpet, cello and violin achieved an average accuracy of 84.5 % using SVM. In their work on evaluating the performance of 3 feature extraction techniques and 3 classifiers on 4 dissimilar instruments, Jeyalakshimi, *et al* [11] reported that MFCCs and HMM gave the best accuracy
of 90.5 %.

Between machine learning classifiers and neural network classifiers, neural network classifiers were more widely used and performed better. Various forms of CNN were able to achieve more than 90 % accuracy with MFCCs features in monophonic musical instrument [2, 6, 7, 9, 15] and mixed results in predominant instrument identification in polyphonic music [22-25] using various dataset.

2.3. Instrument Recognition in Monophonic Music

For musical instruments recognition in monophonic music, Toghiani-Rizi and Windmark, [19], collected audio samples from the London Philharmonic Orchestra dataset and converted them into frequency domain signals. The authors trained an artificial neural network (ANN) model and carried out experiments analyzing the effect of different audio signal features on recognition accuracy – the whole audio sample, only the attack of audio signal, all characteristics of signal except attack, the first 100 Hz of the signal in frequency domain and the following 900 Hz of the same spectrum. The results showed that the ANN model with complete music sample reached the highest average accuracy of 93.5 % among the experiments.

Chakraborty and Parekh, [13], extracted 6 audio descriptors including Mel Frequency Cepstral Coefficients (MFCC), linear predictive coding (LPC), cepstral coefficients (CC) and spectral centroids as spectral descriptors, pitch salience and hierarchical clustering on principal components (HPCP) as tonal descriptors. All the features respectively acted as inputs to 4 different classifiers, namely, k-nearest neighbour (k-NN), support vector machine algorithm (SVM), ANN and random forest. As a result, the ANN classifier showed the best performance with better tuning. Meanwhile, CC was considered the best approach among other features, even the regular approaches with MFCC.

In [20], the authors introduced the implementation of an ANN-based model with specific transient length selection which contained most significant music information. Two experiments were carried out on string instruments recognition where the first experiment had applied the first 1000 Hz normalized frequency spectrum with selected onset position as input data and achieved an accuracy of 96.47 %. As for the second experiment, the authors had extended the signal attack length which included signal peak amplitude that carried more fundamental information, resulting in the final accuracy of 98.22 %.

Kakarwal and Chaudhary [26] proposed a k-NN model on classifying different Indian string instruments by analyzing features of audio samples including zero crossing rate (ZCR), roughness, spectral centroid, brightness and roll-off. The research claimed that the average accuracy of the classifier with the combination of more than 3 audio features was much higher than that with only a single feature extraction, which are 90 % and 67 % respectively. It showed that the advantages of audio feature extraction involving multiple temporal and spectral features is to provide sufficient music information retrieval.

With the motivation from previous LSTM-based tasks, Yun and Bi, [2], implemented a single layer (LSTM) network on the task of musical instruments through temporal features extraction of audio signals. Like CNN approach, the authors trained the LSTM model with log Mel-spectrogram but with temporal features. The LSTM model was then compared with a LeNet CNN model on the performance of instrument recognition and both models obtained the accuracies of 81.85 % and 83.04 % respectively, which were pretty close.

2.4. Instrument Recognition in Monophonic Music

In the topic of polyphonic musical instruments identification and analysis, Han et al. [24] had done notable works exploiting a convolution neural network (CNN) model with Mel-spectrogram as structure input on instrument recognition. The authors had also proposed a t-distributed stochastic neighbor embedding (t-SNE) visualization to show clustering results of each convolutional block. By demonstrating comparison with SVM, the conventional algorithm, Han et al.'s proposed model with CovNet structure had shown better performance with higher micro and macro F1-measure – 0.619 and 0.513 respectively. However, due to imbalance in sample sizes for each instrument type, there was low identification on cello and clarinet in instrument recognition on IRMAS dataset.

Using Han et al.'s research as baseline, Yu et al., [23], introduced a CovNet-ABC model where implementing a CNN model embedded with auxiliary classification based on instrument grouping, batch normalization and center loss. With the help of

auxiliary classification, the problem of poor identification on cello and clarinet presented in [8] had been minimized. As a drawback, some instruments belonging to the same instrument family such as saxophone and trumpet had increased the classification difficulty due to overlapping and great similarities. Anyhow, the model performance with 0.685 micro F1-measure and 0.597 macro F1-measure had shown a slight improvement over [24].

By using the aforementioned IRMAS dataset, Solanki and Pandey, [25], proposed a CNN model with ALEXNET structure containing 8 convolutional layers with learnable parameters to fit large datasets. The proposed model had implemented a 2-dimensional (2D) convolutional neural network with varying filters on instrument recognition. Similar to [24], input audio signals were converted into the form of Mel-spectrogram whilst rectified linear unit (ReLU) activation, max-pooling and dropout regulation were applied to the proposed model. As the results presentation, SoftMax function was utilized to identify the posterior accuracy of musical instruments classification. With the dropout of 0.2, the proposed model achieved an excellent result with 92.80 % accuracy, yet during training and testing period, overfitting occurred during 55th and 56th epoch from a total of 60 epochs. Additionally, a comparison was made between the ReLU and leaky ReLU activation function, as a result, the ReLU activation performed better with accuracy of 92.61 %.

Besides, with the same dataset, IRMAS, Racharla, *et al*., [17], proposed an approach to classify predominant musical instruments from polyphonic music based on spectral features. In this paper, traditional classification algorithms were implemented including decision tree, Logistic Regression, Light BGM, XG Boost and Random Forest. Additionally, the machine learning technique, SVM, was applied as well to extract MFCC and spectral features inclusive of ZCR, spectral centroid, bandwidth and roll-off. As an outcome of the research and comparison, recognition model with SVM algorithm yielded the highest accuracy of 76 % over the traditional classification approaches. Nonetheless, Racharla et al. had shown the effectiveness of multi-spectral features extraction on the improvement in classification accuracy.

Distinctively, Gururani et al. [22], proposed an approach to detect music instruments activity in polyphonic music on a close-grained temporal scale instead of focusing on predominant instruments prediction in a given audio sample as per other works. For auto generation of annotations for instrument activity, the authors picked multitrack audio from MedleyDB and Mixing Secrets over the IRMAS dataset used in past related works. The reason behind was that manual annotation in IRMAS dataset might contain more errors and IRMAS consisted of only a single instrument in each track which might be unideal for instrument activity detection. In their work, 3 different deep neural networks (DNNs) models: multilayer perceptron (MLP), convolution neural networks (CNN), and convolution recurrent neural networks (CRNN) were proposed and compared. The performance of models was summarized using area under curve (AUC) presentation and it was presented that both the CRNN and CNN outperformed MLP algorithm in 18 instruments detection.

In the subsequent year, instead of strongly labelled data, Gururani et al., [27], chose weakly labelled data in the OpenMIC dataset containing 20 distinct musical instruments for instrument recognition. The authors had compared an attention-based model against the other 3 baseline classification models including binary relevance random forest (RF_BR), recurrent neural network (RNN), and fully connected neural networks (FCNN). Undoubtedly, the model incorporating attention mechanism had shown an overall improvement in classification performance where the model attended to specific time intervals in the track associated with each instrument, resulting in interpretable results.

Motivated by these past researchers, this paper explored the performance of a LSTM network and two convolution networks in recognizing musical instruments in monophonic music taken from a standard dataset. With the intent of improving its performance, the spatiotemporal features, in the form of log Mel-Spectrograms of the audio data, were used for training and testing.

3. System Architecture

The research framework is illustrated in Fig. 1. Pre-labeled monophonic audio samples were first preprocessed before the audio tracks were converted into their corresponding Mel-Spectrograms. The Bi-LSTM and CNN models were developed and trained using these Mel-Spectrograms. Training and validation losses were then used to tune the model parameters until an optimum performance was obtained for each model. The performance of these trained model was then evaluated on a new set of testing data and the results compared.

The dataset used in this project was extracted from monophonic audio samples in Kaggle dataset developed by Dibakar Sil [28]. The audio samples of 14 instruments from the family of string, woodwind and brass were categorized into different subfolders annotated with the instrument name. Due to the quantity of audio tracks available in the dataset is different for each instrument, uneven number of samples were used in this study. The distribution of these instruments are shown in Fig. 2. Violin occupies the highest number of samples of 278 while viola the lowest at 108 in the training and validation dataset. Similarly, the quantity of samples in the testing dataset follow the proportion, with the former high at 100 while the latter lower at 50.

The models built for evaluation are Bi-LSTM, one-dimensional CNN (1D CNN) and two-dimensional CNN (2D CNN).

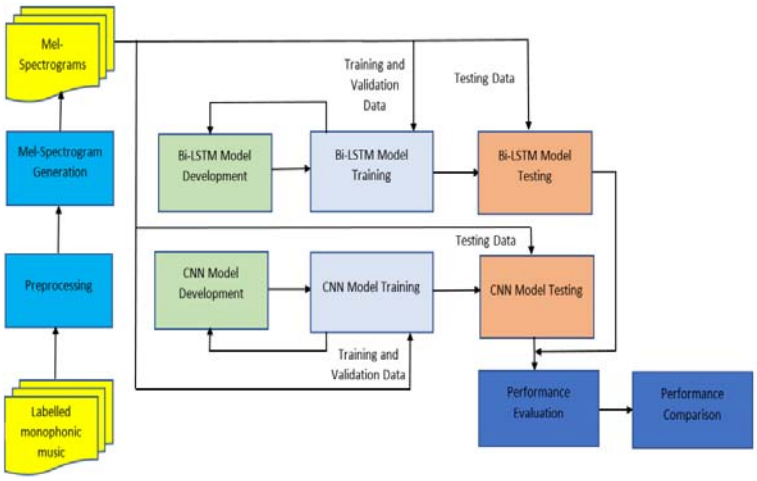


Fig. 1. Proposed research framework.

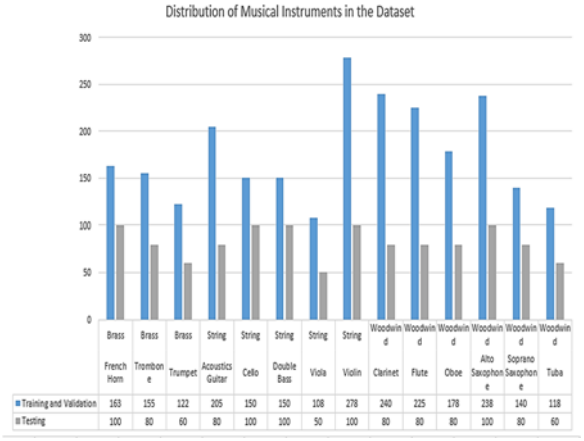


Fig. 2. Distribution of musical instrument audio samples.

3.1. Audio Signal Preprocessing

The process of extracting spatiotemporal features in the form of a log Mel-spectrograms from the audio data is illustrated in Fig. 3. The audio samples in 16-bit WAV format were first enveloped to remove silence in order to avoid the problem of every audio clip having similar patterns. Then each clip was trimmed into pieces of 1 second to form the actual input data. Consequently, the total number of audio clips for training and validation was 2470 clips of audio signal of equal length, in which 80 % was used for training and 20 % for validation. A fresh set of data was prepared in a similar manner for testing.

The trimmed audio clips were down-sampled with a rate of 22050 Hz which was the Nyquist rate, and the stereo audio was converted into mono input by taking the average of the left and right audio channels. The sampling rate of 22050 Hz allows the use of up to 11025 Hz as Nyquist frequency which can include most of the instrument information. Noises above the frequency range were removed.

In this research, a Mel-spectrogram was used as the time-frequency represented input data to the system and hence the implementation of short-time Fourier transform (STFT) with 512 sample length was introduced. Each of the 1-s audio signals was further divided into short term sequences of fixed window size of 25 ms and a hop size of 10 ms, and then the fast Fourier transform (FFT) was applied on the signals. This approach is similar to that proposed by S. Rajesh and N. J. Nalini [30].

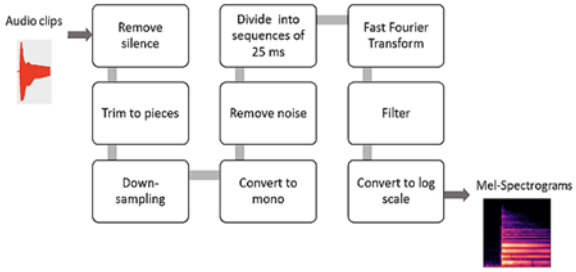
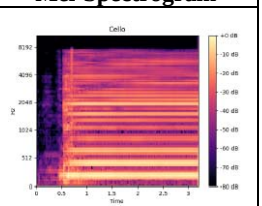
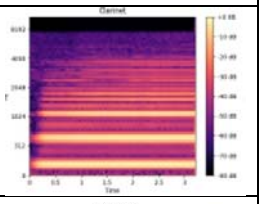
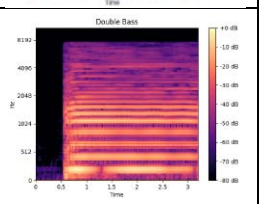
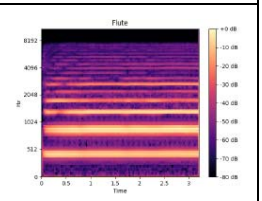
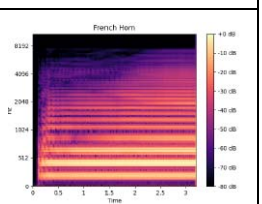
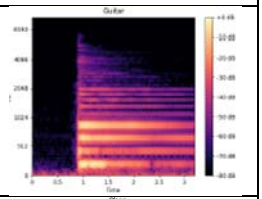
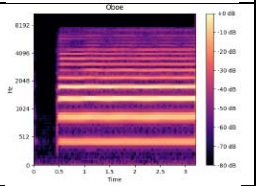
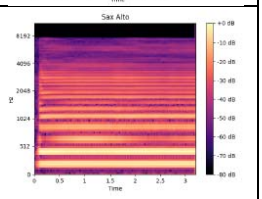
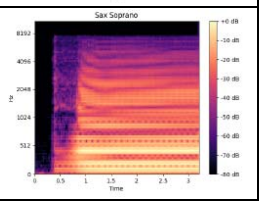
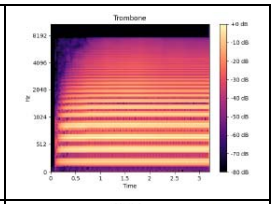
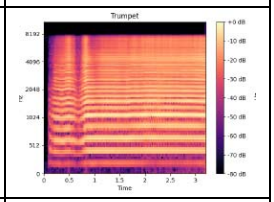
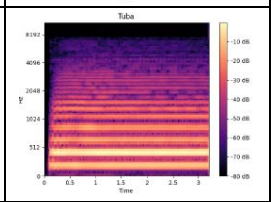
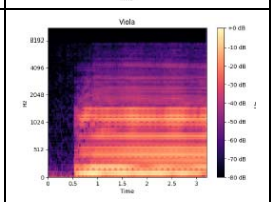
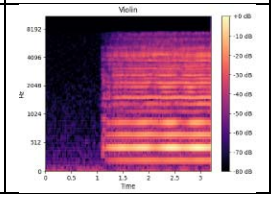


Fig. 3. Audio signal pre-processing and Mel-Spectrogram generation.

In the spectrogram of the audio signal, where the frequency axis is linear, the higher harmonics were not equally represented as the lower harmonics, in which the same differences in low frequencies are perceived as bigger differences in high frequencies. Furthermore, due to the nonlinearity of the human audio perception, where the pitch and loudness are logarithmic in nature, the frequency and the intensity that represent the loudness of the signal were converted into logarithmic scale. Hence, the corresponding Mel-spectrograms were generated by using a Mel filter-bank with 128 bands to reduce the dimensionality of the spectrogram generated, as suggested by [24]. The Mel-spectrogram was then converted into logarithmic scale for better audio visualization. The waveforms and the log Mel-spectrograms of the instruments used are illustrated in Table 1.

Table 1. Waveforms and Mel-Spectrograms of the instruments.

Instrument	Waveform	Mel-Spectrogram
Cello		
Clarinet		
Double Bass		
Flute		
French Horn		
Guitar		
Oboe		
Sax Alto		
Sax Soprano		
Trombone		
Trumpet		
Tuba		
Viola		
Violin		

There are 3 axes in the Mel-spectrogram. The x-axis is the time; the y-axis is the frequency while the z-axis is the intensity of the audio signal in colour. It can be perceived as the representation of the audio signal in image form. A single point in the image conveys information on single frequency component of the audio signal and its corresponding strength at a specific time. Sliding across the horizontal axis, the temporal features, i.e., the fluctuation in the signal strength for all frequency components as time changes, can be observed; while sliding along the vertical axis reveals the spectral features that represent the variation in the composition of the signal at certain time. With the signal strength and frequency being depicted in logarithmic scale, forms the cepstral features, which can be viewed as time-varying patterns on the Mel-spectrogram, hence the term spatiotemporal features. The conversion of audio signal into image form enables the audio signal to be analyzed as an image and eases the utilization of any AI techniques that work on image. The resultant image size is 128×130 pixels.

3.2. Network Architecture

The models of the three neural networks were implemented using the Kapre (Keras Audio Preprocessing) library developed by Choi et al [29].

3.2.1. Bidirectional Long Short-term Memory (Bi-LSTM) Architecture

In this research, the Bi-LSTM model was implemented with two dense layers in a bidirectional manner, as illustrate in Fig. 4. The audio data was first normalized before going through the actual Bi-LSTM model, a time distributed dense layer was created to filter out the signal feature over time from the input data. Next, a bidirectional layer was constructed to compute the gradient in both forward and backward manner. The output from the bidirectional layer was then concatenated with the features filtered from the previous time-distributed dense layer.

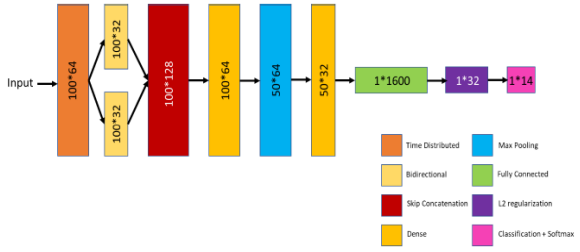


Fig. 4. Architecture of Bi-LSTM.

By doing so, the Bi-LSTM network could not only learn the signal features from the input data but also took reference in the concatenated features from these two layers. Subsequently, two dense layers with ReLU activation were created with a max-pooling layer in between to summarize the features from the first layer by calculating its maximum value. Since The Bi-LSTM was prone to overfitting, according to [30], a dropout of 0.5 was added before the classification layer to minimize the problem. On the classification layer, a SoftMax function was used to indicate the probability for each instrument.

3.2.2. 2D CNN Architecture

A simple 2-dimensional convolutional neural network (2D CNN) model similar to that of computer vision-based approaches like VGG-Net structure was constructed. The Mel-spectrogram formed the input data to the 5 layers CNN model. First layer was the hyperbolic tangent (tanh) convolution layer with a kernel size of 3×3 to learn the small features of data. The kernel size was maintained in the following convolution layers and the number of filters in each layer was increased by a factor of 2 for every two convolution layers to obtain the more specific information of the data as illustrated in Fig. 5. The structure is leaner than VGG16.

Zero padding was applied on the input data for each convolution layer to prevent the loss of some useful information with the increase of convolution layers. Also, a max-pooling method was used after each convolution layer to compute the maximum value of the feature for each patch. However, after the last convolution layer, a global max-pooling layer was constructed to obtain the maximum value of all features before the fully flatten connected layer. After that, a dropout layer was added with the dropout rate of 0.3 as a slight improvement over [25] and eventually a SoftMax function was used for the classification layer. The reason not to use sigmoid function like in [24] is because there was only one instrument presented in each file in this research while sigmoid function was frequently used to handle multiple instruments classification.

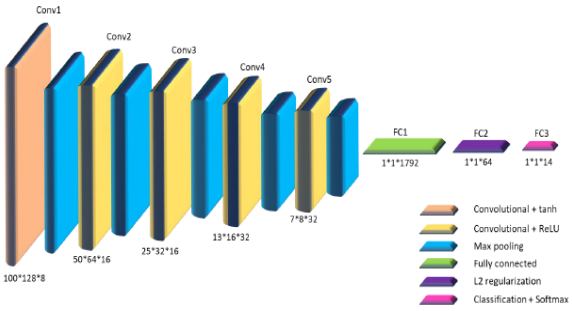


Fig. 5. Architecture of 2D CNN.

3.2.3. 1D CNN Architecture

A 1-dimensional convolutional neural network (1D CNN) was constructed with reference from [21]. The model had 5 time-distributed convolution layers which was the same number of layers as the 2D CNN model. With the time-distributed layers, the model would learn only the temporal features from the Mel-spectrogram. According to the authors, the small convolutional filters with the size of 3 gave the best accuracy. Hence, the kernel size for each convolutional layer was fixed at 3. The number of filters in each layer was increased by a factor of 2 for every layer as presented in Fig. 6. A 1D max-pooling layer was added after each convolution layer for maximum value computation from the previous layer. Similar to the 2D CNN model, a dropout layer was added before the classification layer and SoftMax function was integrated.

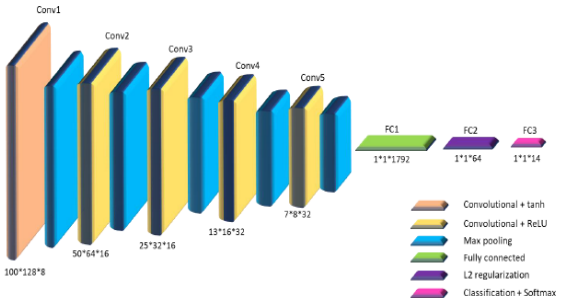


Fig. 6. Architecture of 1D CNN.

3.3. Performance Evaluation

The performance of the models was evaluated through predictions on new audio files taken from the same dataset. To determine accuracy on instrument recognition, the predicted results were visualized in a confusion matrix. The values of True Positive (*TP*), False Positive (*FP*), True Negative (*TN*) and False Negative (*FN*) were used to compute the overall accuracy, precision (*P*) and recall (*R*) of the model using (1) to (3). Following the evaluation method widely used in the past literature [24], the *F1* score in (4) was used to obtain the overall performance of the model, which was the harmonic mean of precision.

$$Accuracy = \frac{TP+TN}{Total\ Predictions}, \quad (1)$$

$$P = \frac{TP}{TP + FP}, \quad (2)$$

$$R = \frac{TP}{TP + FN}, \quad (3)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4)$$

In addition, as the dataset used was multi-class labels with non-uniform distribution, macro-average computation and micro-average computation were carried out. The macro *P* and *R* scores were obtained individually for each class before the calculation of macro P_{macro} and R_{macro}, as illustrated in (5) and (6), where *k* is the total number of classes. $F1_{macro}$ was then computed using P_{macro} and R_{macro}, as shown in (7).

$$P_{macro} = \frac{\sum_{i=1}^{k} P_i}{k}, \quad (5)$$

$$R_{macro} = \frac{\sum_{i=1}^{k} R_i}{k}, \quad (6)$$

$$F1_{macro} = \frac{2 \times P_{macro} \times R_{macro}}{P_{macro} + R_{macro}} \quad (7)$$

On the other hand, the precision and recall score for micro-average were calculated individually from *TP*, *TN*, *FP* and *FN* of each instrument class. The micro precision (P_{micro}) was obtained from the sum of *TP*s for all the classes divided by the sum of all *TP*s and *FP*s. As for the micro recall score (R_{micro}), it would be the division of the total *TP*s for all the classes by the sum of *TP*s and *FN*s. Micro *F1* score was then computed using P_{micro} and R_{micro}. The equations to compute them are given in equations (8) to (10).

$$P_{micro} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FP_i}, \quad (8)$$

$$R_{micro} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FN_i}, \quad (9)$$

$$F1_{macro} = \frac{2 \times P_{macro} \times R_{macro}}{P_{macro} + R_{macro}} \quad (10)$$

With the overall accuracy and F1 measure obtained, a comparison was made among the Bi-LSTM model and the CNN models in musical instrument classification based on monophonic audio signals.

4. Results and Discussions

All three trained models were able to achieve validation accuracy of 1 after 30 epochs of training. It is worth to note that 2D CNN achieved the best accuracy as early as the 8th epoch, whereas the Bi-LSTM model only managed its best accuracy on the 12th epoch. However, it was unstable and fluctuated before stabilizing on the 24th epoch. The 1D CNN model took the longest time to achieve a stable best accuracy on the 29th epoch. These can be seen
in Fig. 7.

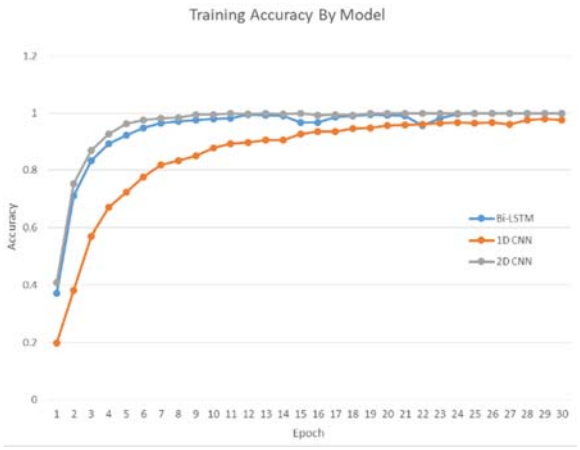


Fig. 7. Training accuracy.

4.1. Predictive Performance of Models

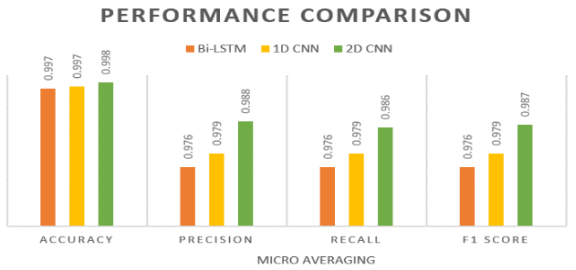
A new set of data was prepared for the musical instruments prediction to examine the practical performance of the three models. The test results are illustrated in Fig. 8. All the three models produced very similar accuracies. In the macro computation of accuracy, precision, recall and F1 score, the predictive accuracy of B-LSTM is the same as the 1D CNN model of 0.997 while the overall accuracy of the 2D CNN model is 0.998.

The macro precision of Bi-LSTM and 1D CNN are similar at 0.977 while 2D CNN is slightly better at 0.987. However, the macro recall of Bi-LSTM is the

worst at 0.975, 1D CNN is marginally better at 0.978 while 2D CNN is the winner at 0.984. The macro F1 score in ascending order is therefore Bi-LSTM at 0.976, 1D CNN at 0.977 and 2D CNN at 0.985.

The micro precision, recall and F1 score of the 3 models follows the similar trend in macro with 2D CNN performing the best while Bi-LSTM the worst with a small difference 0.01 or 1 %.

(a) Macro



(b) Micro

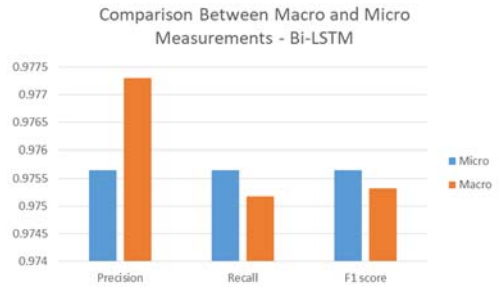
Fig. 8. Overall performance.

A deeper analysis into the performance of each model revealed the difference between micro and macro precision is larger than the differences between the micro and macro values in recall and F1 score in Bi-LSTM, as shown in Fig. 9(a). This suggests there could be more false positives in at least one of the instruments in this model. Interestingly, looking at Fig. 9(b) and Fig. 9(c), 1D CNN and 2D CNN exhibit the opposite trend, i.e., the micro results are better than the macro, implying that these two models perform more consistently across all the instruments as compared to Bi-LSTM. This led us into analyzing the instrument-wise performance of these 3 classifiers.

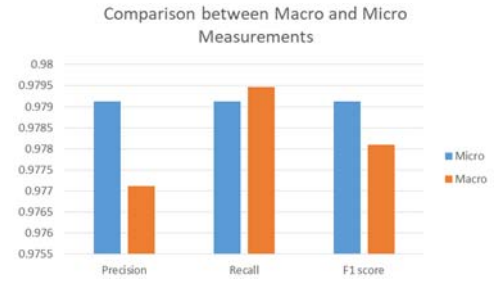
4.2. Performance on Instrument-wise Identification

Fig. 10 shows the predictive performance of Bi-LSTM in all 14 instruments in the form of a confusion matrix while Table 2 summarizes the accuracy, precision, recall and F1 score of all the instruments.

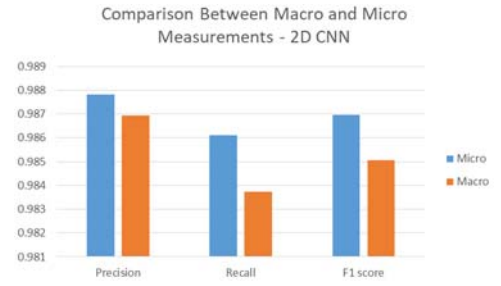
Bi-LSTM performed the best on cello and trombone with an accuracy and an F1 score of 1 and the worst in sax soprano with the lowest F1 score of 0.91. The low F1 score is mainly attributed by the low precision 0f 0.833 caused by high false positive (FP) where the model wrongly predicted 12 cases of violin and 1 case of flute, French horn, trumpet and tuba each as sax soprano. The model is found to perform poorly on violin in recall due to high false negative with 12 cases wrongly predicted as sax soprano mentioned above. The model was found to confuse between sax soprano and violin, probably due to some similarity between the Mel-Spectrograms of these two instruments in some musical notes.



(a) Bi-LSTM



(b) 1D CNN



(c) 2D CNN

Fig. 9. Results of macro and micro measurements.

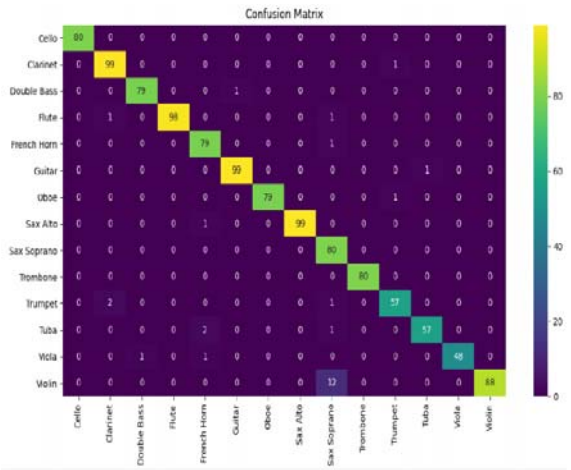


Fig. 10. Confusion matrix of Bi-LSTM.

Table 2. Performance of Bi-LSTM.

	Accuracy	Precision	Recall	F1 score
Cello	**1.000**	1.000	1.000	**1.000**
Clarinet	0.997	0.971	0.990	0.980
Double Bass	0.998	0.988	0.988	0.988
Flute	0.998	1.000	0.980	0.990
French Horn	0.996	0.952	0.988	0.969
Guitar	0.998	0.990	0.990	0.990
Oboe	0.999	1.000	0.988	0.994
Sax Alto	0.999	1.000	0.990	0.995
Sax Soprano	0.986	0.833	1.000	0.909
Trombone	**1.000**	1.000	1.000	**1.000**
Trumpet	0.996	0.966	0.950	0.958
Tuba	0.997	0.983	0.950	0.966
Viola	0.998	1.000	0.960	0.980
Violin	0.990	1.000	0.880	0.936
Overall	0.997	0.977	0.975	0.975

Using the same input as the Bi-LSTM classifier, the performance of 1D CNN is fairly consistent across all the instruments. As shown in Fig. 11 and Table 3, it obtained a perfect accuracy and an F1 score in trumpet, but the lowest F1 score of 0.941 in viola, due to a low precision of 0.923 caused by high false positive with 3 cases of cello and 1 case of violin wrongly predicted as viola. Note that all three are string instruments and it is possible that the features of some musical notes are similar.

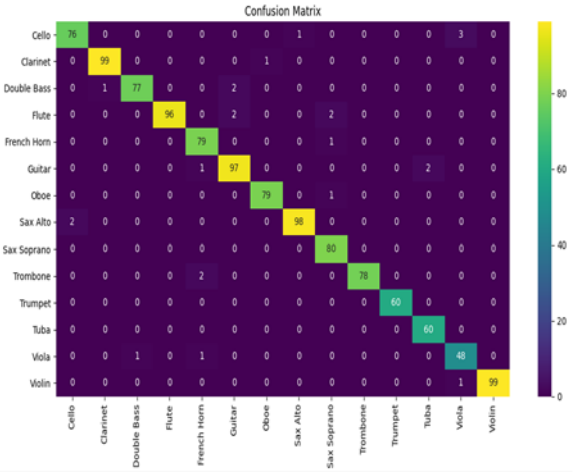


Fig. 11. 1D CNN confusion matrix.

The results of 2D CNN are shown in Fig. 12 and Table 4. In contrary to Bi-LSTM, 2D CNN was observed to perform the best in violin with a perfect accuracy and an F1 score of 1. The same perfect score was also obtained in guitar and trombone. However, the model did poorly in sax soprano with a low precision of 0.909 and F1 score of 0.952, due to cases where cello, clarinet, flute, French horn and tuba were wrongly predicted as sax soprano.

Fig. 13 compares the F1 score of the 3 classifiers across the 14 instruments. While Bi-LSTM out-performed others in cello, flute, oboe and viola, it's extremely poor performance in sax soprano, trumpet and violin resulted in the lowest F1 score among all 3 classifiers. It is worth to note that both Bi-LSTM and 2D CNN classifier did best in trombone and worst in sax soprano, while 1D CNN did best in trumpet and worst in viola.

Table 3. Performance of 1D CNN.

	Accuracy	Precision	Recall	F1 score
Cello	0.995	0.974	0.950	0.962
Clarinet	0.998	0.990	0.990	0.990
Double Bass	0.997	0.987	0.963	0.975
Flute	0.997	1.000	0.960	0.980
French Horn	0.996	0.952	0.988	0.969
Guitar	0.996	0.980	0.970	0.975
Oboe	0.997	0.963	0.988	0.975
Sax Alto	0.997	0.990	0.980	0.985
Sax Soprano	0.997	0.952	1.000	0.976
Trombone	0.998	1.000	0.975	0.987
Trumpet	**1.000**	1.000	1.000	**1.000**
Tuba	0.998	0.968	1.000	0.984
Viola	0.995	0.923	0.960	0.941
Violin	0.999	1.000	0.990	0.995
Overall	0.997	0.977	0.979	0.978

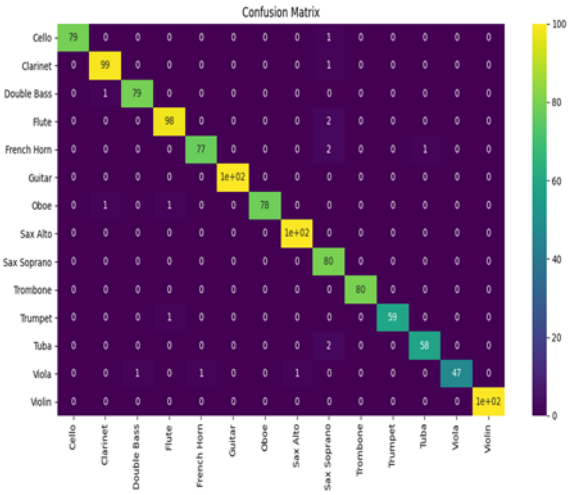


Fig. 12. 2D CNN confusion matrix.

Table 4. Performance of 2D CNN.

	Accuracy	Precision	Recall	F1 score
Cello	0.999	1.000	0.988	0.994
Clarinet	0.997	0.980	0.990	0.985
Double Bass	0.998	0.988	0.988	0.988
Flute	0.998	0.980	0.980	0.980
French Horn	0.997	0.987	0.963	0.975
Guitar	**1.000**	1.000	1.000	**1.000**
Oboe	0.998	1.000	0.975	0.987
Sax Alto	0.999	0.990	1.000	0.995
Sax Soprano	0.993	0.909	1.000	0.952
Trombone	**1.000**	1.000	1.000	**1.000**
Trumpet	0.999	1.000	0.983	0.992
Tuba	0.997	0.983	0.967	0.975
Viola	0.997	1.000	0.940	0.969
Violin	**1.000**	1.000	1.000	**1.000**
Overall	0.998	0.987	0.984	0.985

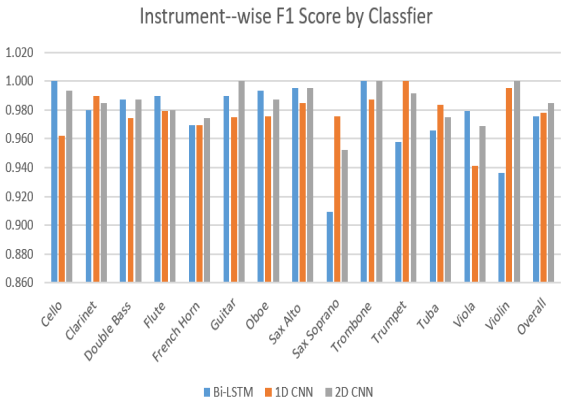


Fig. 13. F1 score of all instruments.

This research has demonstrated that 3 neural networks properly trained with the necessary spatiotemporal features are able to perform well with an accuracy and F1 score higher than 0.996 and 0.975 respectively in monophonic music. All 3 classifiers developed out-performed prior work [2, 7] that used CNN with Mel-spectrograms and [6, 15] that employed CNN with MFCCs.

5. Conclusions

Recognizing the correct musical instruments in audio tracks is important for the purpose of musical information retrieval. This paper investigated the combination of spatiotemporal features coupled with deep learning neural networks to classify musical instruments. Unlike the traditional approaches, the Mel-spectrograms of the musical signals were depicted as images.

However, musical data is complex, with the spectral characteristics overlapping temporal features. Moreover, music is a continuous signal too and it can be of different lengths with its pattern varying according to the sound tuning and filtering. Hence to cater for this particular consideration, the spatiotemporal features that were generated from the spectral and temporal characteristics of the musical notes were used in the musical instruments recognition model.

Three neural networks, namely Bi-LSTM recurrent neural network, 1-dimensional convolution network and 2-dimensional convolution neural network were examined using this approach to recognize 14 musical instruments through the monophonic musical tones. The log Mel-spectrogram of audio samples in Kaggle dataset were used for training and testing. All three models exhibited strong performance, with accuracy exceeding 0.996. Specifically, the 2D CNN achieved an F1 score of 0.985, while the 1D CNN and Bi-LSTM achieved scores of 0.977 and 0.976, respectively.

For future works, the proposed model are to be trained with polyphonic audio samples to investigate their ability to identify musical instruments on polyphonic music.

Acknowledgements

The authors are grateful to Tunku Abdul Rahman University of Management and Technology for various support in this research.

References

[1]. C. X. Chiah, L. C. Tay, W. K. Lai, Classifying Musical Instruments Using Temporal and Spectral Features with Deep Neural Nets, in *Proceedings of the 5th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI'2023)*, 7-9 June 2023, Adeje, Tenerife (Canary Islands), Spain 2023, pp. 224-229.

[2]. M. Yun, J. Bi, Deep Learning for Musical Instrument Recognition, *University of Rochester*, 2017.

[3]. T. Zhang, C. J. Kuo, Audio feature extraction and analysis, in Content-Based Audio Classification and Retrieval for Audiovisual data Parsing, Vol. 1, *Springer*, 2001, pp. 35-54.

[4]. F. H. Foomany, K. Umapathy, Classification of music instruments using wavelet-based time-scale features, in *Proceedings of the IEEE Int. Conference Multimed. Expo Work. (ICMEW'13)*, 2013, pp. 1-4.

[5]. R. S. Kothe, D. G. Bhalke, P. P. Gutal, Musical instrument recognition using k-nearest neighbour and Support Vector Machine, in *Proceedings of the IEEE Int. Conference Adv. Electron. Commun. Comput. Technol. (ICAECCT'16)*, 2016, pp. 308-313.

[6]. A. Huaysrijan, S. Pongpinigpinyo, Deep Convolution Neural Network for Thai Classical Music Instruments Sound Recognition, in *Proceedings of the 25th Int. Comput. Sci. Eng. Conference (ICSEC'21)*, 2021, pp. 283-288.

[7]. D. Szeliga, P. Tarasiuk, B. Stasiak, P. S. Szczepaniak, Musical Instrument Recognition with a Convolutional Neural Network and Staged Training, *Procedia Comput. Sci.*, Vol. 207, 2022, pp. 2493-2502.

[8]. S. K. Mahanta, A. F. U. Rahman Khilji, P. Pakray, Deep neural network for musical instrument recognition using MFCCs, *Comput. y Sist.*, Vol. 25, Issue 2, 2021, pp. 351-360.

[9]. N. Rodin, D. Pincic, K. Lenac, D. Susanj, The Comparison of Different Feature Extraction Methods in Musical Instrument Classification, in *Proceedings of the International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'23)*, 2023, pp. 1136-1141.

[10]. S. Masood, S. Gupta, S. Khan, Novel approach for musical instrument identification using neural network, in *Proceedings of the 12th IEEE Int. Conference Electron. Energy, Environ. Commun. Comput. Control (E3-C3'15)*, 2015, pp. 2-6.

[11]. C. Jeyalakshmi, B. Murugeshwari, M. Karthick, HMM and K-NN based automatic musical instrument recognition, in *Proceedings of the Int. Conference IoT Soc. Mobile, Anal. Cloud (I-SMAC'18)*, 2018, pp. 350-355.

[12]. D. G. Bhalke, C. B. Rama Rao, D. S. Bormane, Dynamic time warping technique for musical instrument recognition for isolated notes, in *Proceedings of the Int. Conference Emerg. Trends Electr. Comput. Technol. (ICETECT'11)*, 2011, pp. 768-771.

[13]. S. S. Chakraborty, R. Parekh, Improved musical instrument classification using cepstral coefficients

and neural networks, in Methodologies and Appplication Issues of COntemporary Computing Framework, *Springer*, 2018, pp. 123-138.
[14]. M. Joshi, S. Nadgir, Extraction of feature vectors for analysis of musical instruments, in *Proceedings of the Int. Conference Adv. Electron. Comput. Commun. (ICAECC'14)*, 2014, pp. 1-6.
[15]. M. Blaszke, B. Kostek, Musical Instrument Identification Using Deep Learning Approach, *Sensors*, Vol. 22, Issue 8, 2022, 3033.
[16]. Y. M. G. Costa, L. S. Oliveira, C. N. Silla, An evaluation of Convolutional Neural Networks for music classification using spectrograms, *Appl. Soft Comput. J.*, Vol. 52, 2017, pp. 28-38.
[17]. K. Racharla, V. Kumar, C. B. Jayant, A. Khairkar, P. Harish, Predominant musical instrument classification based on spectral features, in *Proceedings of the 7th Int. Conference Signal Process. Integr. Networks (SPIN'20)*, 2020, pp. 617-622.
[18]. C. Joder, S. Essid, G. Richard, Temporal Integration for Audio Classification With Application to Musical Instrument Classification, *IEEE Trans. Audio, Speech Lang. Process.*, Vol. 17, Issue 1, 2009, pp. 174-186.
[19]. B. Toghiani-Rizi, M. Windmark, Musical Instrument Recognition Using Their Distinctive Characteristics in Artificial Neural Networks, *arXiv Preprint*, 2016, arXiv:1705.04971.
[20]. P. Roy, S. Roy, D. De, TMIR: Transient Length Extraction Strategy for ANN-inspired Musical Instrument Recognition, in *Proceedings of the IEEE Int. Women Eng. Conference Electr. Comput. Eng. (WIECON-ECE'20)*, pp. 267-271, 2020.
[21]. S. Allamy, A. L. Koerich, 1D CNN Architectures for Music Genre Classification, in *Proceedings of the IEEE Symposium Ser. Comput. Intell. (SSCI'21)*, 2021, pp. 01-07.
[22]. S. Gururani, C. Summers, A. Lerch, Instrument activity detection in polyphonic music using deep neural networks, in *Proceedings of the 19th Int. Soc. Music Inf. Retr. Conference (ISMIR'18)*, 2018, pp. 569-576.
[23]. D. Yu, H. Duan, J. Fang, B. Zeng, Predominant Instrument Recognition Based on Deep Neural Network With Auxiliary Classificatio, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 28, 2020, pp. 852-861.
[24]. Y. Han, J. Kim, K. Lee, Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music, *IEEE/ACM Trans. Audio Speech Lang. Process.*, Vol. 25, 2017, pp. 208-221.
[25]. A. Solanki, S. Pandey, Music instrument recognition using deep convolutional neural networks, *Int. J. Inf. Technol.*, Vol. 14, Issue 3, May 2022, pp. 1659-1668.
[26]. C. S. Kakarwal S, Analysis of Musical String Instruments using k_NN, in *Proceedings of the International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC'20)*, 2020, pp. 283-286.
[27]. S. Gururani, M. Sharma, A. Lerch, An attention mechanism for musical instrument recognition, in *Proceedings of the 20th Int. Soc. Music Inf. Retr. Conference (ISMIR'19)*, Vol. 2, 2019, pp. 83-90.
[28]. Dibakar, Music_Instruments_and_2D_figures, https://www.kaggle.com/datasets/dibakarsil/music-instruments-and-2d-figures. [Accessed: 20-Aug-2021].
[29]. K. Choi, D. Joo, J. Kim, Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras, in *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, 2017.
[30]. S. Rajesh, N. J. Nalini, Musical instrument emotion recognition using deep recurrent neural network, *Procedia Comput. Sci.*, Vol. 167, 2020, pp. 16-25.

Published by International Frequency Sensor Association (IFSA) Publishing, S. L., 2023 (http://www.sensorsportal.com).