

EMPIRICAL RESEARCH

Open Access



Transformer-based ensemble method for multiple predominant instruments recognition in polyphonic music

Lekshmi Chandrika Reghunath * and Rajeev Rajan

Abstract

Multiple predominant instrument recognition in polyphonic music is addressed using decision level fusion of three transformer-based architectures on an ensemble of visual representations. The ensemble consists of Mel-spectrogram, modgdgram, and tempogram. Predominant instrument recognition refers to the problem where the prominent instrument is identified from a mixture of instruments being played together. We experimented with two transformer architectures like Vision transformer (Vi-T) and Shifted window transformer (Swin-T) for the proposed task. The performance of the proposed system is compared with that of the state-of-the-art Han's model, convolutional neural networks (CNN), and deep neural networks (DNN). Transformer networks learn the distinctive local characteristics from the visual representations and classify the instrument to the group where it belongs. The proposed system is systematically evaluated using the IRMAS dataset with eleven classes. A wave generative adversarial network (WaveGAN) architecture is also employed to generate audio files for data augmentation. We train our networks from fixed-length music excerpts with a single-labeled predominant instrument and estimate an arbitrary number of predominant instruments from the variable-length test audio file without any sliding window analysis and aggregation strategy as in existing algorithms. The ensemble voting scheme using Swin-T reports a micro and macro F1 score of 0.66 and 0.62, respectively. These metrics are 3.12% and 12.72% relatively higher than those obtained by the state-of-the-art Han's model. The architectural choice of transformers with ensemble voting on Mel-spectro-/modgd-/tempogram has merit in recognizing the predominant instruments in polyphonic music.

Keywords: Predominant, Modified group delay, Mel-spectrogram, Modgdgram, Tempogram, Shifted window

1 Introduction

Music information retrieval (MIR) is a growing field of research with lots of real-world applications and is applied well in categorizing, manipulating, and synthesizing music. An important MIR task of predominant instrument recognition is addressed in this paper. Predominant instrument recognition refers to the problem where the prominent instrument is identified from a mixture of instruments being played together [1].

The task of identifying the leading instrument in polyphonic music is challenging due to the presence of interfering partials in the orchestral background. The auditory scene produced by a musical composition can be regarded as a multi-source environment, where different sound sources are played at various pitches and loudness, and even the spatial position of a given sound source may vary with respect to time [2]. Automatic identification of lead instruments is important, since the performance of the source separation can be improved significantly by knowing the type of the instrument [1]. If the instrument information is included in the tags, it allows people to search for music with the specific instrument they want. Audio enhancement based on instrument-specific equal-

*Correspondence: clekshmir04@gmail.com

Department of Electronics and Communication Engineering, College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Trivandrum, India

ization is also in high demand in music processing. It also helps to enhance fundamental MIR tasks like auto-tagging [3], and automatic music transcription [4].

An extensive review of approaches for isolated musical instrument classification can be found in [5]. Non-negative matrix factorization (NMF) model [6], end-to-end model [7], fusion model with spectral, temporal, and modulation features [8] can be referred to as initial attempts for the proposed task in a polyphonic environment. More recent works deal with instrument recognition in polyphonic music, which is a more demanding and challenging problem. A method for automatic recognition of predominant instruments with support vector machine (SVM) classifiers trained with features extracted from real musical audio signals is proposed in [2]. Bosch et al. improved this algorithm with source separation in a preprocessing step [9]. Han et al. [1] developed a deep CNN for instrument recognition based on Mel-spectrogram inputs and aggregation of multiple outputs from sliding windows over the audio data. Pons et al. [10] analyzed the architecture of Han et al. in order to formulate an efficient design strategy to capture the relevant information about timbre. Both approaches were trained and validated by the IRMAS dataset of polyphonic music excerpts. Detecting the activity of music instruments using a deep neural network (DNN) through a temporal max-pooling aggregation is addressed in [11]. Dongyan Yu et al. [12] employed a network with an auxiliary classification scheme to learn the instrument categories through multitask learning. Gomez et al. [13] investigated the role of two source separation algorithms as pre-processing steps to improve the performance in the context of predominant instrument detection tasks. It was found that both source separation and transfer learning could significantly improve the recognition performance, especially for a small dataset composed of highly similar musical instruments. In [14], the Hilbert-Huang transform (HHT) is employed to map one-dimensional audio data into two-dimensional matrix format, followed by CNN to learn the affluent and effective features for the task. The proposed work in [15] employed an attention mechanism and multiple-instance learning (MIL) framework to address the challenge of weakly labeled instrument recognition in the OpenMIC dataset.

The modified group delay feature (MODGDF) is proposed for pitched musical instrument recognition in an isolated environment in [16]. While the commonly applied Mel-frequency cepstral coefficients (MFCC) feature is capable of modeling the resonances introduced by the filter of the instrument body, it neglects the spectral characteristics of the vibrating source, which also play their role in human perception of musical sounds [17]. Incorporating phase information attempts to preserve this neglected component. It has already been estab-

lished in the literature that the modified group delay function emphasizes peaks in spectra well [18]. It has also been shown in [19] that sinusoids in noise can be estimated well using group delay function. Furthermore, it was shown that even for shorter windows, the phase spectrum could contribute as much as the magnitude spectrum to speech intelligibility [20]. In our work, we are introducing phase-based modgdgram as a complementary feature to magnitude-based spectrogram in recognizing predominant instruments from a polyphonic environment. The source information is completely suppressed in the modgdgram compared to the spectrogram, and the system-specific information is retained, which is a vital clue in instrument identification.

Tempo-based features are employed in various music information retrieval tasks. Grosche et al. point out the potential of integrating the concept of tempo representation into music structural segmentation [21]. Tempo-based features have also been used for cross-version novelty detection in [22]. In [23], an ensemble of VGG-like CNN classifiers were trained on non-augmented, pitch-synchronized, tempo-synchronized, and genre-similar excerpts of IRMAS for the proposed task. They employed tempo-syncing as one of the data augmentation techniques and achieved better results than the baseline model.

The fusion of multiple modalities can offer significant performance gains over using a modality alone and is widely used in recent music processing applications [24–26]. The performance of the various features depends on the instrument characteristics and other unknown factors, and no one feature consistently outperforms all others. Consequently, researchers have investigated the possibility of fusing multiple features to take advantage of their strengths. In our work, we utilize transformer architectures to learn instrument-specific characteristics using Mel-spectro-/modgd-/tempogram to estimate predominant instruments from polyphonic music. Transformer-based systems have outperformed previous approaches for various natural language processing (NLP) and computer vision tasks [27],[28].

2 Contributions

The major contributions of the proposed experiment can be summed up as:

- 1 Introducing modgdgram and tempogram as complementary features to the conventional Mel-spectrogram representation for predominant instrument recognition. The proposed ensemble voting technique makes use of the potential of three visual representations in making a final decision on recognizing predominant instruments in a polyphonic environment.

- 2 We present a high capacity transformer model for Mel-spectrogram inputs. Our model is derived from [29] with some significant changes as described in Section 4, and it outperforms the existing models, including [1]. The efficacy of transformer models and attention mechanisms are demonstrated by comparison with CNN and DNN architectures.
- 3 We explore the time-domain strategy of synthetic music audio generation for data augmentation using WaveGAN. The proposed task is addressed with and without data augmentation.
- 4 In the development phase, the performance is evaluated using various schemes like Mel-spectrogram, modgdgram, and tempogram followed by ensemble voting.

The outline of the rest of the paper is as follows. Section 3 explains the proposed system. The model architectures are described in Section 4. The performance evaluation is explained in Section 5 followed by the analysis of results in Section 6. The paper is concluded in Section 7.

3 System description

The proposed method of Vision transformer (Vi-T) and Shifted window transformer (Swin-T) are shown in Figs. 2 and 3 respectively. In the proposed model, transformers are used to learn the distinctive characteristics of Mel-spectro/modgd/tempo-gram to identify the leading

instrument in a polyphonic context. As a part of data augmentation, additional training files are generated using WaveGAN (Fig. 1). The probability values reported at the nodes of the trained model are mapped as the scores for a test file input. The final decision on the test file is based on soft voting. Soft voting involves summing the predicted probabilities for class labels (from three networks) followed by thresholding. The candidates above the particular threshold were considered as predominant instruments. The performance of the proposed system is compared with that of the state-of-the-art Han's model and a DNN model. A detailed description of each phase is given in the following subsections.

3.1 Feature extraction

3.1.1 Mel-spectrogram

Mel-spectrogram is widely used in speech and music processing applications [30], [31]. Mel-spectrogram approximates how the human auditory system works and can be seen as the spectrogram smoothed, with high precision in the low frequencies and low precision in the high frequencies [32]. All audio files in the IRMAS dataset are in a 16-bit stereo .wav format with a sampling rate of 44,100 Hz. The time-domain waveform is converted to a time-frequency representation using a short-time Fourier transform (STFT) with a frame size of 50 ms and hop size of 10 ms. Then the linear frequency scale obtained spectrogram is converted to a Mel-scale using 128 for the number of Mel-frequency bins.

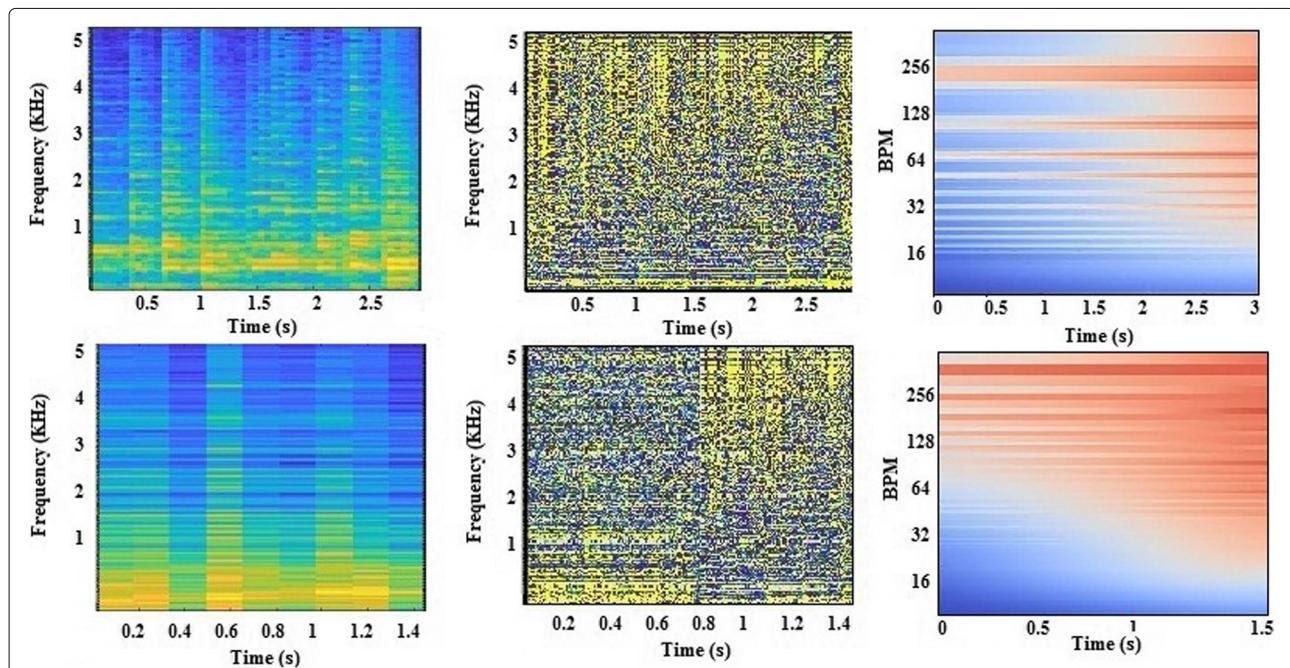


Fig. 1 Visual representation of an audio excerpt with acoustic guitar as leading, Upper pane represents the Mel-spectrogram, modgdgram, and tempogram of the original audio file and lower pane represents the WaveGAN generated files

3.1.2 Modified group delay functions and modgdgram

Group delay features are being employed in numerous speech and music processing applications [18],[33]. The group delay function is defined as the negative derivative of the unwrapped Fourier transform phase with respect to frequency. Group delay functions, $\tau(e^{j\omega})$ are mathematically defined as

$$\tau(e^{j\omega}) = -\frac{d\{\arg(X(e^{j\omega}))\}}{d\omega} \quad (1)$$

where $X(e^{j\omega})$ is the Fourier transform of the signal $x[n]$ and $\arg(X(e^{j\omega}))$ is the phase function. It can be computed directly from the signal, $x[n]$ by [34],

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|X(e^{j\omega})|^2} \quad (2)$$

where the subscripts R and I denote the real and imaginary parts and $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the Fourier transforms of $x[n]$ and $n \cdot x[n]$ (signal multiplied with index), respectively. The spiky nature of the group delay spectrum due to zeros that are located close to the unit circle can be suppressed by replacing the term $|X(e^{j\omega})|$ in the denominator of Eq. (2) with its cepstrally smoothed version, $S(e^{j\omega})$ thereby resulting in modified group delay functions (MODGD) [18]. The modified group delay functions are obtained by,

$$\tau_m(e^{j\omega}) = \left(\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|} \right) (|\tau_c(e^{j\omega})|)^\alpha, \quad (3)$$

where,

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2\gamma}}. \quad (4)$$

Two new parameters, α and γ ($0 < \alpha \leq 1$ and $0 < \gamma \leq 1$) are introduced to control the dynamic range of MODGD [18]. Modgdgram is the visual representation of MODGD with time and frequency in the horizontal and vertical axis, respectively. In a third dimension, the amplitude of the group delay function at a particular time is represented by the intensity or color of each point in the image. Modgdgrams are computed with a frame size of 50 ms and a hop size of 10 ms. The parameters α and γ have been empirically chosen as 0.9 and 0.5, respectively. Mel-spectrograms and modgdgrams are implemented using MATLAB.

Typically in spectrograms, we can see pitch components and their harmonics as striations along with formant structure. But system-specific information (formant tracks) is enhanced in modgdgram by suppressing the source information. In music, the body of the musical instrument is the counterpart of the vocal tract (system) in speech. Davis et al. [35] claim that timbres are properties of musical instruments which rely on the physical characteristics of the instrument. Thus, timbre makes a particular musical instrument or the human voice and

produces a different sound from another, even when they play or sing the same note.

3.1.3 Tempogram

A tempogram is a time-pulse representation of an audio signal laid out such that it indicates the variation of pulse strength over time given a specific time lag l or a beats per minute (BPM) value.[36]. It is a time-tempo representation that encodes the local tempo of a music signal over time. The calculation of the tempogram is based on the assumption that music exhibits coherent and locally periodic patterns. These patterns may be characterized by peaks in the autocorrelation function (ACF) of the onset detection function [36] at certain time lags. The training and testing audio files are read and processed using the Librosa framework. The principle of autocorrelation is used to estimate the tempo at every segment in the novelty function [37]. Autocorrelation tempograms are computed with *librosa.feature.tempogram* using a 2048 point FFT window and a hop size of 512.

4 Model architectures

4.1 DNN

A DNN framework on musical texture features (MTF) is experimented with to examine the performance of deep learning methodology on handcrafted features. MTF includes MFCC (13 dim), spectral centroid, spectral bandwidth, root mean square energy, spectral roll-off, and chroma STFT. The features are computed with a frame size of 40 ms and a hop size of 10 ms using Librosa framework¹. The DNN consists of seven layers, with increasing units from 8 to 512. Regarding the activation function, ReLU has been chosen for hidden layers and softmax for the output layer. The approach attempted in [38] has been customized for multi-label classification and has been experimented with to analyze the role of machine learning techniques, especially using the MTF-SVM framework.

4.2 CNN

CNN uses a deep architecture with repeated convolutions followed by max-pooling. A total of five layers are used with the number of filters starting from 32 to 512 for Mel-spectrogram processing. The first two layers used 5×5 filters, and the remaining layers used 3×3 filters. Using filters of different shapes seems an efficient way of learning spectrogram-based CNNs [10]. To achieve the best performance, the optimal filter size is usually chosen empirically by either experimental validation or visualization for each convolutional layer [39]. The initial layers help to extract general features and also help in noise reduction. The last convolutional layers used 3×3 filters as later layers reveal more specific and complex

¹<https://librosa.org/doc/latest/tutorial.html>

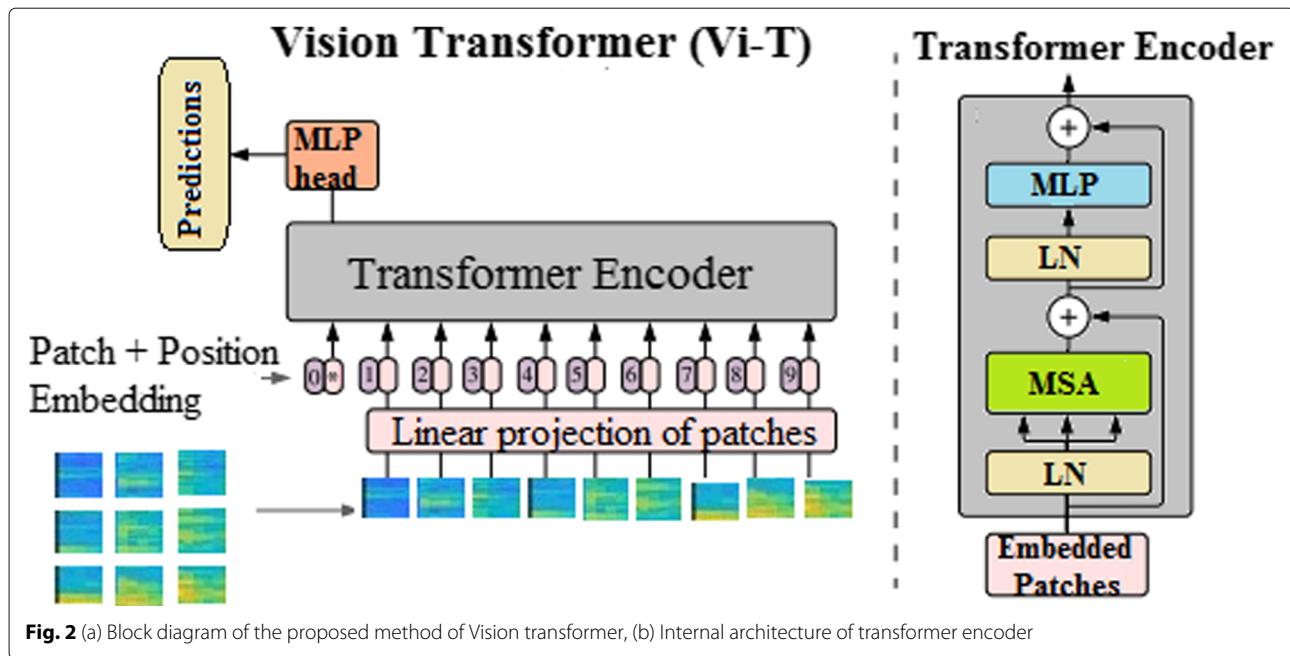


Fig. 2 (a) Block diagram of the proposed method of Vision transformer, (b) Internal architecture of transformer encoder

patterns and final layers activations help to recognize the predominant instruments from accompaniments. Global max-pooling is adopted in the final max-pooling layer, which is then fed to a fully connected layer. For modgram processing, we used six convolutional layers with the number of filters increasing from 8 to 256, followed by 2×2 max pooling. We used filters of size 3×3 in all six layers with a fixed stride size of one. For tempogram processing, we used the same model as Mel-spectrogram. A dropout of 0.5 is introduced after the fully connected layer to avoid overfitting in all processing. Leaky ReLU with $\alpha = 0.33$ in hidden layers has been empirically chosen for optimum performance in Mel-spectrogram processing. But in modgram and tempogram processing, the best performance is obtained for ReLU. Softmax is used as the activation function for the output layer.

4.3 Vi-T

Inspired by the success of Transformer [27] in various natural language processing tasks, Vision Transformers (ViT) [40] constitute the first pure transformer-based architecture that can achieve good performance on the image recognition task. Figure 2 shows the architecture of our proposed method. As shown in Fig. 2(a), the input image $x \in R^{H \times W \times L}$, where H, W, and L represent the height, width, and the number of channels of the image x . The input image is partitioned into non-overlapping patches, called tokens. In our work, we choose $M = 6 \times 6$, where M is the size of a patch. Then each patch is linearly projected to a dimension of 64, along with position embeddings, and feeds the resulting sequence of vectors to a standard transformer encoder. The number of patches, $P=HW/M^2$. The various hyperparameters selected for our proposed

method are shown in Table 1. Position embeddings are added to the patch embeddings to retain positional information. The Transformer encoder is shown in Fig. 2(b) and consists of alternating layers of multi-headed self-attention (MSA) with eight attention heads and two multi layer perceptron (MLP) layers with 2048 and 1024 nodes with Gaussian error linear unit (GELU) nonlinearity in between. Layernorm (LN) is applied before every MSA and MLP layer, and residual connections are placed after each module. MSA is defined in [27] as

$$MSA(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_8) W^O \quad (5)$$

where,

$$\text{head}_i = \text{Attention} \left(QW_i^Q, KW_i^K, VW_i^V \right) \quad (6)$$

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (7)$$

Table 1 Various hyperparameters chosen for Vi-T and Swin-T

Hyperparameter	Vi-T	Swin-T
Image size	72×72	72×72
Patch dimension	6×6	4×4
Hyper parameter (C)	64	96
Number of heads	8	8
Number of windows	NA	4
Number of MLP nodes	2048, 1024	256, 256
Mini batch-size	256	32

where Q , K , V represents the query key and value vector respectively and W_i^Q , W_i^K , W_i^V and W^O are the weight matrices of query, key, value, and output vectors, respectively [27], while d_k is the dimension of the query vector. The outputs from all 8 attention heads are concatenated to form a single output vector before passing it through the feed-forward network. The model is then trained on instrument classification in a supervised manner.

4.4 Swin-T

The main drawback of ViT is that it produces feature maps of a single low resolution and has quadratic computation complexity to input image size due to computation of self-attention globally. Also, the tokens are of fixed scale and are thus unsuitable for vision applications. Unlike other transformers Swin-T [29] has a hierarchical architecture and has linear computational complexity through the proposed shifted window-based self-attention approach. The computational complexity of Vi-T is given by [29]

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \quad (8)$$

$$\Omega(W - MSA) = 4hwC^2 + 2M^2hwC \quad (9)$$

The computational complexity drops for Swin-T as per the Eq. (9) above. MSA has quadratic computational complexity to patch number hw , while W-MSA has linear computational complexity due to the shifted window approach [29]. Figure 3 shows the architecture of our proposed method. The input image is partitioned into non-overlapping patches, called tokens during patch partitioning. In our work, we choose $M = 4 \times 4$, where M is the size of a patch. The second step is linear embedding, in which the eigenvalues in the feature map are projected to a C dimensional vector. The hyperparameter C has been empirically chosen as 96 for our work. The various selected hyperparameters for our proposed method are shown in Table 1. The output of the patch embedding layer leads to two Swin Transformer networks. The output of the second Swin-T network is applied to a patch merging layer. Patch merging works in a similar

way to CNN's pooling layer by concatenating the features of each group of neighboring patches and applying a linear embedding layer to change the output dimension to $2C$. Hence the output of patch merging layer is $(\frac{H}{8} \times \frac{W}{8} \times 2C)$ and is followed by global average pooling and a dense layer with 11 nodes and a softmax activation function. Figure 3(b) shows the internal architecture of the Swin-T block. Shifted windows approach is used in the encoder to address the multi-head self-attention (MSA) scheme. The output of the patch embedding layer is divided into non-overlapping windows (in our work, we choose $N = 4$, where N is the number of windows). Here to compute the self-attention of a given patch within that window, we ignore the rest of the patches in other windows. As illustrated in Fig. 3(b), W-MSA is the windowed multi-head self-attention in which we divide the patched image into non-overlapping windows and compute attention for patches within the window. In SW-MSA, the window is stride forwarded by two patches just like the kernel striding in CNN and computing attention within that window. For the empty patches, the process was repeated after zero padding. W-MSA and SW-MSA are followed by a 2-layer MLP each with 256 nodes and GELU nonlinearity in between. LN is applied before each MSA module and MLP, and a residual connection is applied after each module. The modified equation for attention [29] is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k} + B}\right)V \quad (10)$$

where B is the relative position of window.

5 Performance evaluation

5.1 Dataset

The performance of the proposed system is evaluated using the IRMAS (Instrument Recognition in Musical Audio Signals) dataset, developed by the Music Technology Group (MTG) of Universitat Pompeu Fabra (UPF). It consists of more than 6000 musical audio excerpts from various styles with annotations of the predominant instruments present. All audio files in the IRMAS dataset are

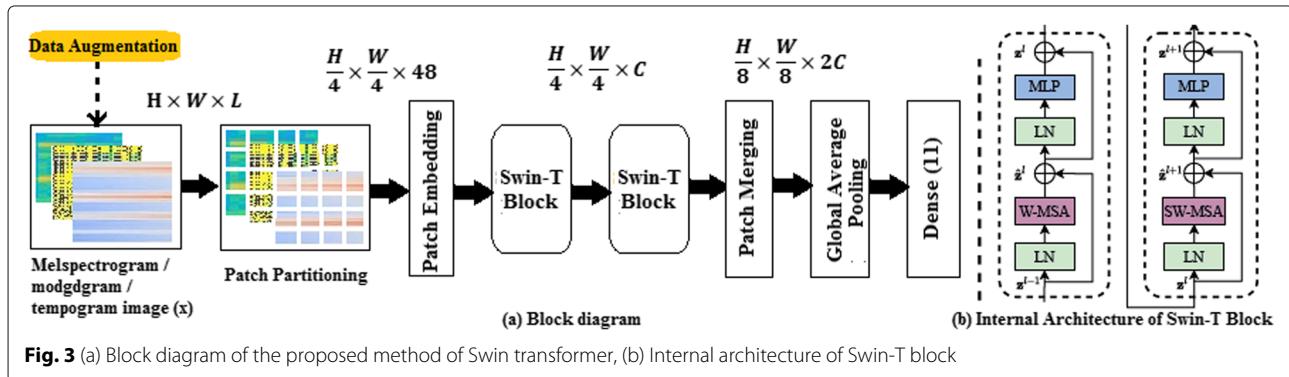


Fig. 3 (a) Block diagram of the proposed method of Swin transformer, (b) Internal architecture of Swin-T block

in a 16-bit stereo .wav format with a sampling rate of 44,100 Hz. IRMAS dataset [2] contains separate training and testing set of eleven classes. The classes include cello (Cel), clarinet (Cla), flute (Flu), acoustic guitar (Gac), electric guitar (Gel), organ (Org), piano (Pia), saxophone (Sax), trumpet (Tru), violin (Vio), and human singing voice (Voice). The training data are single-labeled and consist of 6705 audio files with excerpts of 3 s from more than 2000 distinct recordings. On the other hand, the testing data are multi-labeled and consist of 2874 audio files with lengths between 5 and 20 s and contain multiple predominant instruments. This dataset has two disadvantages when training models. First, the number of audio files available for certain instruments like cello, clarinet, and flute is less than 500, and the models trained with the data are hardly generalizable. Second, the dataset is not well balanced in terms of either musical genre or instrumentation. However, this may not be a problem if the datasets were larger and the distribution represented the real world. Data augmentation offers an excellent solution to this issue. Data augmentation means training the deep network with additional diverse data. This increases the generalization capability of the network and thus reduces overfitting.

5.2 Data augmentation using WaveGAN

Generative adversarial networks (GAN) have been successfully applied to a variety of problems in image generation [41] and style transfer [42]. WaveGAN architecture is similar to deep convolutional GAN (DCGAN), which is used for Mel-spectrogram generation in various music processing applications. The DCGAN generator uses transposed convolution to iteratively upsample low-resolution feature maps into a high-resolution image. In WaveGAN architecture, the transposed convolution operation is modified to widen its receptive field. Specifically, longer one-dimensional filters of length 25 are used instead of two-dimensional filters of size 5×5 and the intermediate representation is upsampled by a factor of four instead of two at each layer. The input to the generator is a random sample taken from a uniform distribution between -1 and 1 and is projected and reshaped to the dimension 16×1024 . This is followed by six transpose convolution layers that upsample the input feature map to a fine and detailed output. The output of the generator is 65,536 samples (corresponding to 4.01 s of audio at 16 kHz). It is also capable to produce 1.49 s of audio at 44.1khz by choosing the slice length of 65536 samples. The output of the generator is directly applied to the input of the discriminator. The discriminator is an efficient CNN that discriminates between real and generated samples. The discriminator is also modified similarly, using length-25 filters in one dimension and increasing stride from two to four which results in WaveGAN architecture [43].

The transposed convolution in the generator produces checkerboard artifacts [43]. To ensure that the discriminator does not learn these artifacts, we use phase shuffle operation (with hyperparameter $n=2$) as suggested in [43]. ReLU is used as the activation for transposed convolution layers and LReLU with $\alpha = 0.2$ is chosen for convolution operation. Finally, the system is trained using the Wasserstein GAN with gradient penalty (WGAN-GP) strategy [44] to tackle the vanishing gradient problem and enhance training stability. For training, the WaveGAN optimizes WGAN-GP using Adam for both generator and discriminator. A constant learning rate of 0.0001 is used with $\beta_1 = 0.5$ and $\beta_2 = 0.9$.

WaveGAN is trained for 2000 epochs on the three-sec audio files of each class to generate similar audio files based on a similarity metric (s) [45] with an acceptance criterion of $s > 0.1$. The values of parameters and hyperparameters associated with WaveGAN for our experiments are listed in Table 2. A total of 6585 audio files with cello (625), clarinet (482), flute (433), acoustic guitar (594), electric guitar (732), organ (657), piano (698), saxophone (597), trumpet (521), violin (526), and voice (720) are generated. Training files available in the corpus are denoted by $Train_{DB}$ and the generated files are added to the available training corpus, and the augmented corpus is denoted by $Train_{AugDB}$. Mel-spectrogram, modgdgram, and tempogram of natural and generated audio files for acoustic guitar are shown in Fig. 1. The experiment details and a few audio files can be accessed at <https://sites.google.com/view/audiosamples-2020/home/instrument>.

The quality of generated files is evaluated using a perceptual test. It is conducted with ten listeners to assess the quality of generated files for 275 files covering all classes. Listeners are asked to grade the quality by choosing one among the five opinion grades varying from poor to excellent quality (scores, 1 to 5). A mean opinion score (MOS) of 3.64 is obtained. This value is comparable to the MOS score obtained in [43] and [46] using WaveGAN.

5.3 Experimental set-up

The experiment is progressed in four phases, namely Mel-spectrogram-based, modgdgram-based, and tempogram-

Table 2 Various hyperparameters chosen for WaveGAN

Name	Value
WavGAN Latent dimension	100
Number of channels	1
WavGAN dimension	32
Training batch size	64
Kernel length	25
Generation length	65,536 samples
Loss	WGAN-GP ($\lambda = 10$)
D updates per G updates	5

based, followed by soft voting. Hard or majority voting is not used in our method since the presence of simultaneously occurring partials degrades its performance [1]. Han's model [1] is implemented with 1 s slice length for performance comparison. In their approach, sigmoid outputs obtained by sliding window analysis on Mel-spectrogram inputs were aggregated followed by thresholding, and the candidates above that particular threshold were considered as predominant instruments. In our proposed method of soft voting, the predicted probabilities from three networks are summed followed by thresholding. We choose a threshold value of 0.5 empirically as it helps to recognize most of the predominant instruments [1].

5.3.1 Training configuration

The DNN network is trained with categorical cross-entropy loss function using Adam optimizer with a learning rate of 0.001 and a mini-batch size of 128. For CNN networks, we choose a batch size of 128 and an Adam optimizer with a categorical cross-entropy loss function. For Vi-T, we used categorical cross-entropy loss function using Adam optimizer, with a learning rate of 0.001 and weight decay of 0.0001, and the mini-batch size was set to 256. For Swin-T we used categorical cross-entropy using the Adam optimizer, with a learning rate of 0.001 and gradient clip value of 0.5, and the mini-batch size was set to 32. 20% of training data is used for tuning the hyperparameters during validation for all the models. The training was stopped when the validation loss did not decrease for more than two epochs.

5.3.2 Testing configuration

2874 polyphonic files of variable length with multiple predominant instruments are used for the testing phase.

Since the number of annotations for each class was not equal, we computed precision, recall, and F1 measures for both the micro and the macro averages. For the micro averages, we calculated the metrics globally, thus giving more weight to the instrument with a higher number of appearances. On the other hand, we calculated the metrics for each label and found their unweighted average for the macro averages.

6 Results and analysis

The overall performance of different phases of the Swin-T experiment with data augmentation $Train_{AugDB}$ is tabulated in Table 3. Our proposed method of Voting-Swin-T achieved micro and macro F1 measures of 0.66 and 0.62, respectively, which are 3.12% and 12.72% relatively higher than those obtained for Mel-spectrogram-based Han's model. The performance of the various features depends on the instrument characteristics and other unknown factors, and none of the features consistently outperforms all others. The proposed Mel-spectrogram-Swin-T framework shows superior performance for seven instrument classes than Han's model. Our proposed Modgdgram-Swin-T framework shows a competing performance with the state-of-the-art Han's model. While the Han model reports a macro-F1 score of 0.55, our proposed Modgdgram-Swin-T gives 0.51. In the case of modgdgram processing, instruments like the electric guitar, organ, saxophone, trumpet, and violin show enhanced performance over the Mel-spectrogram-Swin-T. They showed improved performance for four instruments than Han's model. It shows the promise of the image processing aspect of modgdgram for predominant instrument recognition. Also, our proposed Tempogram-Swin-T shows similar performance as that of the Mel-spectrogram

Table 3 Precision (P), recall (R), and F1 score for the Swin-T experiments and Han's model with data augmentation

Class	Han's Model			Mel-spectrogram			Modgdgram			Tempogram			Voting		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Cel	0.55	0.55	0.55	0.52	0.58	0.55	0.27	0.40	0.32	0.52	0.46	0.49	0.61	0.62	0.61
Cla	0.11	0.65	0.18	0.47	0.76	0.58	0.24	0.50	0.33	0.44	0.79	0.56	0.36	0.77	0.49
Flu	0.33	0.61	0.43	0.81	0.83	0.82	0.52	0.63	0.57	0.81	0.80	0.80	0.57	0.80	0.66
Gac	0.84	0.63	0.72	0.43	0.62	0.51	0.64	0.47	0.54	0.30	0.64	0.41	0.59	0.60	0.59
Gel	0.69	0.69	0.69	0.70	0.52	0.60	0.57	0.55	0.56	0.78	0.42	0.55	0.73	0.59	0.66
Org	0.45	0.46	0.45	0.59	0.53	0.56	0.44	0.55	0.49	0.67	0.53	0.59	0.53	0.58	0.55
Pia	0.76	0.61	0.67	0.61	0.54	0.57	0.71	0.47	0.56	0.51	0.50	0.51	0.81	0.51	0.63
Sax	0.62	0.61	0.61	0.68	0.55	0.61	0.53	0.57	0.55	0.78	0.48	0.59	0.61	0.51	0.56
Tru	0.47	0.42	0.44	0.59	0.68	0.63	0.50	0.72	0.59	0.62	0.66	0.64	0.58	0.74	0.65
Vio	0.41	0.57	0.48	0.53	0.59	0.56	0.40	0.63	0.49	0.56	0.55	0.56	0.59	0.73	0.65
Voice	0.94	0.78	0.85	0.70	0.79	0.75	0.57	0.59	0.58	0.77	0.80	0.78	0.69	0.90	0.78
Macro	0.56	0.60	0.55	0.60	0.63	0.61	0.49	0.55	0.51	0.62	0.60	0.59	0.61	0.67	0.62
Micro	0.64	0.64	0.64	0.62	0.62	0.62	0.54	0.54	0.54	0.58	0.58	0.58	0.66	0.66	0.66

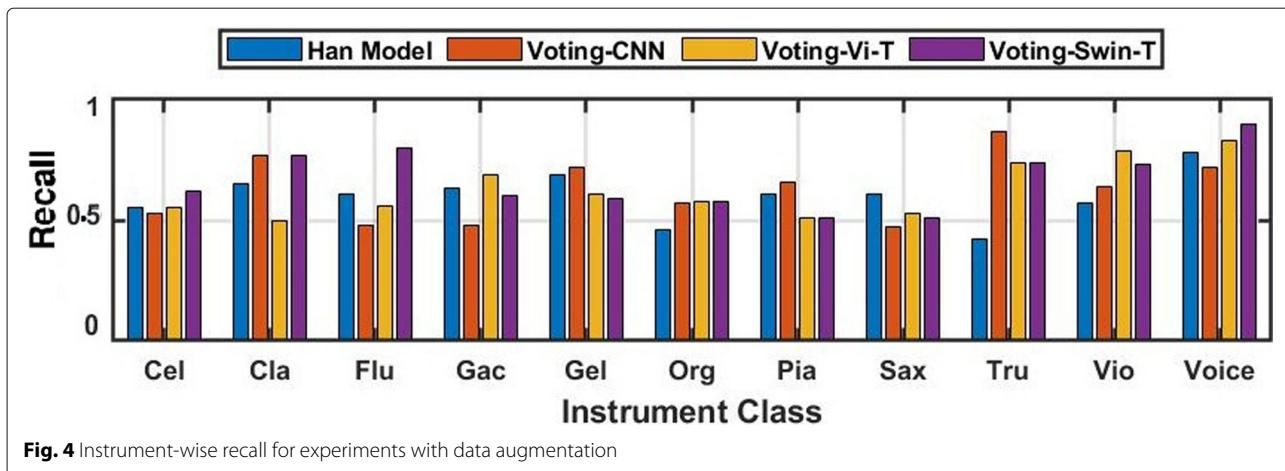


Fig. 4 Instrument-wise recall for experiments with data augmentation

network and reports a better macro score than Han's model. It shows superior performance for five instrument classes than Han's model and three instruments over our proposed Mel-spectrogram-Swin-T network. Thus our proposed voting-Swin-T and Mel-spectrogram-Swin-T showed improved performance than the state-of-the-art Han's model.

6.1 Analysis of instrument-wise identification performance

The instrument-wise recall for all our voting experiments with data augmentation is shown in Fig. 4. The proposed Voting frameworks showed superior performance to the state-of-the-art Han's model. In the case of ensemble voting using CNN, instruments like the clarinet, electric guitar, piano, and trumpet show improved performance over Han's model. In the case of voting using transformers, seven instruments showed improved performance over Han's model. For all the voting techniques, the voice reports a high recall due to its distinct spectral characteristic [1].

6.2 Effect of data augmentation

For deep learning, the number of training examples is critical for the performance compared to the case of using hand-crafted features because it aims to learn a feature from the low-level input data [1]. The problem with small datasets is that models trained with them do not generalize well from the validation and test set [47]. Han's model using ($Train_{DB}$) reports a low F1 score of about 0.20 for cello, and they suggest that it is due to the insufficient number of training samples [1]. The same experiment when repeated using $Train_{AugDB}$ and our Mel-spectrogram-Swin-T showed an improved F1 score validates the claim in [1].

The significance of data augmentation in the proposed model can be analyzed from Table 4. While the proposed method of Voting-Swin-T, without data augmentation

($Train_{DB}$), reports micro and macro F1 score of 0.59 and 0.60, respectively, the metrics improved to 0.66 and 0.62, respectively, for the data augmentation scheme. It shows an improvement of 11.86% and 3.33% relatively higher than that obtained for experiments with $Train_{DB}$. Similar performance improvement is observed for Han's model and MTF-SVM and DNN frameworks.

6.3 Effect of transformer architecture and attention

The instrument-wise F1 scores for all the Mel-spectrogram experiments are shown in Fig. 5. The model using CNN alone does not show improved performance as expected; this is mainly because of the difficulty in predicting the multiple predominant instruments from the variable-length testing file while training with single predominant fixed-length training files. Only the instruments with distinct spectral characteristics and voice show good performance. On the other hand, experiments with transformer architecture showed improved performance for all the instruments. This is mainly because the transformer architecture with a multi-head attention mechanism helps to focus or attend to specific regions of the visual representation for predominant instruments recognition. Another important point is that it requires very few trainable parameters to learn the model, which

Table 4 Effect of data augmentation. The highest values are highlighted

SL.No	Model	$Train_{DB}$		$Train_{AugDB}$	
		Micro F1	Macro F1	Micro F1	Macro F1
1	Han Model[1]	0.60	0.50	0.64	0.55
2	K.Racharla et al. [38] (MTF-SVM)	0.22	0.19	0.25	0.23
3	MTF-DNN	0.32	0.28	0.38	0.35
4	Proposed method -Voting-Swin-T	0.59	0.60	0.66	0.62

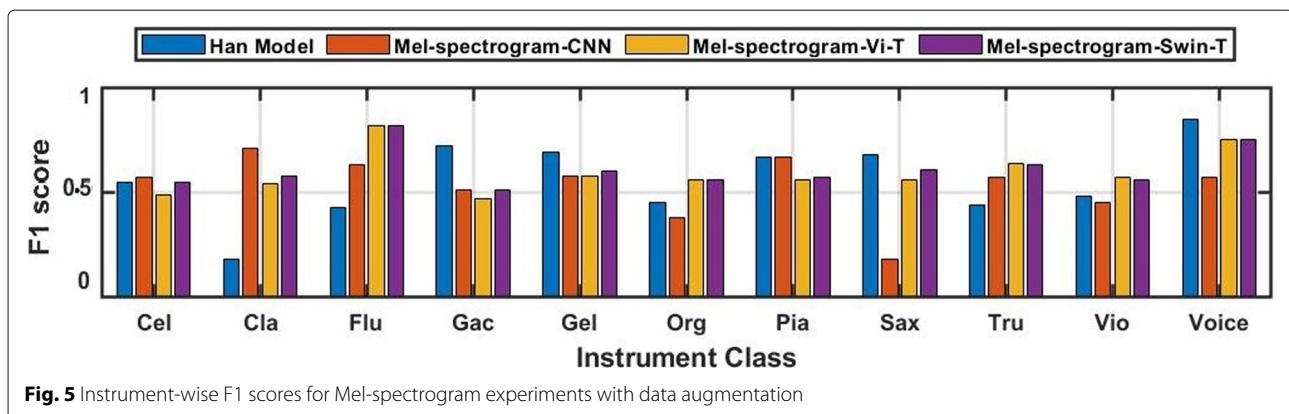


Fig. 5 Instrument-wise F1 scores for Mel-spectrogram experiments with data augmentation

helps to reach convergence faster than the models employing CNN alone. For polytimbral music instrument recognition attention model focuses on specific time segments in the audio relevant to each instrument label. The ability of the attention model to weigh relevant and suppress irrelevant predictions for each instrument leads to better classification accuracy [15]. Compared to self-attention multi-head attention gives the attention layer multiple representation subspaces, and as the image passes through different heads, predictions about the predominant instruments are more refined than employing single head self-attention. In the case of ViT, we have to compute the self-attention for a given patch with all the other patches in an input image. On the other hand, Swin-T with shifted window scheme gives the effect of kernel striding in CNN which along with multi-head attention helps to recognize multiple predominant instruments with linear computational complexity.

We also conducted an ablation study of the architecture in order to gain a better understanding of the network's behavior. We investigated the performance by changing the number of heads, patch size, projection dimension, and the number of MLP nodes. The results are tabulated in Table 5. The optimal parameters obtained through Mel-spectrogram analysis are applied to the modgdgram and

tempogram architectures through a similar ablation study. To summarize, the results show the potential of Swin-T architecture and the promise of alternate visual representations other than the conventional Mel-spectrograms for predominant instruments recognition tasks.

6.4 Effect of voting and ablation study of ensemble

Several studies [48, 49] have demonstrated that by consolidating information from multiple sources, better performance can be achieved compared to uni-modal systems which motivated us to perform the ensemble voting method. We also conducted the ablation study of the ensemble to evaluate the contribution of the individual parts in the proposed ensemble classification framework for predominant instrument recognition. Since there are three visual representations, we have experimented with different fusion schemes as shown in Table 6. Table 6 reports F1 measures for different fusion strategies trained with $Train_{AugDB}$. Spect, Modgd, and Tempo refer to Mel-spectrogram-Swin-T, Modgdgram-Swin-T, and Tempogram-Swin-T respectively.

It is important to note that Spect + Modgd and Modgd + Tempo show improvement in macro measures compared to Mel-spectrogram-based Han's model. This shows the importance of phase information in the proposed

Table 5 Ablation study of the Mel-spectrogram architecture showing the effect of number of heads, patch size, projection dimension, and number of MLP nodes. Highest values are highlighted

SL.No	Architecture Spec.	Option	Vi-T		Swin-T	
			F1 Micro	F1 Macro	F1 Micro	F1 Macro
1	Number of heads	4	0.58	0.52	0.62	0.53
		8	0.59	0.57	0.62	0.61
2	Patch size	4 x 4	0.55	0.50	0.62	0.61
		6 x 6	0.59	0.57	0.60	0.53
3	Projection dimension	64	0.59	0.57	0.56	0.51
		96	0.58	0.51	0.62	0.61
4	Number of MLP nodes	2048,1024	0.59	0.57	0.62	0.55
		256,256	0.57	0.51	0.62	0.61

Table 6 Ablation study of ensemble with data augmentation. Spect, Modgd, and Tempo refer to Mel-spectrogram-Swin-T, Modgdgram-Swin-T, and Tempogram-Swin-T, respectively. + denotes soft voting

Sl.No	Ensemble	F1 Micro	F1 Macro
1	Spect + Modgd.	0.59	0.60
2	Spect + Tempo.	0.64	0.60
3	Modgd + Tempo.	0.57	0.55
4	Spect + Modgd + Tempo.	0.66	0.62

task. Conventionally, the spectrum-related features used in instrument recognition take into account merely the magnitude information. However, there is often additional information concealed in the phase, which could be beneficial for recognition as seen in [16]. In the case of tempogram, Spect + Tempo showed improved performance over Han's model. The advantage of onsets in extracting informative cues about musical instrument recognition is proposed in [50]. Human listeners can easily identify instrument sounds from onset portions compared to other portions of the sound. Cemgil et al. [51] define the "tempogram" which induces a probability distribution over the pairs (pulse period, pulse phase) given the onsets. In most automated tempo and beat tracking approaches, the first step is to estimate the positions of note onsets within the music signal. Results of the experiments described in [52] suggested that the presence of onsets was beneficial, in particular for instrument sounds. Since onset detection is the primary step in computing tempogram, it can provide useful information about predominant instruments. The experimental results validate the claim in [52]. The advantage of voting is that it is unlikely that all classifiers will make the same mistake, as long as every error is made by a minority of the classifiers, an optimal classification can be achieved [53]. Since the ensemble soft voting of three representations results in better performance, we opted for the same as the final scheme and our proposed ensemble frameworks outperform the state-of-the-art Han's model.

6.5 Comparison to existing algorithms

The performance metrics for various algorithms on the IRMAS corpus are reported in Table 7. The number of trainable parameters is also indicated.

Bosch et al. [9] modified the Fuhrmann's algorithm [2] and used typical hand-made timbral audio features with their frame-wise mean and variance statistics to train SVMs with a source separation technique called flexible audio source separation framework (FASST) in a pre-processing step. It reports a micro and macro F1 score of 0.50 and 0.43 respectively, and it is evident that the proposed ensemble frameworks outperform the hand-crafted

Table 7 Performance comparison on IRMAS dataset. The best result in the proposed scheme is highlighted in red

Sl.No	Model / parameters	F1 Micro	F1 Macro
1	Bosch et al. [9]	0.50	0.43
2	Han et al. [1] /1446k	0.60	0.50
3	Pons et al. [10] /743k	0.59	0.52
4	Proposed method - Voting-CNN / 2040k	0.63	0.57
5	Proposed method - Voting-ViT / 1079k	0.65	0.60
6	Proposed method - Voting-Swin-T / 350k	0.66	0.62

features. The MTF-SVM approach [38] has not shown good performance as expected. The state-of-the-art Han's model (*TrainDB*) [1] reports micro and macro F1 score of 0.60 and 0.50 respectively. Han Model (*TrainAugDB*) reports micro and macro F1 score of 0.64 and 0.55 respectively. The proposed voting model using Swin-T reports micro and macro F1 scores of 0.66 and 0.62 respectively. These values are 3.12% and 12.72% relatively higher than the state-of-the-art Han's model. Han et al. [1] developed a deep CNN for instrument recognition based on Mel-spectrogram inputs and aggregation of multiple outputs from sliding windows over the audio data. Pons et al. [10] customized the architecture of Han et al. and introduced two models, namely, single-layer and multi-layer approaches. They used the same aggregation strategy as that of Han's model by averaging the softmax predictions and finding the candidates with a threshold of 0.2. As different from the existing approaches, we estimated the predominant instrument using the entire Mel-spectrogram without sliding window and aggregation analysis. Better micro and macro measures show that it is possible to predict multiple instruments from the visual representations without any sliding window analysis. Also, our proposed Swin-T for Mel-spectrogram requires approximately four times fewer trainable parameters than Han's model [54]. In [15], the usage of an attention layer was shown to improve classification results in the OpenMIC dataset when applied to a set of Mel-spectrogram features extracted from a pre-trained VGG net. While the work focuses on Mel-spectrogram, we experimented with the effect of phase and tempo information along with magnitude information. Our proposed ensemble voting technique outperformed existing algorithms and the MTF DNN and SVM framework on the IRMAS dataset for both the micro and the macro F1 measure.

7 Conclusion

We presented a transformer-based predominant instrument recognition system using multiple visual representations. Transformer models are used to capture

the instrument-specific characteristics and then do further classification. We experimented with Vi-T and the recent Swin-T architectures with a detailed ablation study and our proposed experiments using Swin-T outperform existing algorithms with very less trainable parameters.

We introduced an alternate visual representation to conventionally used Mel-spectrograms. Our study shows that visual representation in terms of modgdgram can be explored in many applications. We believe that optimum parameters may potentially lead to a better visual representation for modified group delay functions. It is worth noting that many recent deep learning schemes in image processing such as transfer learning, attention mechanism, and transformers are transferable to the audio processing domain. Modified group delay functions can be computed directly from the music signal and also from the flattened music spectrum. It is known as direct-modgdgram (or simply "modgdgram") and source-modgdgram, respectively. Direct modgdgram emphasizes system information and source-modgdgram provides information about the multiple sources present in the music signal [55]. Source-modgdgram has been effectively used for melody extraction [56] and multi-pitch estimation [57]. Since we need system information to track the presence of instruments, we employ the direct-modgdgram for the task of instrument recognition.

The proposed method is evaluated using the IRMAS dataset. As observed in many music information retrieval tasks, the data augmentation strategy has also shown its promise in the proposed task. The time-domain strategy of synthetic music generation for data augmentation using WaveGAN is explored. WaveGAN data augmentation for instrument detection is probably a new attempt in predominant instrument recognition. As future work, we would like to focus on synthesizing high-quality audio files using recent high fidelity audio synthesis approaches discussed in [58] and to compare the pipeline of traditional audio augmentations used in many tasks [23] with adversarial audio synthesis. The ensemble voting framework outperforms the existing state-of-the-art algorithms and music texture features DNN and SVM frameworks. The results show the potential of the ensemble voting technique in predominant instrument recognition in polyphonic music.

Acknowledgements

The authors would like to acknowledge Juan J. Bosch, Ferdinand Fuhrmann, and Perfecto Herrera (Music Technology Group - Universitat Pompeu Fabra) for developing the IRMAS dataset and making it publicly available.

Authors' contributions

LCR and RR jointly designed, implemented, and interpreted the computer simulations. RR implemented the modgdgram algorithm. All authors contributed to writing the manuscript and further read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the zenodo repository (<https://www.upf.edu/web/mtg/irmas>) and are publicly available.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 2 December 2021 Accepted: 22 April 2022

Published online: 16 May 2022

References

1. Y. Han, J. Kim, K. Lee, Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Trans. Audio, Speech Lang. Process.* **25**(1), 208–221 (2017)
2. F. Fuhrmann, P. Herrera, in *Proc. of 13th International Conference on Digital Audio Effects (DAFx10) Graz, Austria, September 6-10, 2010*. Polyphonic instrument recognition for exploring semantic similarities in music, (2010), pp. 1–8
3. J.-Y. Liu, Y.-H. Yang, in *Proc. of the 24th ACM Multimedia Conference Amsterdam, Netherlands October 15 - 19, 2016*. Event localization in music auto-tagging (Association for Computing Machinery, New York, 2016), pp. 1048–1057
4. Z. Duan, J. Han, B. Pardo, Multi-pitch streaming of harmonic sound mixtures. *IEEE/ACM Transactions on Audio. Speech Lang. Process.* **22**(1), 138–150 (2013)
5. G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, S. McAdams, The timbre toolbox: Extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.* **130**(5), 2902–2916 (2011)
6. P. Smaragdis, J. C. Brown, in *Proc of IEEE Workshop on Applications of Signal Process. Audio Acoust., New Paltz, NY, 2003*. Non-negative matrix factorization for polyphonic music transcription, (2003), pp. 177–180
7. P. Li, J. Qian, T. Wang, Automatic instrument recognition in polyphonic music using convolutional neural networks. arXiv preprint arXiv:1511.05520 (2015)
8. T. Kitahara, M. Goto, K. Komatani, T. Ogata, H. G. Okuno, Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps. *EURASIP J. Adv. Signal Proc.* **2007**, 1–15 (2006)
9. J. J. Bosch, J. Janer, F. Fuhrmann, P. Herrera, in *Proc. of the 13th International Society for Music Information Retrieval Conference, ISMIR, Porto, Portugal*. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals, (2012), pp. 552–564. <https://doi.org/10.5281/zenodo.1416076>
10. J. Pons, O. Slizovskaia, R. Gong, E. Gomez, X. Serra, in *Proc. of 25th European Signal Processing Conference Kos International Convention Centre (KICC), Psalidi, Kos Island, August 28 to September 2, 2017*. Timbre analysis of music audio signals with convolutional neural networks (IEEE, 2017), pp. 2744–2748
11. S. Gururani, C. Summers, A. Lerch, in *Proc. of 19th International Society for Music Information Retrieval Conference Paris, France. September 23-27, 2018*. Instrument activity detection in polyphonic music using deep neural networks, (2018), pp. 569–576. <https://doi.org/10.5281/zenodo.1492479>
12. D. Yu, H. Duan, J. Fang, B. Zeng, Predominant instrument recognition based on deep neural network with auxiliary classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 852–861 (2020)
13. J. S. Gómez, J. Abeßer, E. Cano, in *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR, Paris, France September 23-27, 2018*. Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning, (2018), pp. 577–584. <https://doi.org/10.5281/zenodo.1492481>
14. X. Li, K. Wang, J. Soraghan, J. Ren, in *Proc of International Conference on Computational Intelligence in Music Sound Art and Design (Part of EvoStar)*. Fusion of hilbert huang transform and deep convolutional network for predominant musical instruments recognition vol. 12103 of Lecture Notes in Computer Science (Springer, 2020), pp. 80–89
15. K. Watcharasupat, S. Gururani, A. Lerch, Visual attention for musical instrument recognition. arXiv preprint arXiv:2006.09640 (2020)

16. A. Diment, P. Rajan, T. Heittola, T. Virtanen, in *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research, Marseille, France, October 15-18, 2013*. Modified group delay feature for musical instrument recognition (LMA, 2013), pp. 431–438. <http://www.cmmr2013.cnrs-mrs.fr/Docs/CMMR2013Proceedings.pdf>
17. F. Fuhrmann, et al., *Automatic musical instrument recognition from polyphonic music audio signals*. (PhD thesis, Universitat Pompeu Fabra, 2012)
18. H. A. Murthy, B. Yegnanarayana, Group delay functions and its applications in speech technology. *Sadhana*. **36**(5), 745–782 (2011)
19. B. Yegnanarayana, H. A. Murthy, Significance of group delay functions in spectrum estimation. *IEEE Trans. Signal Process.* **40**(9), 2281–2289 (1992)
20. K. K. Paliwal, L. D. Alsteris, On the usefulness of stft phase spectrum in human listening tests. *Speech Commun.* **45**(2), 153–170 (2005)
21. P. Grosche, M. Muller, F. Kurth, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 2010-Mar 15-19, Dallas, Texas, USA*. Cyclic tempogram—a mid-level tempo representation for music signals, (2010), pp. 5522–5525. <https://doi.org/10.1109/ICASSP.2010.5495219>
22. M. Muller, T. Pratzlich, J. Driedger, in *Proc. of 13 th International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal, October 8th-12th, 2012*. A cross-version approach for stabilizing tempo-based novelty detection, (2012), pp. 427–432
23. A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi, P. Maragos, in *Proc. of 28th European Signal Processing Conference(EUSIPCO 2020), Virtual, January 18-22, 2021*. Augmentation methods on monophonic audio for instrument classification in polyphonic music, (2021), pp. 156–160. <https://doi.org/10.23919/Eusipco47968.2020.9287745>
24. O. Slizovskaia, E. Gómez, G. Haro, in *Proc. of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR'17, June 6-9, 2017, Bucharest, Romania*. Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture, (2017), pp. 226–232. <https://doi.org/10.1145/3078971.3079002>
25. S. Oramas, F. Barbieri, O. Nieto Caballero, X. Serra, Multimodal deep learning for music genre classification. *Trans. Int. Soc. Music Inf. Retr.* **1**, 4–21 (2018). <https://doi.org/10.5334/tismir.10>
26. C. Chen, Q. Li, A multimodal music emotion classification method based on multi feature combined network classifier. *Math. Probl. Eng.* **2020** (2020). <https://doi.org/10.1155/2020/4606027>
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017) Long Beach, CA, USA*. Attention is all you need (Curran Associates, Inc, pp. 5998–6008. <http://arxiv.org/abs/1706.03762>
28. T. Zhong, S. Zhang, F. Zhou, K. Zhang, G. Trajcevski, J. Wu, Hybrid graph convolutional networks with multi-head attention for location recommendation. *World Wide Web*. **23**(6), 3125–3151 (2020)
29. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030* (2021)
30. M. Sukhavasi, S. Adapa, Music theme recognition using cnn and self-attention. *arXiv preprint arXiv:1911.07041* (2019)
31. D. Ghosal, M. H. Kolekar, in *Proc. of Interspeech, Hyderabad, India, September 2-6, 2018*. Music genre recognition using deep neural networks and transfer learning, (2018), pp. 2087–2091. <https://doi.org/0.21437/Interspeech.2018-2045>
32. W. J. Poser, *Douglas o'shaughnessy, speech communication: Human and machine*. (Addison-wesley publishing company, Reading, Massachusetts, 1987), pp. 52–54
33. R. Rajan, H. A. Murthy, Two-pitch tracking in co-channel speech using modified group delay functions. *Speech Comm.* **89**, 37–46 (2017)
34. A. V. Oppenheim, R. W. Schafer, *Discrete Time Signal Processing*. (Prentice Hall, Inc, New Jersey, 1990)
35. S. Davies, Perceiving melodies and perceiving musical colors. *Rev. Philos. Psychol.* **1**, 19–39 (2009). <https://doi.org/10.1007/s13164-009-0007-2>
36. M. Tian, G. Fazekas, D. A. Black, M. Sandler, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. On the use of the tempogram to describe audio content and its application to music structural segmentation, (2015), pp. 419–423
37. M. Muller, *Fundamentals of Music Processing Audio, Analysis, Algorithms, Applications*, vol. 5. (Springer International Publishing, Cham, 2015)
38. K. Racharla, V. Kumar, C. B. Jayant, A. Khairkar, P. Harish, in *Proc. of 7th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India*. Predominant musical instrument classification based on spectral features, (2020), pp. 617–622. <https://doi.org/10.1109/SPIN48934.2020.9071125>
39. M. D. Zeiler, R. Fergus, in *Proc. of European conference on computer vision (ECCV)*. T visualizing and understanding convolutional networks (Springer International Publishing, Switzerland, 2014), pp. 818–8331
40. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., in *Proc. of 9th International Conference on Learning Representations (ICLR)-Virtual mode from May 3-7 (2021)*. An image is worth 16x16 words: Transformers for image recognition at scale, (2021), pp. 1–21. OpenReview.net
41. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Aaron Courville, Y. Bengio, Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, 2672–2680 (2014)
42. T. Kim, M. Cha, H. Kim, J. K. Lee, J. Kim, in *Proc. of 34th International conference on machine learning, Sydney, Australia. 06–11 August 2017*. Learning to discover cross-domain relations with generative adversarial networks, vol. 70 (PMLR, 2017), pp. 1857–1865. <https://proceedings.mlr.press/v70/kim17a.html>
43. C. Donahue, J. J. McAuley, M. Puckette, in *Proc. of 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. Adversarial audio synthesis, (2019), pp. 1–16. OpenReview.net
44. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, in *Proc. of the 31st International Conference on Neural Information Processing Systems, Long Beach California USA December 4 - 9, 2017*. Improved training of wasserstein GANs (Curran Associates Inc., Morehouse Lane Red Hook NY, 2017)
45. A. Madhu, S. Kumaraswamy, in *Proc. of 27th European Signal Processing Conference (EUSIPCO), 2-6 September 2019 in A Coruña, Spain*. Data augmentation using generative adversarial network for environmental sound classification, (2019), pp. 1–5
46. G. Atkar, P. Jayaraju, Speech synthesis using generative adversarial network for improving readability of hindi words to recuperate from dyslexia. *Neural Comput. Appl.* **33**, 9353–9362 (2021). <https://doi.org/10.1007/s00521-021-05695-3>
47. L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017)
48. M. Lee, J. Lee, J.-H. Chang, Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition. *Digit. Sig. Process.* **85**, 1–9 (2019)
49. L. Nanni, G. Maguolo, S. Braham, M. Paci, An ensemble of convolutional neural networks for audio classification. *arXiv preprint arXiv:2007.07966* (2020)
50. K. Siedenburg, M. R. Schadler, D. Hulsmeier, Modeling the onset advantage in musical instrument recognition. *J. Acoust. Soc. Am.* **146**(6), 523–529 (2019)
51. A. T. Cemgil, B. Kappen, P. Desain, H. Honing, On tempo tracking: Tempogram representation and kalman filtering. *J. New Music. Res.* **29**(4), 259–273 (2000)
52. M. Ogg, L. R. Slevc, W. J. Ildsardi, The time course of sound category identification: insights from acoustic features. *J. Acoust. Soc. Am.* **142**(6), 3459–3473 (2017)
53. M. S. Mohd Azmi, M. N. Sulaiman, Accelerator-based human activity recognition using voting technique with nbtree and mlp classifiers. *Int. J. Adv. Sci. Eng. Inf. Technol.* **7**(1), 146–152 (2017)
54. S. Paul, P.-Y. Chen, Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581* (2021)
55. R. Rajan, *Estimation of Pitch in Speech and Music Using Modified Group delay Functions*. (Ph.D. thesis, Submitted to Indian Institute of Technology, Madras, 2017). http://compmusic.upf.edu/system/files/static_files/Rajan-Rajeev-PhD-thesis-2017.pdf
56. R. Rajan, H. A. Murthy, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*. Group delay-based melody monopitch extraction from music (Group delay-based melody monopitch extraction from music, 2013), pp. 186–190
57. R. Rajan, H. A. Murthy, Two-pitch tracking in co-channel speech using modified group delay functions. *Speech Commun.* (2017). 89.10.1016/j.specom.2017.02.004

58. J. Kong, J. Kim, J. Bae, in *Proc.of 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, vol. 33 (Curran Associates, Inc, 2020), pp. 17022–17033

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
