Dynamic Emotion Analysis in Piano Music Based on Performance Techniques Recognition

Yueyan Wu
School of Science and Engineering
the Chinese University of Hong Kong, Shenzhen
Shenzhen, China
e-mail: 121090620@link.cuhk.edu.cn

Clement Leung
School of Science and Engineering
the Chinese University of Hong Kong, Shenzhen
Shenzhen, China
e-mail: clementleung@cuhk.edu.cn

***Abstract*—The relationship between music and emotion has always been essential in musicology and psychology. This study aims to automatically identify the playing technique in piano performance through deep learning technology and analyze its influence on the dynamic change of emotion. We propose a technique recognition method based on a deep Convolutional Neural Network (CNN), which can accurately identify different techniques (such as octave, vibrato, glissando, etc.). In addition, we design a simple temporal analysis model to analyze the evolution of emotion over time based on the dynamic change of playing technique. The experimental results show that the identification of playing techniques achieves nearly 86% accuracy, outperforming traditional methods, and specific playing techniques are significantly related to certain emotions. There are also results on the dynamic emotion analysis task. This study not only provides a new perspective and method for the field of music emotion recognition but also provides a new tool and method for music analysis and music education.**

Keywords-Performance Techniques Recognition*; *Convolutional Neural Network (CNN)*; *Music Emotion Recognition (MER).

I. Introduction

Music, as a vital part of human culture, has long been regarded as a ’language of emotions’ [1]. Therefore, it is natural to associate music with emotions and classify it based on emotional content. Music Emotion Recognition (MER) refers to the use of computers to extract and analyze music features, establish mapping relationships between these features and emotion spaces, and recognize the emotions expressed in music [2]. In recent years, significant progress has been made in MER, especially with the development of deep learning techniques. For instance, a bimodal Deep Belief Network (DBN) model that combines audio and lyrics has shown improved accuracy in emotion recognition [3].

Additionally, Convolutional Neural Networks (CNNs) have become widely used in Music Emotion Recognition (MER) due to their ability to automatically extract music features, reducing the need for manual feature extraction [4]. Liu et al. transformed the audio signal into a spectrogram using Short-Time Fourier Transform (STFT), which was then processed through convolution, pooling, and hidden layers, followed by Softmax for emotion prediction. The innovation of the method is that the use of CNN reduces the burden of artificial feature extraction and uses convolution to capture local time and frequency patterns in the spectrogram. However, a major drawback is that it is difficult to interpret which features are most relevant to identifying the emotions in the music [5].

Despite the growing body of research on music and emotion, much of the existing work primarily focuses on lyrics, volume, and dynamics, with little attention given to how performance techniques affect the emotional expression of music. Performance techniques, such as vibrato, glissando, and arpeggio, play a crucial role in shaping the emotional content of a musical piece. For example, vibrato is often associated with expressiveness and tension, while glissando can evoke a sense of excitement or anticipation [6]. However, the relationship between specific performance techniques and emotional expression remains underexplored, particularly in the context of dynamic emotion analysis. Most studies rely on holistic emotion assessments, overlooking the temporal evolution of emotions within individual audio segments. This gap in the literature limits our understanding of how emotions fluctuate over time in response to different performance techniques.

Furthermore, existing methods for performance technique recognition face significant challenges. Traditional approaches, such as spectral analysis and cepstral analysis, can detect fundamental frequencies and harmonics but are limited by trade-offs between time and frequency resolution [7]. Moreover, harmonic relationships in Western music can cause spectral overlap, reducing the accuracy and reliability of recognition. Recent advances in deep learning, such as CNNs and Long Short-Term Memory Networks (LSTMs), have shown promise in capturing complex performance gestures by integrating performance gestures and timbral information [8]. However, these methods still face two major challenges: (1) the lack of datasets with annotated performance technique labels, and (2) the complexity and time-consuming nature of integrating non-audio factors, such as performer gestures and contextual information.

This study aims to address these gaps by proposing a deep learning model that not only automatically identifies various performance techniques in piano music but also analyzes how these techniques influence dynamic emotional changes over time. Our approach leverages a CNN to recognize performance techniques and a temporal analysis model to track the evolution of emotions within segmented audio clips. By combining these two components, we provide a novel framework for understanding the dynamic interplay between performance techniques and emotional expression in piano music.

The remainder of this paper is organized as follows: Section 2 outlines the proposed method and model architecture. Section 3 presents the experimental results and data analysis, and Sections 4 and 5 conclude the paper with a summary of findings and future directions for research.

II. Related work | Methods

A. Data Preprocessing

Data preprocessing steps have been applied to ensure the consistency, quality, and efficiency of the audio data used in our analysis.

1) Data Format Conversion

To ensure consistency and quality, all audio files were converted to WAV format, a widely supported, uncompressed format that guarantees high-quality, lossless audio representation. The following standardization steps were applied:

- **Sampling Rate:** All audio files were resampled to 44.1 kHz to balance quality and computational efficiency.
- **Bit Depth:** Audio files were stored with a 16-bit depth to preserve quality while maintaining manageable file sizes.
- **Mono Channel:** Audio was converted to mono format to eliminate potential issues from stereo channels.

This standardization ensured compatibility with the feature extraction and neural network training pipelines.

2) Audio Segmentation

To improve recognition accuracy, the audio files were divided into segments of 1.5 seconds and 3 seconds, chosen based on the characteristics of the relevant performance techniques:

- **1.5-second segments:** Used for techniques like glissando and octave, which typically occur rapidly within a short time frame.
- **3-second segments:** Used for techniques like arpeggios and vibrato, which generally involve longer durations and require more time to capture fully.

This dual-segmentation strategy accommodates the unique temporal characteristics of different performance techniques, enhancing the model's recognition capabilities.

3) Data Augmentation

To enhance the model's generalization and robustness, two data augmentation techniques were applied:

- **Time Shifting:** Each audio sample had a 50% chance of being shifted randomly along the time axis by -500 to +500 samples, simulating different starting points.
- **Adding Gaussian Noise:** Each audio sample had a 50% probability of having Gaussian noise added, with a standard deviation of 0.5% of the original signal's amplitude, simulating real-world noisy conditions.

B. Performance Techniques Recognition Model

The model aims to accurately identify piano performance techniques, such as glissando, vibrato, octave, and arpeggio, using state-of-the-art machine learning and deep learning techniques. The recognition process involves data collection, feature extraction, model training, and evaluation.

1) Data Collection

We built the dataset by collecting additional audio samples using the following methods:

- **Online Audio Collection:** We gathered audio recordings from online platforms such as YouTube and audio libraries. These recordings specifically highlight piano performance techniques, including glissando, octave, arpeggio and vibrato.
- **Self-recorded Data:** We also recorded our own piano performances, specifically designed to feature the techniques listed above.

2) Feature Extraction

Feature extraction is a critical step in our audio classification pipeline, where both static and dynamic features are extracted from raw audio signals to capture spectral and temporal information. Specific methods are applied for each playing technique—glissando, vibrato, octave, and arpeggio—based on their unique characteristics.

a) Mel-Spectrogram

To obtain a time-frequency audio signal representation, we utilize the Mel-spectrogram, computed with a sampling rate of 22,050 Hz and 128 Mel bands. The Mel-spectrogram transforms the audio signal into the Mel scale, which aligns more closely with human auditory perception.

$$\text{Mel-spectrogram}(y, \text{sr} = 22050, \text{n\_mels} = 128) \tag{1}$$

b) Decibel Conversion

We convert the power spectrogram to decibel (dB) units to enhance the dynamic range of the Mel-spectrogram, using the following transformation:

$$\text{Mel-spectrogram}_{\text{dB}} = 10 \cdot \log_{10}(\text{Mel-spectrogram} + \epsilon) \tag{2}$$

where ϵ is a small constant (e.g., 10^{-6}) to avoid taking the logarithm of zero. This conversion normalizes the amplitude variations, making the spectrogram more suitable for neural network training.

c) Glissando Feature Extraction

Glissando is a playing technique characterized by rapid and continuous pitch changes within a short time frame. To capture these dynamic changes, we extract Delta and Delta-Delta features from the Mel-spectrogram:

- **Delta Features**: Calculated as the first-order temporal derivative of the Mel-spectrogram to capture the rate of change in spectral features.

$$\Delta X(t) = X(t+1) - X(t) \tag{3}$$

- **Delta-Delta Features**: Calculated as the second-order temporal derivative of the Mel-spectrogram to capture the acceleration of changes in spectral features.

$$\Delta^2 X(t) = \Delta X(t+1) - \Delta X(t) \tag{4}$$

d) Vibrato Feature Extraction

Vibrato is a technique involving slight and continuous pitch fluctuations over a longer duration. To effectively recognize vibrato, we extract frequency modulation features based on the Mel-spectrogram:

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-239-5

- **Modulation Frequency Features**: The modulation frequency refers to the rate at which pitch fluctuates over time, while the modulation amplitude describes the extent of these fluctuations, helping to capture the distinctive characteristics of vibrato in musical performance.

$$\text{Modulation Frequency} = \frac{df}{dt} \tag{5}$$

e) Octave Feature Extraction

Octave playing involves the simultaneous occurrence of two notes separated by an octave. To capture the frequency relationships between these notes, we employ harmonic spectrum features:

- **Harmonic Analysis**: We apply harmonic decomposition methods to extract the harmonic components of the audio signal, analyzing the relationships between harmonic frequencies.

$$\text{Harmonic Components}(t) = \sum_{k=1}^{N} A_k \cdot \sin(2\pi k f_0 t) \tag{6}$$

f) Arpeggio Feature Extraction

Arpeggio involves playing the notes of a chord in sequence rather than simultaneously. The main features we extract for arpeggio detection include Delta, Delta-Delta Features and Time Interval Features.

- **Time Interval Features**: The time interval features are calculated by detecting the onset of each note in the arpeggio and computing the time intervals between successive onsets.

g) Feature Stacking and Normalization

For each playing technique, we stack the extracted features to form a multi-channel input tensor. For example, glissando features include the Mel-spectrogram, Delta, and Delta-Delta features. Figure 1 shows an example of glissando features. Similarly, for vibrato, octave, and arpeggio, we stack the respective features accordingly. All features are standardized before stacking to ensure zero mean and unit variance, which stabilizes the training process and accelerates convergence.

Feature Stacking: For example, glissando features are stacked as follows:

$$\text{Mel combined} = \text{Stack}(\text{Mel-spectrogram}_{\text{dB}}, \Delta, \Delta^2) \tag{7}$$

This results in a tensor of shape $[3, 128, 65]$, where 3 channels correspond to the Mel-spectrogram, Delta, and Delta-Delta.128 Mel bands represent the frequency dimension. 65 frames represent the temporal dimension.

For other techniques, the stacking procedure is similar, with the number of channels adjusted based on the features extracted. For instance, vibrato features may include Mel spectrograms and frequency modulation features, resulting in a 2-channel input tensor, while octave and arpeggio may use 4 channels, incorporating Mel spectrograms, harmonic features, and time-related features.

Normalization: After stacking the features, we normalize them to ensure all input features are on a similar scale:

$$\text{Mel combined} = \frac{\text{Mel combined} - \mu}{\sigma + \epsilon} \tag{8}$$

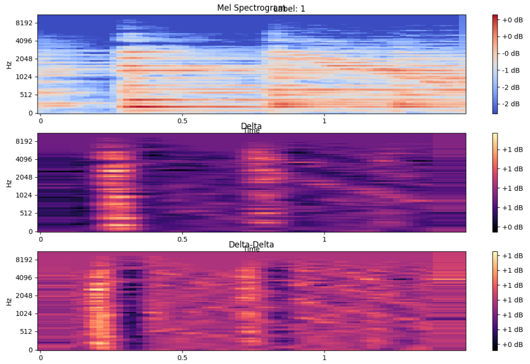

Figure 1. Example of glissando features.

where μ is the mean and σ is the standard deviation of the combined features across the dataset, and ϵ is a small constant (e.g., 10^{-6}) to avoid division by zero.

3) Model Architecture and Loss Function

We designed four Convolutional Neural Networks (CNN) to recognize different piano performance techniques, including glissando, vibrato, arpeggio, and octave. The architecture of the model consists of several sequential layers. For specific details of Vibrato detection, refer to Table I.

TABLE I
CNN ARCHITECTURE FOR BINARY CLASSIFICATION (GLISSANDO DETECTION)

Layers	**Operator**	**Input Size**	**Output Size**
Conv1	Conv2D 3×3	$3 \times 128 \times 65$	$32 \times 128 \times 65$
MaxPool	MaxPool 2×2	$32 \times 128 \times 65$	$32 \times 64 \times 32$
Conv2	Conv2D 3×3	$32 \times 64 \times 32$	$64 \times 64 \times 32$
MaxPool	MaxPool 2×2	$64 \times 64 \times 32$	$64 \times 32 \times 16$
Conv3	Conv2D 3×3	$64 \times 32 \times 16$	$128 \times 32 \times 16$
MaxPool	MaxPool 2×2	$128 \times 32 \times 16$	$128 \times 16 \times 8$
Conv4	Conv2D 3×3	$128 \times 16 \times 8$	$256 \times 16 \times 8$
MaxPool	MaxPool 2×2	$256 \times 16 \times 8$	$256 \times 8 \times 4$
AAP	AdaptiveAvgPool	$256 \times 8 \times 4$	$256 \times 1 \times 1$
Flatten	Flatten	$256 \times 1 \times 1$	256
FCL1	Fully Connected	256	128
ReLU, Dropout	Dropout	128	128
FC1	Fully Connected	128	1

a) Convolutional Layers

The model utilizes a series of convolutional layers that applies filters to the input feature maps. Each convolutional layer is followed by a Batch Normalization layer and a Rectified Linear Unit (ReLU) activation function to improve convergence and introduce non-linearity. The operation for a convolutional layer can be described as:

$$\mathbf{H}_i = \text{ReLU}\left(\text{BatchNorm}(\text{Conv2D}(\mathbf{X}_{i-1}, W_i, b_i))\right) \tag{9}$$

where $\mathbf{X}_{i-1}$ is the output of the previous layer, W_i and b_i are the weights and bias of the i-th convolutional layer, and $\mathbf{H}_i$ is the output of the convolutional layer.

b) Pooling Layers

After each convolutional block, max pooling is applied to reduce the spatial dimensions of the feature maps while preserving the most relevant features. The max pooling operation can be described as:

$$\mathbf{H}_i^{\text{pool}} = \text{MaxPooling}(\mathbf{H}_i) \tag{10}$$

where $\mathbf{H}_i$ is the feature map after convolution, and $\mathbf{H}_i^{\text{pool}}$ is the output of the pooling layer.

c) Adaptive Pooling

An adaptive average pooling layer is applied at the end of the convolutional layers to reduce the feature map to a fixed size, regardless of the input dimensions. The adaptive pooling operation is:

$$\mathbf{H}_{\text{final}} = \text{AdaptiveAvgPool2d}(\mathbf{H}_{\text{pool}}) \tag{11}$$

where $\mathbf{H}_{\text{pool}}$ is the pooled feature map, and $\mathbf{H}_{\text{final}}$ is the fixed-size output feature map.

d) Fully Connected Layers

After the feature maps are extracted, they are flattened into a one-dimensional vector and passed through fully connected layers. The output of the fully connected layer can be written as:

$$\mathbf{z}_1 = \text{ReLU}(W_1 \cdot \mathbf{H}_{\text{final}} + b_1) \tag{12}$$

where W_1 and b_1 are the weights and bias of the first fully connected layer, and $\mathbf{z}_1$ is the output of this layer. The second fully connected layer produces the final output:

$$\mathbf{z}_2 = W_2 \cdot \mathbf{z}_1 + b_2 \tag{13}$$

and the final classification output is obtained using a sigmoid activation:

$$\mathbf{y}_{\text{pred}} = \text{Sigmoid}(\mathbf{z}_2) \tag{14}$$

e) Output Layer

The final output of the model is a probability value between 0 and 1, indicating whether a specific performance technique (such as vibrato, glissando, etc.) is present in the audio segment.

f) Loss Function

The model is trained using the binary cross-entropy loss, which is appropriate for the binary classification task of detecting the presence or absence of a musical technique. The binary cross-entropy loss can be defined as:

$$\mathcal{L} = -\left(y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right) \tag{15}$$

where y is the ground truth label (0 or 1), and $\hat{y}$ is the predicted probability of the model. This loss function is minimized during training using optimization algorithms like Adam [9].

C. Correlation Analysis between Performance Techniques and Emotions

In this section, we describe the methodology used to analyze the relationship between various piano performance techniques and the emotional expressions conveyed through the audio data. The goal of this analysis is to understand how different performance techniques, such as glissando, tremolo, arpeggio, and octave, influence emotional expression, based on musical features such as pitch, rhythm, and dynamics.

1) Emotion Labeling using GEMS (Geneva Emotional Music Scales)

For emotion labeling, we employed the Geneva Emotional Music Scales (GEMS), a comprehensive model specifically designed for music-induced emotion. GEMS includes 45 emotional tags, which are divided into nine distinct categories [8]:

- **Amazement**, **Solemnity**, **Tenderness**, **Nostalgia**, **Calmness**, **Power**, **Joyful Activation**, **Tension**, and **Sadness**.

Emotion labels for the performance techniques were manually annotated by professional musicians with expertise in emotional interpretation in music. These musicians listened to the performances techniques and assigned appropriate emotion labels based on their auditory perception of the emotional content.

2) Pearson Correlation Analysis

To explore the relationship between performance techniques and emotional expression, we performed a Pearson correlation analysis. Pearson's correlation coefficient (r) quantifies the linear relationship between two variables, ranging from -1 to $+1$, where $+1$ indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship.

We calculated the Pearson correlation coefficient between the following variables:

- **Performance Techniques:** Glissando, tremolo, arpeggio, and octave.
- **Emotions:** Amazement, Solemnity, Tenderness, Nostalgia, Calmness, Power, Joyful Activation, Tension, and Sadness.

The Pearson correlation coefficient for each pair indicates the strength and direction of the relationship between each performance technique and the corresponding emotional expression. A positive correlation suggests that the performance technique is associated with the emotion, while a negative correlation suggests the opposite.

D. Dynamic Emotion Analysis

Dynamic emotion analysis aims to capture the temporal evolution of emotional expression in piano performances. Given the segmented audio clips, each representing a 3-second segment, we analyze the emotional changes as a function of the performance techniques detected in the audio.

To quantify emotional progression, each audio clip was assigned an emotion vector based on a specific performance technique, and the emotion vector of each clip was tracked throughout the performance. Finally we visualized the temporal progression of emotions to show how emotional intensity evolves throughout the performance. This analysis helps to identify the emotional peaks and transitions generated by specific techniques and how they relate to the performance dynamics.

1) Emotion Weighting Based on Performance Techniques

To enhance the precision of emotion analysis, the emotional contribution of each performance technique is weighted according to its decibel level. The decibel level of each technique

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-239-5

reflects its relative prominence in the audio, thus affecting the emotional expression of the segment. The weight w_i for each technique is computed using the following formula:

$$w_i = \frac{d_i}{\sum_{i=1}^{n} d_i}, \quad d_i = 20\log_{10}\left(\frac{P_i}{P_{\text{ref}}}\right) \tag{16}$$

where d_i is the decibel value associated with the technique i, and $\sum_{i=1}^{n} d_i$ is the total sum of decibel values for all techniques in the segment. The weighted emotional vector $\mathbf{E}_{\text{final}}$ for each segment is then computed by:

$$\mathbf{E}_{\text{final}} = \sum_{i=1}^{n} w_i \cdot \mathbf{E}_i \tag{17}$$

where $\mathbf{E}_i$ is the emotional vector associated with technique i, and w_i is the weight determined by its decibel level. This ensures that techniques with higher decibel values contribute more to the final emotional expression of the segment.

III. Results

A. Performance Metrics

Table II summarizes the classification performance of our `AudioClassifier` model.

TABLE II
Performance Metrics for Piano Performance Techniques

Technique	Accuracy	Precision	Recall	F1-score
Glissando	89.5%	88.3%	87.6%	89.9%
Octave	86.2%	88.1%	84.7%	86.4%
Arpeggio	83.0%	82.9%	84.3%	83.1%
Vibrato	85.8%	83.7%	88.2%	85.9%

Glissando performed best in accuracy, accuracy, recall and F1 score, especially in the F1 score of 89.9%. Octave accuracy is the highest at 88.1%, but the overall F1 score is slightly lower than that of the glissando. Arpeggios performed the worst among the indicators, with the lowest accuracy of 83.0%. The vibrato performed better in recall and F1 scores, but still fell short of the glissando and octaves.

B. Pearson Correlation Analysis Results

The Pearson correlation coefficients between the performance techniques and emotions are summarized in Table III. Glissando has a strong positive correlation with pleasure activation, surprise and power. Vibrato are highly associated with nostalgia and tenderness, and are positively associated with sadness. Arpeggios were positively correlated with nostalgia and tenderness, but negatively correlated with tension and sadness. The octave shows a strong sense of power and pleasure activation, and is negatively associated with tenderness and sadness.

C. Dynamic Emotion Analysis Results

In this section, we present the results of the dynamic emotion analysis applied to the performance of Czerny Op. 365 No. 33, a Polish dance. For the purpose of this analysis, the 1-minute audio was segmented into 20 equal parts, each lasting 3 seconds. These segments were analyzed for the presence of specific performance techniques and their corresponding emotional expressions. The emotional vectors for each segment were determined based on the techniques detected. Specifically:

- The *octave* technique, present in the majority of the segments, was predominantly associated with the emotion of *joyful*.
- The *vibrato* technique, observed in segments 4, 5, 13, 15, 16, 17, 19, and 20, was associated with emotional expressions such as *amazement*, *tension*, and *sadness*.
- The *glissando* technique, detected in segments 7 and 10, elicited emotions like *joyful* and *activation*.

The results shows in Figure 2.

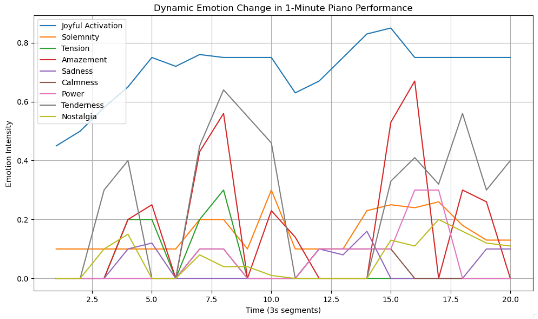

Figure 2. Dynamic emotion Cchange in 1-minute piano performance.

Peaks in emotional intensity were found to correlate with specific techniques, highlighting how the performer's use of these techniques influenced the emotional flow of the piece. The emotional transitions between segments revealed that the piece, while maintaining an overall joyful tone due to the dominance of *octave*, also incorporated dramatic shifts, reflecting the emotional depth of the work.

IV. Discussion | Evaluation

The CNN-based model achieves high accuracy in classifying piano performance techniques, with training accuracy reaching 96% and validation accuracy stabilizing at 86% (Figure 3), indicating robust generalization without overfitting. However, two limitations persist:

- **Overlapping Spectral Features:** Techniques with similar harmonic patterns, such as arpeggios and trills, are occasionally misclassified. For instance, trills involve rapid note alternations that may overlap with arpeggio harmonics in the Mel-spectrogram.
- **Independent Technique Detection:** The current framework processes each technique independently, leading to redundant computations. A unified multi-label classification approach could better capture inter-technique dependencies (e.g., vibrato often co-occurs with legato phrasing).

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-239-5

TABLE III
PEARSON CORRELATION COEFFICIENTS BETWEEN PERFORMANCE TECHNIQUES AND EMOTIONS.

Performance Technique	Joyful Activation	Calmness	Tension	Amazement	Sadness	Solemnity	Power	Tenderness	Nostalgia
Glissando	0.65	0.21	0.71	0.55	-0.31	-0.47	0.62	-0.20	-0.60
Vibrato	0.30	-0.25	0.65	0.53	0.65	-0.40	-0.35	0.78	0.80
Arpeggio	0.62	-0.10	-0.26	-0.55	-0.32	-0.30	0.13	0.73	0.82
Octave	0.75	-0.45	0.60	0.60	-0.50	0.54	0.90	-0.80	-0.56

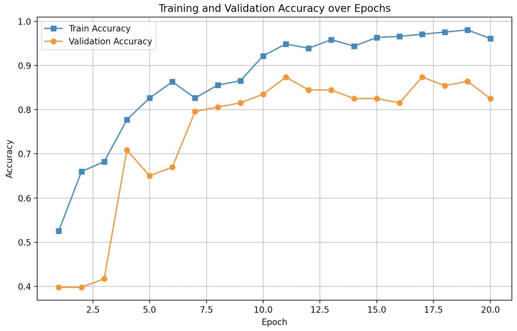

Figure 3. Training and validation accuracy over epochs.

Our analysis of the association between playing skills and emotion reveals some interesting findings, suggesting that different playing skills are significantly associated with specific emotions. This analysis provides a valuable perspective for further understanding of emotional expression in piano performance. However, the perception of musical emotion is highly subjective. Even though we invited professional musicians to conduct data annotation, there is still some disagreement. Different listeners or players may have different emotional responses to the same playing technique. It is worth noting that while there is a correlation between playing technique and emotion, the same technique may trigger different emotions in different musical contexts. For example, a glissando technique may elicit anger in a fast-paced part, while a slow-paced part may convey anticipation. Identifying emotions accurately is still tricky.

Dynamic emotion analysis, which combines technical recognition with emotion time series tracking, provides a valuable perspective on the evolution of emotion over time in piano performance. By tracking emotional changes in the temporal dimension, we could observe fluctuations in emotional intensity and identify the impact of playing techniques on emotional dynamics. However, when performing sentiment analysis, we combined the decibel level of each technique for weighted analysis. While this provides some basis for quantifying emotional intensity, there are still some problems. First of all, simply weighting by decibel intensity may oversimplify the expression of emotion because changes in emotion are not only affected by volume but also related to pitch, rhythm, performance expression, and other factors. Second, decibel levels can have different effects on players and sound equipment, leading to sentiment analysis bias. Therefore, future research needs to explore a more integrated approach to sentiment analysis that may include more audio features.

V. CONCLUSION AND FUTURE WORK

In this study, we propose a deep learning approach for dynamic emotion analysis of piano music by combining piano performance technique recognition with emotion time-series tracking. Our CNN-based model effectively identifies various performance techniques and achieves high classification accuracy. We found that different techniques are strongly associated with specific emotional expressions, though emotional perception remains subjective and context-dependent. Despite the model's strong performance, challenges remain, such as distinguishing overlapping techniques and simplifying sentiment analysis based on decibel levels. These results demonstrate the potential of this approach but also highlight areas for further improvement.

Future research could focus on integrating multiple performance techniques into a single model and expanding the range of performance techniques recognized. Additionally, incorporating more audio features, such as tone, timbre, and rhythm, could provide a more comprehensive understanding of emotional expression. Real-time emotion tracking during performance could also open up new applications in music education and interactive environments. Lastly, developing larger and more diverse annotated datasets would enhance model generalization and improve recognition accuracy.

REFERENCES

[1] C. C. Pratt, "Music as the Language of Emotion," *The Library of Congress*, December 1950.

[2] X. Y. Yang, Y. Z. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365–389, 2018.

[3] M. Y. Huang, W. G. Rong, T. Arjannikov, J. Nan, and Z. Xiong, "Bi-modal deep Boltzmann machine-based musical emotion classification," in *Proceedings of the 25th International Conference on Artificial Neural Networks*, pp. 199–207, 2016.

[4] P. T. Yang, S. M. Kuang, C. C. Wu, and J. L. Hsu, "Predicting music emotion by using convolutional neural networks," in *Proceedings of the 22nd HCI International Conference*, pp. 266–275, 2020.

[5] X. Liu, Q. C. Chen, X. P. Wu, Y. Liu, and Y. Liu, "CNN-based music emotion classification," *arXiv Preprint* arXiv:1704.5665, 2017.

[6] M. Blaszke and B. Kostek, "Musical instrument identification using deep learning approach," *Sensors (Basel, Switzerland)*, vol. 22, no. 8, pp. 3033, 2022.

[7] V. Lostanlen, J. Anden, and M. Lagrange, "Extended playing techniques: the next milestone in musical instrument recognition," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, pp. 1–10, September 2018.

[8] M. Chekowska-Zacharewicz and M. Janowski, "Polish adaptation of the Geneva Emotional Music Scale (GEMS): Factor structure and reliability," *Psychology of Music*, vol. 57, no. 6, pp. 427–438, 2020.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-239-5