BRAININFO 2025

The Tenth International Conference on Neuroscience and Cognitive Brain Information

ISBN: 978-1-68558-239-5

March 9th –13th, 2025

Lisbon, Portugal

BRAININFO 2025 Editors

Petre Dini, IARIA, USA/EU

BRAININFO 2025

Forward

The Tenth International Conference on Neuroscience and Cognitive Brain Information (BRAININFO 2025), held between March 9th, 2025, and March 13th, 2025, in Lisbon, Portugal, continued a series of international events that evaluate current achievements and identify potential ways of making use of the acquired knowledge in neuroscience, brain connectivity, brain intelligence paradigms, cognitive information, and specific applications.

The complexity of the human brain and its cognitive actions prompted extensive research for decades. Most of the findings were adapted in virtual/artificial systems with the idea of modeling them and used in a brain-like manner for human-centered medical cures, especially for neurotechnology. Information representation, retrieval, and internal data connections still constitute a domain where solutions are either missing or in a very early stage.

We take here the opportunity to warmly thank all the members of the BRAININFO 2025 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to BRAININFO 2025. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the BRAININFO 2025 organizing committee for their help in handling the logistics of this event.

We hope that BRAININFO 2025 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the areas of neuroscience and cognitive brain information.

BRAININFO 2025 Chairs

BRAININFO 2025 Steering Committee

Erwin Lemche, Institute of Psychiatry, Psychology & Neuroscience | King's College School of Medicine and Dentistry, UK
Yu-Dong (Eugene) Zhang, University of Leicester, UK
Chih Lai, University of St. Thomas Minnesota, USA
William Speier, UCLA David Geffen School of Medicine, USA
Sule Yildirim-Yayilgan, Norwegian University of Science and Technology, Norway
Ricardo Ron-Angevin, University of Málaga, Spain
Rory Lewis, University of Colorado at Colorado Springs, USA

BRAININFO 2025 Publicity Chairs

Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
José Miguel Jiménez, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politècnica de València, Spain

BRAININFO 2025
Committee

BRAININFO 2025 Steering Committee

Erwin Lemche, Institute of Psychiatry, Psychology & Neuroscience | King's College School of Medicine and Dentistry, UK
Yu-Dong (Eugene) Zhang, University of Leicester, UK
Chih Lai, University of St. Thomas Minnesota, USA
William Speier, UCLA David Geffen School of Medicine, USA
Sule Yildirim-Yayilgan, Norwegian University of Science and Technology, Norway
Ricardo Ron-Angevin, University of Málaga, Spain
Rory Lewis, University of Colorado at Colorado Springs, USA

BRAININFO 2025 Publicity Chairs

Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
José Miguel Jiménez, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politècnica de València, Spain

BRAININFO 2025 Technical Program Committee

Aftab Ahmad, John Jay, The City University of New York (CUNY), USA
Mariano Alcañiz Raya, Institute of Research and Innovation in Bioengineering (I3B) - Universidad Politécnica Valencia, Spain
Saad Alqithami, Albaha University, Saudi Arabia
Panagiotis Bamidis, Aristotle University of Thessaloniki, Greece
Nadejda Bocheva, Institute of Neurobiology | Bulgarian Academy of Sciences, Bulgaria
Gian Carlo Cardarilli, University of Rome Tor Vergata, Italy
German Castellanos-Dominguez, Universidad Nacional de Colombia, Sede Manizales, Colombia
Manoj Chandrasekaran, Meta Platforms Inc., USA
Ahana Roy Choudhury, Florida State University, USA
Gopikrishna Deshpande, Auburn University, USA
Shiaofen Fang, Indiana University Purdue University Indianapolis, USA
Mai Gamal, German University in Cairo, Egypt
Faizal Hajamohideen, University of Technology and Applied Sciences, Oman
Monte Hancock, 4Digital Inc., USA
Cosimo Ieracitano, University Mediterranea of Reggio Calabria, Italy
Kavikumar Jacob, Universiti Tun Hussein Onn Malaysia, Malaysia
M. Shamim Kaiser, Jahangirnagar University, Bangladesh
Stamatis Karlos, University of Patras, Greece
Peter Kieseberg, St. Poelten University of Applied Sciences, Austria
Hongzhi Kuai, Maebashi Institute of Technology, Japan
Chih Lai, University of St. Thomas Minnesota, USA
Erwin Lemche, Institute of Psychiatry, Psychology & Neuroscience | King's College School of Medicine / Dentistry Department of Psychosis Studies | Section of Cognitive Neuropsychiatry, UK

Copyright Information

Table of Contents

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:

- **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:

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

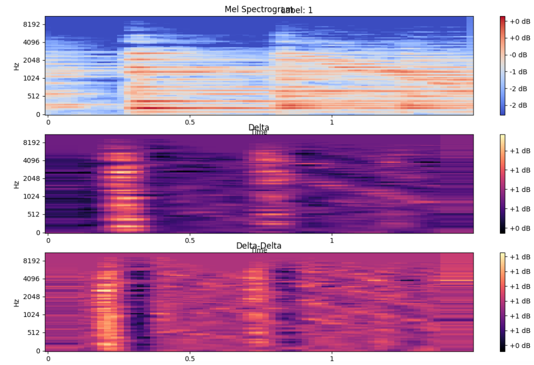

Figure 1. Example of glissando features.

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 pre-

 ISBN: 978-1-68558-239-5

serving 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

 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.

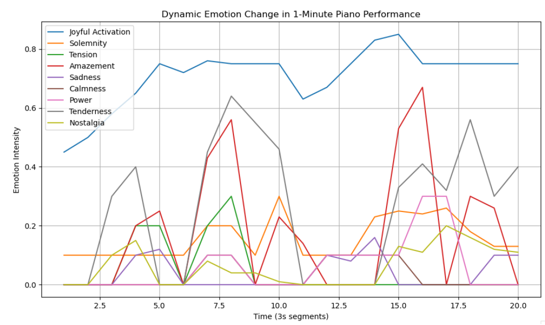

Figure 2. Dynamic emotion Cchange in 1-minute piano performance.

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).

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

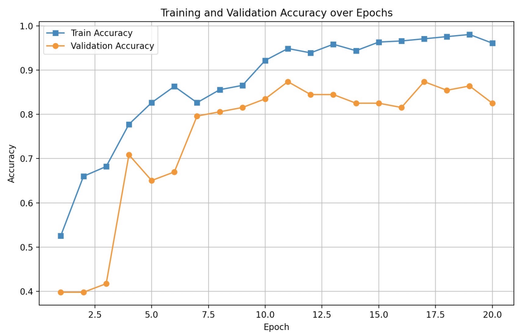

Figure 3. Training and validation accuracy over epochs.

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.

Predicting Emotion States Using Markov Chains

Clement Leung
School of Science and Engineering and
Guangdong Provincial Key Laboratory of
Future Networks of Intelligence
Chinese University of Hong Kong
Shenzhen, China
Email: clementleung@cuhk.edu.cn

Zhifei Xu
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
Email: zhifeixu1@link.cuhk.edu.cn

***Abstract*—In a wide range of tasks, especially those involving critical safety considerations, it is crucial that human participants maintain appropriate emotional conditions. As a result, accurate recognition of these emotional states has become a central research challenge, with mainstream methods frequently utilizing Pre-trained Language Models (PLMs) to incorporate emotional understanding. With the emergence of Large Language Models (LLMs) like ChatGPT, we have seen remarkable advancements in various natural language processing applications. However, the potential of ChatGPT's zero-shot capabilities for image-based emotion recognition and analysis has not been thoroughly explored. In this study, we focus on classifying and predicting emotional states, specifically distinguishing between positive and negative emotions, and we examine ChatGPT4's ability to interpret emotions directly from images. Our experiments show that ChatGPT4 can effectively predict changes in emotional states over time, surpassing expectations in identifying the progression of positive and negative emotions. Nonetheless, we identified shortcomings in its capacity to accurately recognize specific negative emotions, indicating room for further improvement.**

Keywords-Image Emotion Prediction; Large Language Model; ChatGPT4; zero-shot; Markov Chain; Emotion Stability Parameter.

I. Introduction

In human communication, accurately representing and interpreting emotions is crucial. Emotions foster meaningful connections and reveal an individual's mental state and intentions. Over the past decade, extensive research has focused on integrating emotional insight into human-computer dialogue systems [1]. Concurrently, the advent of ChatGPT [2] and Instruct-GPT [3] has sparked interest in their capacity for precise emotion recognition. Emotional support is increasingly essential in scenarios like personal conversations, mental health assistance, and customer interactions. Accordingly, our study investigates how effectively ChatGPT4 [4] can discern emotions from facial expressions.

Emotion recognition and prediction have gained prominence for promoting safety, supporting mental well-being, and enhancing user experiences. Recognized as a key factor in human safety, emotion recognition has been extensively researched [5] [6]. People naturally communicate emotions through words, text, images, facial cues, and physical gestures.

In a tech-driven world, Artificial Intelligence's (AI) ability to understand and respond to human emotions is indispensable [7]. The significance of emotionally sensitive AI is magnified by societal pressures such as occupational stress, perceived injustices, and the strain of personal breakups [8] [9], which can push individuals to harmful extremes. Evidence of such distress includes suicidal ideation linked to professional demands [8], school shootings, and road rage incidents. High-stakes roles—like surgeons, pilots, and truck drivers—require emotional stability, as demonstrated by a pilot with depression who attempted to shut down an airplane's engines mid-flight [9]. These scenarios underscore the urgent need for advancements in emotion recognition and prediction to bolster safety and mental health [10]. In recent years, generating emotionally responsive outputs through neural networks has become a prominent research focus [11], driven by advancements in online social networks and deep learning technologies. Moreover, the continuous evolution of large language models has triggered a revolution in conversational AI, as exemplified by ChatGPT4. These models exhibit robust, general-purpose linguistic capabilities, offering unprecedented levels of semantic comprehension and nuanced response generation. Consequently, the quality of human-computer interaction has significantly improved. Yet, the extent to which these systems exhibit emotions within their dialogues remains largely unexplored. Our goal involves developing effective conversational strategies in ChatGPT4 and assessing recent progress, strengths, and limitations [12] [13] [14] in the realm of multi-modal emotion recognition and prediction tasks. Employing ChatGPT4 for emotion detection is also considered beneficial for maintaining fairness in experimental settings, as it can interpret data without the biases often seen in human evaluators. Emphasizing this approach not only fosters fairness but also prioritizes safety and well-being. By pinpointing emotional states accurately, interventions can be better tailored and more ethically executed, thereby safeguarding participants.

Furthermore, foundational research on emotion recognition and prediction dates back to Ekman's widely recognized classification model [15], which identified six universal emotions: joy, sadness, fear, anger, surprise, and disgust. Building upon Ekman's work, Plutchik proposed an arrangement of eight primary emotions—joy, trust, fear, surprise, sadness, disgust, anger, and anticipation—in a wheel-shaped model [16]. These approaches represent categorical or discrete models, positing a fixed set of universally understood basic emotions. In contrast, continuum models treat emotions as existing along dynamic dimensions [17] [18], factoring in valence (ranging from positive to negative), arousal (level of excitement or calmness), and dominance (sense of influence or control).

Traditionally, emotion recognition and prediction research focused on a single channel of expression. However, people naturally communicate emotions through multiple modalities: voice, text, images, facial expressions, and body movements, making it challenging to accurately interpret emotions from just one source. To address this complexity, multimodal sentiment analysis integrates various data inputs, such as audio signals, shape changes, and overall appearance [19], often combined with text and images. Employing advanced techniques such as Convolutional Neural Networks (CNNs) [20] [21] or transformers enables more accurate and comprehensive emotion recognition and classification. Since emotions frequently evolve over time, predicting their progression is equally important in understanding real-world emotional dynamics.

Section III shows the results and analysis of the experiment, and Section IV discusses the experimental results. The conclusion and future work are presented in Section V.

II. Materials and Methods

In this paper, we propose an emotion prediction model grounded in Markov chains and emotion stability parameters. By constructing an emotion state transition matrix and incorporating stability parameters, the model integrates eight basic emotional states and forecasts how emotions evolve. To verify its effectiveness, we conducted long-term emotion predictions and thoroughly traced how these emotional states change as time passes. Our experimental results show that this model can effectively capture dynamic emotional fluctuations, providing a novel approach to emotion analysis and prediction.

Emotions play a pivotal role in everyday life and human-computer interactions. Accurately predicting and analyzing shifting emotional states is of great importance in fields like psychology, artificial intelligence, and human-computer interaction [15] [16]. However, many existing methods lack a dynamic perspective and struggle to anticipate how emotions might evolve as time moves forward. To address this gap, we present an emotion prediction model based on Markov chains and emotion stability parameters, aiming to precisely predict long-term changes in emotional states.

In reality, human emotions are continuous and frequently shift from one state to another [22] [23]. Depending on the context, it may be necessary to foresee the emotional states of specific individuals, such as when scheduling surgeries in hospitals, managing pilots during flights, or assigning tasks in high-risk industries. As noted earlier, emotional responses can surface when individuals are fatigued or treated unfairly. In safety-critical jobs, we want the people involved to maintain stable emotional states [9], since those experiencing emotional difficulties can compromise the safety of others.

Within this study, we focus on a classification model guided by Ekman's framework, using six primary emotions: happiness, sadness, fear, anger, surprise, and disgust. These emotions are categorized into positive and negative groups. Happiness and surprise are considered positive (+1), while sadness, fear, anger, and disgust are treated as negative (-1). Although certain high-risk scenarios might warrant a stricter classification—possibly moving surprise into the negative category—this paper retains surprise as a positive emotion.

Our model, denoted as $S(t)$, represents how an individual's emotional state changes over time, with t indicating the temporal dimension. We assign $S(t) = 1$ for positive emotions and $S(t) = -1$ for negative emotions. Human emotional complexity arises from external factors beyond personal control, such as financial stability, relationships, health, workplace conditions, market fluctuations, and family issues. These elements can trigger transitions from positive to negative emotional states or vice versa.

We begin by setting $S(0) = 1$. We then model the moments of emotional shifts (from +1 to -1 or the reverse) using a Poisson Process. Accordingly, $S(t) = 1$ if the number of transitions in the interval $(0, t)$ is even, and $S(t) = -1$ if it is odd. This approach captures the stochastic nature of emotional shifts and lays the groundwork for predicting long-term emotional evolution.

$$P[S(t) = 1|S(t) = 1] = p_0 + p_2 + p_4 + ... + ..., \quad (1)$$

where p_k is the number of Poisson points in $(0, t)$ with parameter λ. That is,

$$\begin{aligned} P[S(t) = 1|S(0) = 1] &= e^{-\lambda t}[1 + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^4}{4!} ... + ...] \\ &= e^{-\lambda t} \cosh \lambda t \end{aligned} \quad (2)$$

Now, $S(t) = -1$ if the number of points in the time interval $(0, t)$ is odd; that is,

$$\begin{aligned} P[S(t) = -1|S(0) = 1]] &= e^{-\lambda t}[1 + \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^5}{5!} ... + ...] \\ &= e^{-\lambda t} \sinh \lambda t \end{aligned} \quad (3)$$

Equation (2) represents the probability that the emotion is still positive at time t given that it was positive at time 0. Equation (3) gives the probability that the emotion is negative at time t given that it was positive at time 0. The parameter λ in both expressions represents a rate at which emotions change or decay over time. A larger value of λ would mean emotions change more rapidly, while a smaller value would mean they change more slowly. This is where we mathematically analyze possible emotional changes and predict them. Also, to verify the idea, we use ChatGPT. As for the experimental evaluation part, the Receiver Operating Characteristic (ROC) method was adopted to analyze the experimental results, and a specific explanation was placed in the experimental part.

Here, we briefly explain equation (3). First, we assume that lambda is 0.3, 0.6, 0.9. Meanwhile, Figure 1 shows the function $e^{-\lambda t} \sinh(\lambda t)$ with λ values of 0.3, 0.6 and 0.9. Displaying the function $e^{-\lambda t} \sinh(\lambda t)$ with λ values of 0.3, 0.6

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

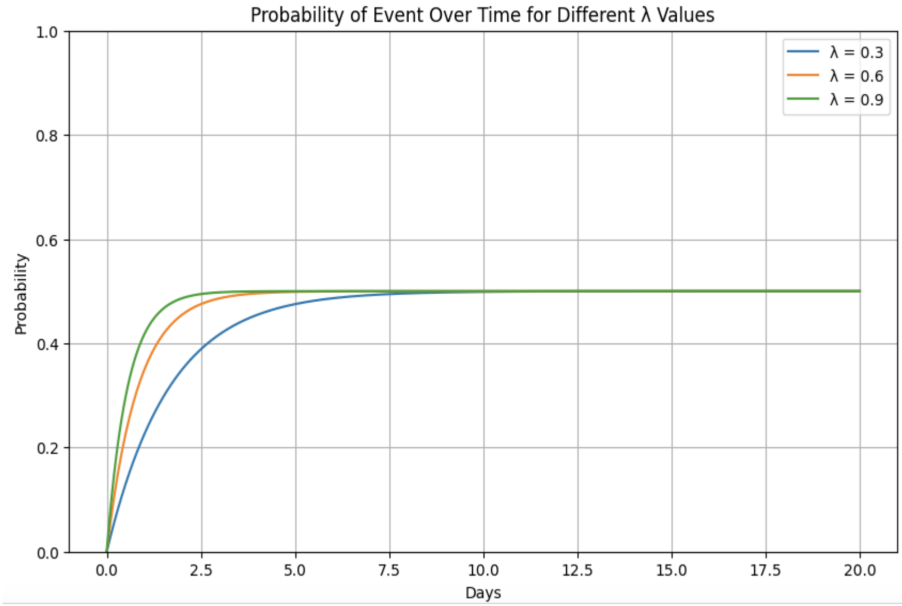

Figure 1. Different λ of the equation 3.

and 0.9, provides insightful observations about the temporal changes in probabilities, especially relevant to emotional states or comparable processes that can either diminish or progress over time. It is evident from the visualization that varying the decay constants (λ) significantly affects how long and intensely certain states persist across several days.

With a decay constant of $\lambda = 0.3$, the probability initially is high but diminishes gradually, indicating a persistent or slowly fading condition. For example, maintaining a negative emotional state translates to a higher likelihood of remaining in this state longer. Within the initial day, the probability stays well above 60%, denoting a strong endurance of the state. By day three, it hovers around 50%, showing a steady, yet noticeable decline. This gradual reduction might symbolize scenarios where the causes behind the emotional state are slow to be addressed or alleviated.

Increasing the decay constant to $\lambda = 0.6$ accelerates the probability's decline. This faster fall suggests a quicker fading of the emotional intensity or the likelihood of sustaining the same state. On the first day, the probability remains elevated but swiftly falls below 60%, nearing 50% by the close of the second day. This quicker reduction may be associated with effective interventions or environmental changes, or possibly better coping strategies that shorten the duration of the negative condition.

At $\lambda = 0.9$, the probability decreases even more swiftly. The graph shows a sharp descent, indicative of scenarios where negative emotional states or similar conditions dissipate very quickly. The probability does not stay above 50% beyond two days, dropping near this mark by the end of the first day. This rapid decline could point to highly effective external support or events that inherently do not have prolonged effects.

By analyzing these curves, one can determine how various strategies or intrinsic elements affect the control or maintenance of specific states—be they emotional, physical, or of another nature. The differing λ values symbolically illustrate the varied speeds at which environments, individuals, or systems either normalize or transition from one state to another. This knowledge is essential in areas such as psychology, where predicting the duration of an individual's negative emotional state is key to developing timely and effective interventions. Insights into these temporal patterns are invaluable for customizing interventions or supports that are sensitive to timing and more closely correspond to the observed rates of change.

This section elaborates on the construction of the emotion prediction model, including the definition of emotion states, the construction of the state transition matrix, the introduction of emotion stability parameters, the calculation of emotion distributions over time steps, and the computation and ranking of emotion change probabilities. We will demonstrate the detailed derivation process of emotion changes over longer time steps $t = 0$ to $t = 5$. The emotion state vector $S(t)$ represents the probability distribution of emotions at time t:

$$S(t) = [P_{E_1}(t), P_{E_2}(t), \ldots, P_{E_8}(t)]^T \tag{4}$$

where $P_{E_i}(t)$ denotes the probability of emotion E_i at time t.

The transition of emotion states is modeled using a Markov chain. The state transition matrix P represents the probability of transitioning from one emotional state to another. The matrix P is an 8×8 probability matrix, where each row sums to 1.

We assume the state transition matrix P to be:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} & p_{18} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} & p_{27} & p_{28} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{81} & p_{82} & p_{83} & p_{84} & p_{85} & p_{86} & p_{87} & p_{88} \end{bmatrix}$$

where p_{ij} represents the probability of transitioning from emotion E_i to emotion E_j, satisfying:

$$\sum_{j=1}^{8} p_{ij} = 1, \quad \forall i \in \{1, 2, \ldots, 8\} \tag{5}$$

In other words, we set the following transition probabilities (the values are illustrative and can be adjusted based on actual conditions):

- Higher probabilities of transition among positive emotions and lower probabilities of transitioning to negative emotions.
- Higher probabilities of transition among negative emotions and lower probabilities of transitioning to positive emotions.

For example, when the emotion is Joy (E_1), the transition probabilities are:

$$p_{1j} = \begin{cases} 0.5, & \text{if } j = 1 \text{ (remain in Joy)} \\ 0.15, & \text{if } j = 2 \text{ (transition to Trust)} \\ 0.15, & \text{if } j = 3 \text{ (transition to Surprise)} \\ 0.1, & \text{if } j = 4 \text{ (transition to Anticipation)} \\ 0.05, & \text{if } j = 5 \text{ (transition to Sadness)} \\ 0.02, & \text{if } j = 6 \text{ (transition to Disgust)} \\ 0.02, & \text{if } j = 7 \text{ (transition to Anger)} \\ 0.01, & \text{if } j = 8 \text{ (transition to Fear)} \end{cases}$$

Transition probabilities for other emotions can be similarly defined, ensuring each row sums to 1. The emotion stability parameter λ_i is used to simulate the volatility of emotions in reality:

- Positive emotions have smaller λ_i, indicating they are more stable.
- Negative emotions have larger λ_i, indicating they are less stable.

We set:

$$\lambda_i = \begin{cases} 0.2, & \text{if } E_i \text{ is a positive emotion} \\ 0.5, & \text{if } E_i \text{ is a negative emotion} \end{cases}$$

Calculation of Emotion Distributions over Time Steps. We consider the Initial Emotion State to be S(0) and we set the initial emotion as Joy (E1):

$$S(0) = [1, 0, 0, 0, 0, 0, 0, 0]^T \tag{6}$$

State Transition Computation

At each time step t, the emotion state is updated using the state transition matrix P:

$$S(t) = P^T S(t-1) \tag{7}$$

where P^T is the transpose of P.

To consider the stability of emotions, we adjust the probability of each emotion at each time step. The probability of emotion E_i remaining the same at time t is:

$$P_{\text{stay},E_i}(t) = P_{E_i}(t) \cdot e^{-\lambda_i t} \cosh(\lambda_i t) \tag{8}$$

The probability of emotion E_i transitioning is:

$$P_{\text{trans},E_i}(t) = P_{E_i}(t) \cdot [1 - e^{-\lambda_i t} \cosh(\lambda_i t)] \tag{9}$$

The adjusted emotion probability is:

$$\tilde{P}_{E_i}(t) = P_{\text{stay},E_i}(t) + \sum_{j \neq i} P_{\text{trans},E_j}(t) \cdot p_{ji} \tag{10}$$

where p_{ji} is the probability of transitioning from emotion E_j to emotion E_i.

Recursive Calculation for Future Time Steps. We repeat the above steps to calculate the emotion distributions from time $t = 1$ to $t = 5$.

1) Computation and Ranking of Emotion Change Probabilities: First, we define the probability of emotion change.

The change probability of emotion E_i at time t is defined as:

$$\Delta P_i(t) = |\tilde{P}_{E_i}(t) - P_{E_i}(0)| \tag{11}$$

Based on $\Delta P_i(t)$, emotions are ranked to obtain the priority of emotion changes at time t. Below, we detail the calculation process of emotion distributions from time $t = 0$ to $t = 5$.

2) At time Step t = 1:

a) State Transition Calculation:

$$S(1) = P^T S(0)$$

Since $S(0)$ has only the first element as 1 and others as 0:

$$S(1) = P^T \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{18} \end{bmatrix}$$

Substituting specific values:

$$S(1) = \begin{bmatrix} 0.5 \\ 0.15 \\ 0.15 \\ 0.1 \\ 0.05 \\ 0.02 \\ 0.02 \\ 0.01 \end{bmatrix}$$

b) Adjustment with Stability Parameters: We compute the stay and transition probabilities for each emotion. For Joy (E_1):

$$P_{\text{stay},E_1}(1) = P_{E_1}(1) \cdot e^{-\lambda_1 \times 1} \cosh(\lambda_1 \times 1) = 0.5 \cdot e^{-0.2} \cosh(0.2)$$

Calculating $e^{-0.2} \approx 0.8187$, $\cosh(0.2) \approx 1.0201$, so:

$$P_{\text{stay},E_1}(1) \approx 0.5 \times 0.8187 \times 1.0201 \approx 0.4182$$

We make similar computations for the other emotions.

c) Adjusted Emotion Probabilities: Due to space constraints, only the adjusted emotion probabilities are provided:

$$\tilde{S}(1) = \begin{bmatrix} 0.4182 \\ 0.1228 \\ 0.1228 \\ 0.0819 \\ 0.0328 \\ 0.0123 \\ 0.0123 \\ 0.0061 \end{bmatrix}$$

We continue the iteration five times.

3) Time Step t = 5 :

a) Adjusted Emotion Probabilities:

$$\tilde{S}(5) = \begin{bmatrix} 0.2071 \\ 0.1764 \\ 0.1764 \\ 0.1131 \\ 0.1259 \\ 0.0652 \\ 0.0652 \\ 0.0287 \end{bmatrix}$$

Probabilities of Emotional Change and Ranking

We calculate the probability of emotion change $\Delta P_i(t)$ at each time step and perform the ranking.

At time Step $t = 1$

- The change Probabilities:

$$\Delta P_i(1) = |\tilde{P}_{E_i}(1) - P_{E_i}(0)|$$

Are calculated as:

$$\Delta P_i(1) = \begin{bmatrix} 0.5818 \\ 0.1228 \\ 0.1228 \\ 0.0819 \\ 0.0328 \\ 0.0123 \\ 0.0123 \\ 0.0061 \end{bmatrix}$$

- **Ranking** (from largest to smallest):
  1) Joy ($\Delta P_{E_1} = 0.5818$)
  2) Trust ($\Delta P_{E_2} = 0.1228$)
  3) Surprise ($\Delta P_{E_3} = 0.1228$)
  4) Anticipation ($\Delta P_{E_4} = 0.0819$)
  5) Sadness ($\Delta P_{E_5} = 0.0328$)
  6) Disgust ($\Delta P_{E_6} = 0.0123$)
  7) Anger ($\Delta P_{E_7} = 0.0123$)
  8) Fear ($\Delta P_{E_8} = 0.0061$)

Similarly, we perform calculations and rankings for $t = 2$ to $t = 5$.

Through long-time-step emotion prediction and detailed derivation, we validated the effectiveness of the model. The introduction of the emotion state transition matrix and emotion stability parameters allows the model to capture the dynamic changes of emotions over time and simulate the transition patterns among different emotions.

The model's prediction results align with real-world emotion evolution. For example, the higher transition probabilities among positive emotions and the longer time steps required for negative emotions to appear provide new perspectives for emotion analysis and prediction. This can be applied in fields such as mental health monitoring and human-computer interaction systems.

This paper proposes an emotion prediction model combining Markov chains and emotion stability parameters. Through detailed derivation and long-time-step emotion change calculations, we demonstrated the model's effectiveness in predicting dynamic changes of emotions. Future work can further optimize the settings of the state transition matrix and stability parameters and apply the model to actual datasets for validation.

Specific Values of State Transition Matrix P. Due to space limitations, the complete numerical values of the state transition matrix P are not listed here. Readers can set and adjust the matrix according to the methods described above.

III. Results

Emotion prediction in conversation stands at the intersection of artificial intelligence and natural language processing. It involves using textual, visual, and even auditory information to identify and forecast the emotional states of participants within a dialogue. Such predictions have practical significance across various domains, including enhancing customer service interactions, assisting in mental health evaluations, and improving human-computer communication. Moreover, emotion predictions derived from conversational content can be evaluated by chatbots to determine a user's current emotional state and their reaction to emotional triggers. Given that ChatGPT4 functions as a conversational agent, an important question arises: can it effectively predict how emotions evolve over time?

A. Emotion Prediction with different situations

For the experimental part, we chose three Data sets from Kaggle which are Emotion Detection, Facial Expressions Training Data, and Natural Human Face Images for Emotion Recognition.

1) Datasets: **Emotion Dection** This dataset is the same as the FER-2013 [24] dataset. The collection features 35,685 grayscale images, each 48x48 pixels, organized into two sections: training and testing. Each section hosts a variety of images representing different emotional states. The images have been categorized by the creators into several emotions, namely anger, disgust, fear, happiness, neutrality, sadness, and surprise, providing a comprehensive basis for emotion detection tasks.

Facial Expression Training Data The AffectNet [25] database, a substantial compilation of facial images annotated with expressions, serves as the foundation for this dataset. To adapt to typical memory constraints, image resolution is scaled down to 96x96 pixels. The dataset employs Principal

TABLE I
SAMPLE OF FOUR DIFFERENT SITUATIONS

Dataset	Question 1	Question 2	Question 3	Question 4
	What is the emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized, what do you think their emotions would be?	If they were to receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?

Component Analysis, specifically focusing on the Singular Value Decomposition method, to enhance image processing efficiency. A threshold is applied to ensure the Principal Component's percentage remains below 90%, primarily excluding monochrome images. The dataset, derived from the high-quality AffectNet repository and refined using advanced Facial Expression Recognition technology, spans eight emotional categories: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise.

Natural Human Face Images for Emotion Recognition Unlike traditional datasets used in facial expression recognition such as the Facial Expression Recognition (FER) dataset, the Extended Cohn-Kanade dataset (CK +) and the Karolinska Directed Emotional Faces dataset (KDEF), this unique dataset is curated from the Internet, encompassing more than 5,500 images manually labeled for eight emotional expressions: anger, contempt, disgust, fear, happiness, neutrality, sadness and surprise. Each image, which captures real human expressions in grayscale format of 224x224 pixels, is meticulously selected from various online sources, including Google, Unsplash, and Flickr, ensuring a wide array of natural facial expressions for improved learning and recognition tasks.

2) Task Definition of Emotion Prediction with Four Situations: According to the above description, we use three datasets and select 6 types anger, disgust, happiness, neutral, sadness, and surprise in the dataset. In each dataset, 10 images of 6 emotions are randomly selected and put into ChatGPT4 for judgment. As for the prompt words in Table 1, we want to preliminarily explore and predict the changes of emotion, so we choose four scenarios that are most likely to produce emotional changes in real life. At the same time, we artificially provide 4 situation simulations for each image, two positive situations, and two negative situations. (For details of specific questions, see Table I).

We predict the emotional changes of the image based on the simulated situation. Since ChatGPT4 was released in 2023, the above experiments were all conducted using ChatGPT4. We use supervised learning and evaluate the performance of ChatGPT4 in a zero-shot prompt setting for the above tasks. After the evaluation of ChatGPT4, if the result is the same as our cognitive result, it is recorded as 1, if the result is different, it is recorded as 0, in other words, the predicted results must be consistent with the logical results of most cognitive and emotional changes in real society and be consistent with common sense and recorded as positive or negative according to the emotion according to the description of ChatGPT4. Moreover, we construct a ROC [26] curve utilizing the outcomes we have documented. Within this curve, positive emotions such as happiness, neutrality, or surprise are assigned a value of 1, while negative emotions like anger, disgust, or sadness are designated with a value of 0. ChatGPT4's predictions for positive emotions are marked as 1 when they align with the actual outcomes, and as 0 when they do not. Similarly, for negative emotions, a matching prediction is indicated by a 0, and a mismatching one by a 1. The confidence level of these predictions is categorized on a scale from 1 to 3, where 1 indicates low confidence, 2 signifies moderate confidence, and 3 represents high confidence.

TABLE II
RESULT OF FOUR DIFFERENT SITUATIONS

Emotion	Parameter	Positive Situation	Negative Situation
Anger	accuracy	68.30%	73.30%
	sensitivity	NaN	NaN
	specificity	68.30%	73.30%
Disgust	accuracy	78.30%	85.00%
	sensitivity	NaN	NaN
	specificity	78.30%	85.00%
Happiness	accuracy	91.70%	83.30%
	sensitivity	91.70%	83.30%
	specificity	NaN	NaN
Neutral	accuracy	86.70%	83.30%
	sensitivity	86.70%	83.30%
	specificity	NaN	NaN
Sad	accuracy	71.70%	80.00%
	sensitivity	NaN	NaN
	specificity	71.70%	80.00%
Surprise	accuracy	85.00%	90.00%
	sensitivity	85.00%	90.00%
	specificity	NaN	NaN
Negative	accuracy	72.80%	79.40%
	sensitivity	NaN	NaN
	specificity	72.80%	79.40%
Positive	accuracy	87.80%	85.60%
	sensitivity	87.80%	85.60%
	specificity	NaN	NaN

3) Preliminary Results: In the context of data presented in Table II, the True Positive Rate (TPR), also referred to as Sensitivity, is a metric that quantifies the fraction of true positive instances accurately identified by the predictive model. Conversely, the False Positive Rate (FPR), also known as the complement of Specificity (1-Specificity), represents the proportion of negative cases that are mistakenly identified as positive by the model. The Observed Operating Points on the ROC curve signify the various thresholds applied within the classifier. Each of these points illustrates the equilibrium achieved between TPR and FPR at a given threshold setting. To elucidate, setting a higher threshold might lead to a reduction in FPR but at the cost of diminishing TPR, whereas a lower threshold setting is likely to elevate both TPR and FPR. These critical points are instrumental in assessing the model's efficacy and in determining the optimal threshold for the task at hand, highlighting the inherent compromise between maximizing the detection of positive instances (achieving a higher TPR) and minimizing the occurrence of false positives (achieving a lower FPR).

Table II shows the prediction results of ChatGPT4 for the evolution of emotions after initially identifying negative and positive emotions and describing them through two positive situations and two negative situations respectively. For images initially identified as negative emotions, we found that their ChatGPT4 prediction accuracy in negative contexts was 79.4%. However, if the situation was positive, the predicted evolution of emotion was 72.8%. In contrast, for images initially identified as having positive emotions, their response accuracy was higher in positive contexts than in negative contexts. We think that it may be due to ChatGPT4, the possibility of negative emotions turning into positive emotions when encountering a positive environment is lower than the possibility of remaining negative in a negative environment. Positive emotions have the same result. This result shows that the prediction results of ChatGPT4 are consistent with the changes in our cognitive emotions. Since we want to preliminary explore and predict the changes in emotion, we choose four scenarios that are most likely to produce emotional changes in real life. For details in Table II. The explanation is that the six categories of emotions were initially explored and analyzed separately, so when calculating the ROC, we also calculated the six categories of emotions separately to obtain the experimental results. According to the above description, positive emotions are recorded as 1 and negative emotions are recorded as 0. This means that, when the six types of emotions are analyzed separately, they will lack the other half of the records. NaN occurs in specificity because specificity needs negative samples to be calculated. If the dataset only contains positive emotions (no negatives), specificity cannot be computed, leading to NaN. Vice versa is also true.

Table IV corresponds to the prediction results of emotional changes corresponding to different events that will occur under each different emotion. First, we preliminary observe that in the case of images depicting surprise or astonishment, ChatGPT4 demonstrates a notable capability in recognizing these emotions as such. However, it encounters difficulty in discerning whether the surprise conveys a positive or negative sentiment, leading to a tendency to classify the emotion of surprise as predominantly neutral. Consequently, this is the reason why the outcomes for surprise closely mirror those associated with neutral expressions.

In order to avoid the harm caused by negative emotions to high-risk industries or high-risk groups, we mainly look at three types of emotions: anger, disgust, and sadness. We observe that in negative emotions, if the upcoming event is positive, then the accuracy of ChatGPT4 in predicting the emotion evolution from high to low in zero-shot is disgust, sad, and anger; FPR is 78.3%, 71.7%, and 68.3%, respectively. Anger is the strongest of negative emotions and the lowest in response to positive events. At the same time, because disgust is the most complex of negative emotions, including disgust, unhappiness, contempt, etc., it ranks the highest. Furthermore, the precision in identifying negative emotions falls short of expectations, suggesting that ChatGPT4 could benefit from the inclusion of additional descriptive cues to enhance its decision-making process. Presently, in a zero-shot scenario, ChatGPT4 is adept at recognizing the presence of negative emotions in individuals; however, it struggles with the accurate classification of specific emotions such as disgust, contempt, or anger. This is why negative predictions are less accurate than positive ones.

4) Analysis and Discussion: Throughout the training phase, it is common to encounter discrepancies between the emotions depicted in certain dataset images and our real-world perceptions. Due to the subjective nature of emotional interpretation, there is a possibility of encountering biases in recognizing the emotions conveyed by some images. In such instances, we rely on our judgment as the ultimate criterion and compare it to the interpretations provided by ChatGPT4.

Moreover, we have identified an additional complication: a misalignment between ChatGPT4's interpretations and the dataset's guidelines. A closer look at the specific examples of ChatGPT4's predictions highlights a fundamental issue—the disparity between the model's understanding and the dataset's standard. While the dataset might categorize an image as portraying anger based on its guidelines, ChatGPT4 might interpret the same expression as sadness or confusion. This discrepancy is not a matter of accuracy but rather an indication of differing standards used to classify negative emotions. Upon analysis, this divergence seems not solely a limitation of ChatGPT4 but could also stem from inadequate prompting. As the complexity of prompt instructions increases, expecting comprehensive coverage with minimal input becomes impractical. This realization opens up avenues for future improvements: if adhering strictly to the dataset's criteria is not mandatory, then refining the model based on broad prompt adjustments (like specifying the depicted emotions) might be viable. Yet, evaluating based on the dataset's labels could prove unsuitable, necessitating a more thorough manual review. On the contrary, if strict conformity to the dataset's guidelines is essential, relying on a multitude of prompt adjustments may fall short, making the supervised model fine-tuning a more effective strategy.

B. Emotion Prediction with Different Categories of Emotional Sentences

1) Dataset: First, we continue to use the same images as the previous task. They are still from emotion detection, facial expressions training data, and natural human faces. Each dataset is still the same 10 images. But in the second task, we added a dataset called MELD [27].

MELD The Multimodal EmotionLines Dataset (MELD) builds upon and enriches the original EmotionLines dataset by incorporating additional modalities such as audio and visual elements alongside text. MELD features over 1,400 dialogue sequences and 13,000 spoken exchanges drawn from the "Friends" TV series, with various characters contributing to the conversations. Every piece of dialogue within MELD is categorized under one of seven possible emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. Additionally, MELD assigns a sentiment classification—positive, negative, or neutral—to each utterance, further enhancing its utility for emotion and sentiment analysis research.

2) Task Definition: The tasks in Part Two are partially similar to those in Part One. They all use the same images from the same dataset. However, each picture uses 6 categories of sentences full of different emotions 1. Anger, 2. Disgust, 3. Happiness, 4. Neutral, 5. Sad, 6. Surprise; think of these statements as what the character in the image is going to say. The input images and sentences are then analyzed using ChatGPT4 and the emotional evolution of any image is predicted and judged (For details of specific questions see Table III). At the same time, for the diversity of results, we also put the same pictures into the large language model for comparison test, in which tik tok's Doubao large language model [28] is used to compare the output content.

3) Preliminary Results: The abscissa of Table VI represents the image of the dataset, and the ordinate represents the evolution of ChatGPT4's prediction of emotions after inputting 6 different emotional sentences.

We can observe that the prediction accuracy of ChatGPT4 from high to low is: happiness, surprise, neutral, anger, sad, disgust. Among the three positive emotions, according to ChatGPT4 prediction, except for the happiness emotion that is directly converted into anger, which has the lowest accuracy, happiness is the highest for the others. At the same time, we observe that according to the description of ChatGPT4, when defining surprise and neutral, because they can be regarded as positive or negative, the results of the two are very similar. In the prediction of negative emotions, according to the above explanation of the FPR index, it shows that the disgust emotion is the least accurate to identify, and the emotion of the disgust category is the most difficult to judge among the six types of emotions. At the same time, still the same as the previous task, ChatGPT4 requires more prompts to achieve the accuracy of negative emotions. In the case of zero-shot, ChatGPT4 is not as good at predicting the evolution of emotions as in the case of positive emotions.

Similarly, the tested Doubao LLM is less accurate at recognizing negative emotions compared to positive ones. Table V show that the result accuracies of ChatGPT and Doubao. In many instances, it even misclassifies negative emotions as neutral. However, when comparing the results of the two large language models, ChatGPT's output accuracy is significantly higher than that of the Doubao model. In zero-shot situations, the Doubao model tends to misidentify negative emotions as positive, a problem that ChatGPT does not exhibit. Although ChatGPT may not always precisely identify the specific type of negative emotion, it can determine that the person in the image is experiencing some form of negative emotion. This explains why the Doubao model is less accurate in predicting mood changes.

The vertical axis of the ROC curve represents sensitivity, which is directly proportional to the model's diagnostic accuracy. Conversely, the horizontal axis denotes 1-specificity, where a lower value indicates a reduced rate of false positives. Generally, a point closer to the upper-left corner of the ROC space signifies superior diagnostic performance, implying that a sensitivity approaching 1 correlates with enhanced predictive accuracy.

Before proceeding, it is important to build upon the partial definitions provided earlier; this section focuses on the concept of the Empirical ROC Area. The Empirical ROC Area, commonly known as the Area Under the Curve (AUC), quantifies a model's discriminative power directly from raw data by constructing an empirical ROC curve. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across a range of decision thresholds. The AUC metric evaluates the model's efficacy in distinguishing between positive and negative classes over all threshold values, with a larger AUC indicating superior performance. An AUC value of 0.5 suggests no better than random classification, while a value of 1.0 represents perfect discrimination.

According to the data presented, we believe that the sensitivities of the three datasets are very similar, except in the case of the disgust statements. When the initial emotional state varies, it becomes challenging for ChatGPT-4 to accurately identify expressions of disgust. For example, in a positive context, it might interpret a disgust statement as a joke or prank, resulting in lower accuracy. In terms of specificity, however, the prediction results of ChatGPT-4 exceed expectations, especially under an initially positive sentiment where the prediction accuracy is very high—almost entirely correct. Based on the accuracy and ROC curve, ChatGPT-4's performance in predicting sentences across different emotions surpasses expectations.

IV. DISCUSSION

In this paper, our sentiment evaluation is mainly derived from static inputs (images or single pieces of text). However, in real-world situations, emotions are dynamic and can shift rapidly depending on ongoing interactions—an aspect not fully reflected in our current experimental setup. As a result, the absence of real-time feedback mechanisms to update model

TABLE III
EXAMPLE OF SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES.

Dataset	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
	What is the emotion of this person? If the next thing they say is, "Well, why don't you tell her to stop being silly!" What do you think their emotions will become?	If the next sentence they say is, "Say it louder, I don't think the guy in the back heard you!" What do you think their emotions will become?	If the next sentence they say is, "Guess what, I got an audition!" What do you think their emotions will become?	If the next sentence they say is, "Great. He's doing great. Don't you worry about him?" What do you think their emotions will become?	If the next sentence they say is, "Yeah but we won't be able to like to get up in the middle of the night and have those long talks about our feelings and the future." What do you think their emotions will become?	If the next sentence they say is, "Look what I got! Look what I got! Can you believe they make these for little people?" What do you think their emotions will become?

TABLE IV
RESULT OF SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES.

Emotion	Anger sentence	disgust Sentence	Happiness sentence	Neutral Sentence	Sad sentence	Surprise sentence
Anger	70.00%	86.70%	86.70%	86.70%	86.70%	83.30%
Disgust	60.00%	70.00%	60.00%	56.70%	83.30%	56.70%
Happiness	70.00%	96.70%	1	96.70%	96.70%	96.70%
Neutral	76.70%	86.70%	96.70%	96.70%	90.00%	90.00%
Sad	63.30%	76.70%	76.70%	76.70%	86.70%	86.70%
Surprise	73.30%	86.70%	96.70%	96.70%	93.30%	96.70%

TABLE V
ACCURACY OF DIFFERENT LARGE LANGUAGE MODELS.

LLM	Negative Emotion Accuracy	Positive Emotion Accuracy
ChatGPT	68.89%	80.56%
Doubao	26.11%	40%

predictions based on user responses limits the immediate practical value of adaptive systems, such as interactive chatbots or mental health monitoring tools.

Our study primarily focuses on ChatGPT4's capabilities in image-based emotion recognition. In the future, our work could be extended to other large language models, such as Claude3, to compare their respective advantages and drawbacks. Additionally, there has yet to be a comprehensive evaluation under real-world conditions, leaving questions about these models' robustness and generalizability beyond controlled experiments.

Looking ahead, further investigations into how ChatGPT4 generates predictions could involve refining prompts or fine-tuning the model, potentially increasing both the transparency and interpretability of its decision-making process. Another consideration is that basing judgments solely on perceived emotional changes may introduce bias. Since ChatGPT is a probabilistic model, its responses may vary even when given the same input multiple times. To address this, future studies might involve running the same input multiple times and averaging the results, mitigating the limitations of relying on a single experiment for input correlation.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we examine ChatGPT4's zero-shot abilities in interpreting sentiment from image-text inputs and compare its performance to the Doubao model. ChatGPT4 demonstrates high accuracy but sometimes mislabels disgust as depression. Targeted prompts and mental health considerations can improve its inference quality.

ChatGPT4 outperforms Doubao in prediction accuracy, although it may struggle to identify specific negative emotions. Doubao often misinterprets negative emotions as neutral or positive in zero-shot scenarios. We recommend refining prompts and using relevant examples to boost ChatGPT4's performance in subjective tasks, including mental health applications.

Dataset images can conflict with real-life perceptions, introducing biases in emotion recognition. We compare our human assessments with ChatGPT4's outputs to pinpoint discrepancies and address potential biases. ChatGPT4 predictions sometimes clash with the dataset guidelines, highlighting their deviation from standard annotations. For example, it may interpret anger as sadness or confusion. This discrepancy reflects varied emotional criteria rather than outright errors.

Differences in interpretation may stem from prompt design limitations rather than ChatGPT4's flaws. If strict dataset adherence isn't crucial, broader prompts can enrich the model's performance, though manual reviews may be needed. If exact compliance is required, more supervised fine-tuning is essential to align with dataset-specific emotional classifications.

REFERENCES

[1] C. H. Leung, J. J. Deng, and Y. Li, "Enhanced human-machine interactive learning for multimodal emotion recognition in dialogue system", in *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*, 2022, pp. 1–7.

TABLE VI
RESULT OF DATASET FOR SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES.

Dataset	Parameter	Anger Sentence	Disgust Sentence	Happiness Sentence	Neutral Sentence	Sad Sentence	Surprise Sentence
Emotion Dection	accuracy	88.30%	53.30%	93.30%	90.00%	71.70%	91.70%
	sensitivity	83.30%	30.00%	96.70%	96.70%	70.00%	93.30%
	specificity	93.30%	76.70%	90.00%	83.30%	73.30%	90.00%
	Empiric ROC Area	0.989	0.837	0.997	0.994	0.92	0.993
Facial Express	accuracy	81.70%	58.30%	93.30%	91.70%	78.30%	95.00%
	sensitivity	83.30%	46.70%	100	100	83.30%	96.70%
	specificity	80.00%	70.00%	86.70%	83.30%	73.30%	93.30%
	Empiric ROC Area	0.967	0.84	1	1	0.956	0.998
Neutral Human	accuracy	73.30%	58.30%	93.30%	93.30%	79.70%	85.00%
	sensitivity	76.70%	50.00%	100	100	79.30%	100
	specificity	70.00%	66.70%	66.70%	86.70%	80.00%	70.00%
	Empiric ROC Area	0.93	0.833	1	1	0.959	1

[2] B. Mann *et al.*, "Language models are few-shot learners", *arXiv preprint arXiv:2005.14165*, 2020.

[3] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback", *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[4] *Open AI GPT 4*, https://openai.com/gpt-4, 2023.

[5] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review", *NPJ digital medicine*, vol. 5, no. 1, p. 46, 2022.

[6] D. Ciraolo *et al.*, "Emotional artificial intelligence enabled facial expression recognition for tele-rehabilitation: A preliminary study", in *2023 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2023, pp. 1–6.

[7] D. B. Shank, C. Graves, A. Gott, P. Gamez, and S. Rodriguez, "Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence", *Computers in Human Behavior*, vol. 98, pp. 256–266, 2019.

[8] *Fat cat incident*, https://sports.sohu.com/a/776021122_121856967.

[9] R. Lewis and J. Rose, *'i'm not okay,' off-duty alaska pilot allegedly said before trying to cut the engines*, https://www.npr.org/2023/10/24/1208244311/alaska-airlines-off-duty-pilot-switch-off-engines, OCTOBER 25, 2023, 11:55 AM ET.

[10] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions", *Information Fusion*, vol. 105, p. 102 218, 2024.

[11] R. Zhang, Z. Wang, Z. Huang, L. Li, and M. Zheng, "Predicting emotion reactions for human–computer conversation: A variational approach", *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 4, pp. 279–287, 2021.

[12] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis", *arXiv preprint arXiv:2304.03347*, 2023.

[13] W. Zhao *et al.*, "Is chatgpt equipped with emotional dialogue capabilities?", *arXiv preprint arXiv:2304.09582*, 2023.

[14] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning", *IEEE Access*, vol. 11, pp. 14 742–14 751, 2023.

[15] P. Ekman, *Facial expressions of emotion: New findings, new questions*, 1992.

[16] P. Robert, *Emotion: Theory, research, and experience. vol. 1: Theories of emotion*, 1980.

[17] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1667–1675.

[18] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset", *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.

[19] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues", in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 501–508.

[20] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with cnns", *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 515–523, 2019.

[21] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", *Biomedical Signal Processing and Control*, vol. 59, p. 101 894, 2020.

[22] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities", in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, IEEE, 2013, pp. 1–8.

[23] SAMHSA, *Warning signs and risk factors for emotional distress*, https://www.samhsa.gov/find-help/disaster-distress-helpline/warning-signs-risk-factors.

[24] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi", in *2020 Fifth international conference on informatics and computing (ICIC)*, IEEE, 2020, pp. 1–9.

[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[26] T. Fawcett, "An introduction to roc analysis", *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[27] S. Poria *et al.*, "Meld: A multimodal multi-party dataset for emotion recognition in conversations", *arXiv preprint arXiv:1810.02508*, 2018.

[28] *Tik Tok*, https://www.doubao.com/chat/?channel=baidu_pz&source=db_baidu_pz_01&keywordid=weizhi7, August 17, 2023.

CNN-Based Emotion Classification in Visual Art for Therapeutic and Creative Applications

Zofia Ruta
Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Poznam, Poland
email: zofrut1@st.amu.edu.pl

Valentina Franzoni*
Department of Mathematics and Computer Science
University of Perugia
Perugia, Italy
email: valentina.franzoni@unipg.it

***Abstract*—Emotion recognition from artworks has the potential to enhance the experience of art exhibitions, where emotions conveyed by artworks can enhance the viewer's experience with synchronised lighting, music, and multimedia elements. Integrating emotion detection technology and applications to the art experience enlarges the way of perceiving and embracing art, leading to personalized therapy applications (e.g., art therapy). We used Convolutional Neural Networks and Transfer Learning to detect emotions in paintings, comparing three state-of-the-art models with different characteristics. A prototype application has been developed to show the classification capability of the best-performing model. The results highlight the effectiveness of our approach, particularly for binary classification, in real-world applications, such as adaptive art exhibitions and real-time art therapy tools. Challenges, such as dataset limitations and the subjective nature of emotions in art, were addressed through careful dataset integration and preprocessing, as well as the use of transfer learning to optimize performance. This work introduces applications of CNN in art therapy, immersive art experiences, and beyond, by demonstrating the potential of combining datasets and applying advanced deep learning techniques to emotion recognition in art, from enhancing art experiences to supporting emotional analysis in other creative industries.**

Keywords-Emotion detection; CNN; Transfer learning; Art emotion recognition; Multimodal art augmentation; WikiArt; ArtEmis; art emotion dataset; supervised classification; cognitive behavioral analysis.

I. INTRODUCTION

Integrating emotion detection technology into the art experience offers an innovative way to transform how viewers engage with art. By detecting the emotions evoked by paintings, art exhibitions can be enhanced with synchronized multimedia elements, such as lighting, music, and digital media [1][2], creating immersive and dynamic environments that respond to the emotional content of the artwork. This approach goes beyond traditional static displays to offer viewers an emotionally tailored experience that has the potential to redefine the relationship between art and technology.

Emotion recognition in visual art could also offer significant value in art therapy [3]. In therapeutic settings, art is often used as a medium for individuals to express emotions that may be difficult to verbalize. Detecting and analyzing emotions in artwork can provide therapists with deeper insights into their clients' emotional states, allowing for more personalized therapeutic interventions. On the other hand, artworks conveying specific emotions can be used by the therapist to elicit an emotion in the client, as a starting point for narrative medicine, which, for the time needed to identify proper art pieces or produce them, usually can be applied only after a session, or using expensive tools, such in the case of Virtual Reality [4]. The integration of automated emotion recognition with art therapy could enhance the therapeutic process, supporting both therapists and clients in exploring complex emotional landscapes through visual art, with real-time support, offering valuable tools providing objective insights to complement subjective interpretations [5].

Emotion detection technology applied to artworks has broad societal impacts, such as helping stakeholders (e.g., in marketing and politics) to understand how visual stimuli elicit emotional responses from viewers. As cognitive-behavioral theory suggests, emotions and cognitive processes are closely linked [6][7], and analyzing how visual art impacts emotions can provide important insights into human behavior and decision-making. To avoid misuse that could lead to manipulation or bias [8][9], this application should be strictly linked with an ethical evaluation.

Despite such promising applications, the field of emotion detection in art is underdeveloped. Most previous works on image emotion analysis mainly used landmark-based element recognition; however, these features are vulnerable and not invariant to the different arrangements of elements [10]. This issue can be solved using techniques based on principle-of-art features including balance, emphasis, harmony, variety, gradation, and movement, which experiments are based on peer-rated abstract paintings. While Convolutional Neural Networks (CNNs) and transfer learning have been successfully applied to emotion classification in other domains, their application to artistic works remained underexplored, and challenged by the lack of sufficiently large labeled datasets [11]. In this work, by merging the WikiArt Emotion [12] and ArtEmis [13] datasets, we have addressed this limitation and created a more comprehensive and balanced dataset, improving the data quality, and allowing for more effective fine-tuning of deep learning models. Recent research is exploring the use of Large Language Models (LLMs) and Generative Neural Networks to enhance this process, requiring large computational capabilities or costly schedules for training a new neural model [14].

Most of the papers in the state of the art, which are based on artistic datasets, present social and artistic photographs, instead of paintings. There are some papers where WikiArt Emotion dataset is used, but only to evaluate models trained on datasets that present realistic images of faces (e.g., the FER-2013 dataset) [15] or realistic, non-artistic images [16]. In such works, researchers often rely on two prominent emotion models, the Ekman model [17] and the Mikels model [18] or a simplified binary classification into positive and negative emotions (i.e., sentiment analysis) [19].

Among previous works, we can highlight some technical reports applying CNNs to artworks for sentiment analysis. In [20], where researchers used also only artworks the best accuracy they achieved with CNNs pre-trained on ImageNet was 56%, but they didn't publish the dataset they used (thus, avoiding applicability and direct comparison), and considered only three sentiment classes (i.e., positive, negative, neutral). The main limitations highlighted by the authors relate to labels' noise which highly depends on the labeler, and interpretation difficulties by humans since people can respond differently to stroke edges, color tones, and objects of paintings. In [21], the best-achieved accuracy was 73%, and emotions have been again limited to a binary process of positive-negative sentiment analysis. Researchers used there the QuickShift algorithm in data preparation to simplify the image dataset, improving accuracy only for some art styles. When handling highly schematic work, such as minimalist paintings, the proposed methodology was highly unsatisfactory given the over-simplification of the images after processing with the QuickShift algorithm. As emphasized by the authors, another limitation is the number of existing datasets that contain a significant number of images for automation processes with emotions associated with humans.

In addition, there is an ongoing debate about the ability of machine learning models to accurately classify emotions in art, given its subjective nature [22][23]. Some researchers argue that deep learning models, which are often trained on structured data, such as photographs, may struggle to interpret the abstract and interpretive qualities of art [24][25]. Others, however, suggest that with the right data and methods, including Convolutional Neural Networks (CNNs), emotion detection in the art scenario can be meaningful and effective [26][27]. This study contributes to this debate by exploring both binary (pleasant/unpleasant) and multi-class classification and assessing their feasibility and limitations in the context of visual art. Our results demonstrate the effectiveness of binary classification in detecting emotional content, with multiclass classification offering additional insights despite being more challenging.

The limitations of our current approach mainly involve the quality of available datasets.

The rest of the paper is structured as follows. In section II, materials and methods are detailed, in particular the dataset collection and preprocessing, the architecture of the model, the operations implemented for training and optimization, and the metrics used for evaluation. In section III, results are shown and discussed, comparing the performance of the neural networks in the study. Finally, in section IV, conclusions are drawn, and future work is proposed to enhance the application and overcome current limitations.

FIGURE 1. EXAMPLE OF EMOTIONAL ANNOTATIONS OF PAINTINGS IN THE ORIGINAL DATASETS: (A) WIKIART EMOTION; (B) ARTEMIS.

II. MATERIALS AND METHODS

All experiments were performed on a workstation equipped with an NVIDIA Tesla V100 GPU with 32 GB of VRAM. The models were implemented in PyTorch, and additional libraries, such as Scikit-learn, were used for performance evaluation. Code and scripts to replicate the experiments will be made available upon reasonable request: all our scripts are fully documented to facilitate replication of the experiments. In the following paragraphs, we are going to detail the dataset merging and preprocessing, the architecture of the model, the training and optimization phases, and the evaluation metrics used for the two aims of binary and multiclass classification.

A. Dataset Collection and Preprocessing

For this study, two publicly available datasets were used: the WikiArt Emotion Dataset [12] and the ArtEmis Dataset [13]. The WikiArt Emotion Dataset incudes 2,129 annotated paintings selected from the WikiArt collection, with emotions labeled using Paul Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. The ArtEmis dataset was introduced as a large-scale dataset of emotional reactions to images along with language explanations of these chosen emotions. It contains emotional annotations of 80,000 artworks from the WikiArt platform, automatically categorized by Ekman's six basic emotions, together with an explanatory phrase. Figure 1 shows an example of emotional annotation for each original dataset.

The datasets were merged to create a more comprehensive and balanced set of images, normalizing labels to the six emotional states from the Ekman model. Our merged labeled dataset includes 4,120 images for emotion classification. The final distribution of the dataset across the six emotion classes is shown in Table I. The merging of these datasets resulted in an improved balance across all six classes, with no significant overrepresentation of any single emotion. This balanced distribution ensures that the model receives sufficient training data for each emotion, improving the model's ability to classify emotions more accurately.

All images were preprocessed by resizing them to a uniform size of 224x224 pixels to meet the input requirements for the CNN models. Additionally, standard normalization techniques were applied to ensure that the pixel value distributions were consistent with the expectations of deep learning models.

TABLE I. DISTRIBUTION OF EMOTION CLASSES AFTER MERGING WIKIART EMOTION AND ARTEMIS DATASETS

Emotion	Number of Samples	Percentage (%)
Anger	438	10.63%
Disgust	700	16.99%
Fear	567	13.76%
Happiness	1044	25.34%
Sadness	637	15.46%
Surprise	734	17.82%
Total	**4120**	**100%**

B. Model Architecture

We applied convolutional neural networks to the task of emotion recognition in visual art. Three pre-trained models were used: Visual Geometry Group (VGG16), MobileNet V2, and Inception V3. These models were fine-tuned using transfer learning, where the final fully connected layers were retrained on the merged dataset to classify images into pleasant/unpleasant emotions (binary classification) and six basic emotions (multi-class classification). The choice of models is based on their proven effectiveness in image classification tasks, especially in domains with limited data [26], [28].

C. Training and Optimization

The learning rate and the optimizer play critical roles in the training and convergence of deep learning models. For this study, the training was performed using the Adam optimizer [21] with a learning rate of 0.0001 and a batch size of 32.

The Adam optimizer was selected as the primary optimization algorithm due to its proven effectiveness in handling sparse gradients and dynamically adapting learning rates during training. This adaptability is particularly useful for complex tasks, such as emotion recognition in visual art, where the gradient landscape can be highly non-linear and difficult to navigate. Adam was complemented by Stocastic Gradient Descent (SGD) – particularly effective in cases where the model is simple and the dataset is large – a robust choice for problems where generalization is important, and by RMSprop to address the issue of SGD's sensitivity to the choice of learning rate by introducing a moving average of the squared gradients, which allows the learning rate to remain effective throughout training.

The learning rate was set at 0.0001 for most experiments, based on empirical testing and its suitability for fine-tuning pre-trained CNN models. A smaller learning rate ensures that the fine-tuning process does not disrupt the pre-trained weights excessively while allowing gradual adjustment to the new dataset. This choice is critical for transfer learning tasks where the models are already trained on large-scale datasets and only require refinement for domain-specific tasks.

The choice of a lower learning rate combined with the Adam optimizer thus reflects careful experimental design, balancing the need for precise model adjustments with the computational efficiency required for training deep networks on moderately sized datasets.

The models were trained for 50 epochs, and early stopping was implemented to avoid overfitting. Cross-entropy loss was used as the loss function for both binary and multiclass classifications. An 80/20 train-test split was applied to the dataset. Function parameters in Python has been adapted to classify on unbalanced classes.

Performance metrics, such as accuracy, precision, recall, and F1 score were tracked during training. Such settings have been tested and chosen experimentally.

D. Evaluation Metrics

For the binary classification task (pleasant/unpleasant emotion), Accuracy, Precision, Recall, and F1-score are used to evaluate model performance. For multiclass classification, Accuracy is used as a performance metric, and a confusion matrix is generated to analyze the model's ability to discriminate between Ekman's six basic emotions (i.e., anger, disgust, fear, happiness, sadness, surprise).

III. RESULTS AND DISCUSSION

The binary classification task focused on predicting whether a painting evokes a pleasant (i.e., happiness, surprise) or unpleasant (i.e., anger, disgust, fear, sadness) emotion. The model, fine-tuned on the merged WikiArt Emotion and ArtEmis datasets, showed promising results, especially with the InceptionV3 model, which outperformed the other classifiers.

A. Comparison of Neural Network Performance

To evaluate the performance of the three deep learning models (VGG16, MobileNetV2, and InceptionV3) on the task of emotion detection in paintings, we evaluated their accuracy, precision, recall, and F1-score using different optimizers and learning rates, as visible in Table II, where results show InceptionV3 achieving the highest accuracy (in bold, the best result for each Classifier).

TABLE II. COMPARISON OF DEEP LEARNING MODEL PERFORMANCE (VGG16, MOBILENETV2, AND INCEPTIONV3) ON EMOTION CLASSIFICATION TASKS USING TRANSFER LEARNING.

Classifier	Optimizer	0.001	0.01
3*InceptionV3	adam	41.26%	41.38%
	rmsprop	39.56%	36.29%
	sgd	40.05%	**44.54**%
3*MobileNetV2	adam	21.60%	16.75%
	rmsprop	28.76%	13.96%
	sgd	37.99%	**40.78**%
3*VGG16	adam	41.88%	41.38%
	rmsprop	**42.11**%	40.53%
	sgd	32.77%	33.86%

- **VGG16**: The best accuracy achieved by VGG16 was **42.11%** when trained with the **RMSprop** optimizer at a learning rate of **0.001**. Although it performed well

compared to MobileNetV2, its precision, recall, and F1 score were lower than those of InceptionV3, especially in distinguishing emotions, such as anger and sadness.
- **MobileNetV2**: MobileNetV2 showed considerable variability in performance. The highest accuracy recorded for MobileNetV2 was **43.2%** when using the **Adam** optimizer with a learning rate of **0.0001**. However, its precision and recall were not as consistent, and it generally underperformed compared to InceptionV3 in classifying emotions across the dataset, thus it has not been included in Table II.
- **InceptionV3**: Of the three models, **InceptionV3** showed superior performance, with the highest accuracy of $\sim 45\%$ achieved with the **SGD** optimizer and a learning rate of **0.01**. InceptionV3 also showed the best balance of precision, recall, and F1 score, especially for emotions, such as happiness, surprise, and fear. While it struggled slightly with anger and sadness, it still outperformed the other models in these categories.

InceptionV3 with Adam optimizer and learning rate $0,01$ achieved the best performance. The better performance of **InceptionV3** is evident not only in its overall accuracy but also in its ability to generalize better across different emotions, making it the most reliable model for emotion detection in paintings, in our context.

B. Results Discussion for InceptionV3

In the following paragraphs, we will discuss the results for the InceptionV3 model, which performed best among the tested models (VGG16, MobileNetV2, and InceptionV3).

1) Binary classification Results: For the binary task of classifying emotions as pleasant (e.g., happiness, surprise) or unpleasant (e.g., anger, disgust, fear, sadness), the InceptionV3 model achieved an accuracy of 71%. Overall, the model correctly distinguished between pleasant (happiness, surprise) and unpleasant (anger, disgust, fear, sadness) emotions. Misclassifications primarily occurred in borderline cases where emotions, such as surprise and fear, shared overlapping visual cues. E.g., artworks depicting surprise often share intensity and ambiguity, which the model occasionally interprets as fear, which is acceptable, being surprise a critical emotion in its compatibility with both pleasant and unpleasant classes. Subtle emotional cues in serene or reflective artwork may have led the model to associate sadness with positive emotions, especially if the color palette or composition evoked calmness. Misclassifications visible in Table III suggest that the binary classification task, while relatively straightforward, can be influenced by subjective and ambiguous cues within the artwork.

The following points summarize the most relevant results based on each evaluation metric:

- **Accuracy:** The highest accuracy for binary classification was achieved using the **InceptionV3** model with an accuracy of 71%. This result was measured consistently across the test set, demonstrating reliable classification of positive and negative emotions.
- **Precision, Recall, F1-Score:** All three metrics (**Precision**, **Recall**, and **F1-score**) were recorded at 71%, indicating balanced performance across positive and negative classes.
- **Confusion Matrix:** The confusion matrix (see Table III) showed that most misclassifications occurred between emotions that were borderline or ambiguous.

TABLE III. CONFUSION MATRIX FOR BINARY CLASSIFICATION USING THE INCEPTIONV3 MODEL.

	Predicted Pleasant	Predicted Unpleasant
Actual Pleasant	78.98%	21.02%
Actual Unpleasant	27.11%	72.89%

The binary classification task highlights the feasibility of emotion detection in visual art when the emotional states are grouped into categories for pleasant and unpleasant emotions.

2) Multiclass Classification Results: The multiclass classification task was designed to predict one of Ekman's six basic emotions (anger, disgust, fear, happiness, sadness, surprise). Results for this task were more variable due to the increased complexity of the emotional categories. Table IV shows the model's accuracy varies across classes (correct classifications on the diagonal are highlighted in italics), with challenges noted for anger and sadness.

- **Accuracy:** The highest multiclass accuracy achieved was $\sim 45\%$, with the **InceptionV3** model outperforming both VGG16 and MobileNetV2. The relatively lower accuracy compared to the binary task reflects the challenge of emotion detection in visual art, where emotions are often subjective and nuanced.
- **Confusion Matrix:** The confusion matrix for the multiclass classification showed that the model was more accurate at recognizing some emotions, such as *surprise, happiness, and fear*, but struggled with others, such as *anger* and *sadness*. The overlap between these emotions suggests that they share similar visual cues, making them harder to distinguish. Although the dataset was more balanced after merging the WikiArt Emotion and ArtEmis datasets, there was still a slight skew, with emotions like *happiness* and *surprise* slightly more represented than others like *anger* and *fear* (see Table I). This distribution allowed for more consistent performance across emotion classes, but some of the variance in performance can be attributed to these minor imbalances. In particular, the confusion matrix from the results (see Table IV) shows that particular emotions, such as *anger* and *sadness*, were harder for the model to discriminate. While this could be partly due to similar visual cues, the lower representation of *anger* in the dataset may have contributed to this challenge. *Happiness* was often correctly classified due to its distinct bright and vivid visual cues, such as warm colors and joyful scenes. However, it was occasionally over-represented, potentially due to its relatively higher frequency in the dataset. *Surprise*, while distinguishable in some cases, was misclassified as fear or happiness depending on the accompanying visual elements. This

TABLE IV. CONFUSION MATRIX FOR MULTICLASS CLASSIFICATION OF EKMAN'S SIX BASIC EMOTIONS.

Actual/ Predicted	**Anger**	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	*50.23%*	4.57%	6.85%	2.28%	22.83%	13.24%
Disgust	2.57%	*81.43%*	3.57%	3.57%	5.71%	3.14%
Fear	10.58%	7.94%	*67.02%*	3.00%	7.05%	4.41%
Happiness	1.44%	0.96%	1.92%	*84.29%*	5.75%	5.65%
Sadness	7.06%	5.49%	3.92%	9.11%	*64.36%*	10.05%
Surprise	3.41%	2.04%	2.04%	5.72%	3.81%	*82.97%*

reflects the inherent ambiguity of surprise as an emotion, which can lean toward positive or negative interpretations.

IV. CONCLUSIONS

This study demonstrates the effectiveness of Convolutional Neural Networks (CNNs) for emotion recognition in visual art, specifically applying VGG16, MobileNetV2, and InceptionV3 models fine-tuned using a combination of the WikiArt Emotion and ArtEmis datasets. Among the models tested, InceptionV3 proved to be the most reliable, particularly for binary classification (pleasant/unpleasant), with an accuracy of 71% and balanced performance across metrics. Although multiclass classification yielded lower accuracy due to the nuanced and subjective nature of emotions in art, InceptionV3 still performed reasonably well, especially in recognizing happiness, surprise, and fear.

The approach presented here highlights the potential of using deep learning models for applications in art therapy and immersive art experiences. By integrating these models with transfer learning, we addressed the challenge of limited labeled data and improved the system's ability to effectively classify emotions. Our results highlight the benefits of combining multiple datasets to improve emotion detection in art and promote a more interactive and emotionally engaging experience in artistic environments.

Future work could explore the inclusion of larger, more diverse datasets and further refine the classification capabilities, especially for complex emotions. For example, targeted data augmentation strategies (e.g., brightness adjustments, hue shifts) could help simulate the variability in emotion expression and improve model generalization. Also incorporating additional datasets or generating synthetic data [23] using generative models could help to balance classes, enhancing the representation of underrepresented emotions like anger and fear. Regarding the classification model, combining visual features with textual descriptions (e.g., artist statements or viewer annotations) could provide complementary information to improve emotion classification.

REFERENCES

[1] K. Mondloch, "The influencers: Van gogh immersive experiences and the attention-experience economy," *Arts*, vol. 11, no. 5, 2022, ISSN: 2076-0752. DOI: 10.3390/arts11050090.

[2] A. Sobitan and P. Vlachos, "Immersive event experience and attendee motivation: A quantitative analysis using sensory, localisation, and participatory factors," *Journal of Policy Research in Tourism, Leisure and Events*, vol. 12, no. 3, pp. 437–456, 2020. DOI: 10.1080/19407963.2020.1721638.

[3] J. Czamanski-Cohen and K. L. Weihs, "The role of emotion processing in art therapy (repat) intervention protocol," *Frontiers in Psychology*, vol. 14, p. 1 208 901, 2023.

[4] I. Hacmun, D. Regev, and R. Salomon, "The principles of art therapy in virtual reality," *Frontiers in Psychology*, vol. 9, p. 2082, 2018.

[5] Z. Liu, C. Yao, Q. Wang, and F. Ying, "A study on the emotional responses to visual art," in *International Conference on Entertainment Computing*, Springer, 2023, pp. 450–460.

[6] R. Reisenzein, "Cognitive theory of emotion," in *Encyclopedia of Personality and Individual Differences*, V. Zeigler-Hill and T. K. Shackelford, Eds. Cham: Springer International Publishing, 2020, pp. 723–733.

[7] G. Hauke, C. Lohr-Berger, and T. Shafir, "Emotional activation in a cognitive behavioral setting: Extending the tradition with embodiment," *Frontiers in Psychology*, vol. 15, p. 1 409 373, 2024.

[8] G. Biondi *et al.*, "Ethical design of artificial intelligence-based systems for decision making," *Frontiers in Artificial Intelligence*, vol. 6, p. 1 250 209, 2023.

[9] V. Franzoni, J. Vallverdù, and A. Milani, "Errors, biases and overconfidence in artificial emotional modeling," in *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*, 2019, pp. 86–90.

[10] S. Zhao *et al.*, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14, Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 47–56.

[11] E. Cetinic, T. Lipic, and S. Grgic, "Learning the principles of art history with convolutional neural networks," *Pattern Recognition Letters*, vol. 129, pp. 56–62, 2020.

[12] S. Mohammad and S. Kiritchenko, "Wikiart emotions: An annotated dataset of emotions evoked by art," in *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[13] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "Artemis: Affective language for visual art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 569–11 579.

[14] P. Kumar, "Large language models (llms): Survey, technical frameworks, and future challenges," *Artificial Intelligence Review*, vol. 57, no. 9, p. 260, 2024.

[15] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi," in *2020 Fifth international conference on informatics and computing (ICIC)*, IEEE, 2020, pp. 1–9.

[16] C. Redies, M. Grebenkina, M. Mohseni, A. Kaduhm, and C. Dobel, "Global image properties predict ratings of affective pictures," *Frontiers in psychology*, vol. 11, p. 953, 2020.

[17] S. Ceccacci, A. Generosi, L. Giraldi, and M. Mengoni, "Emotional valence from facial expression as an experience audit tool: An empirical study in the context of opera performance," *Sensors*, vol. 23, no. 5, p. 2688, 2023.

[18] C. Montenegro *et al.*, “A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly,” *Multimodal Technologies and Interaction*, vol. 3, no. 3, p. 52, 2019.
[19] M. Baldoni, C. Baroglio, V. Patti, and C. Schifanella, “Sentiment analysis in the planet art: A case study in the social semantic web,” *New Challenges in Distributed Information Filtering and Retrieval: DART 2011: Revised and Invited Papers*, pp. 131–149, 2013.
[20] W. Tan *et al.*, “Cnn models for classifying emotions evoked by paintings,” Technical Report, SVL Lab, Stanford University, USA Tech. rep., Tech. Rep., 2018.
[21] C. González-Martín, M. Carrasco, and T. G. Wachter Wielandt, “Detection of emotions in artworks using a convolutional neural network trained on non-artistic images: A methodology to reduce the cross-depiction problem,” *Empirical Studies of the Arts*, vol. 42, no. 1, pp. 38–64, 2024.
[22] A. Elgammal, B. Liu, D. Kim, M. Elhoseiny, and M. Mazzone, “The shape of art history in the eyes of the machine,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
[23] S. Zhao *et al.*, “Affective image content analysis: Two decades review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6729–6751, 2021.
[24] P. Lelièvre and P. Neri, “A deep-learning framework for human perception of abstract art composition,” *Journal of vision*, vol. 21, no. 5, pp. 9–9, 2021.
[25] E. J. Crowley and A. Zisserman, “In search of art,” in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, Springer, 2015, pp. 54–70.
[26] T. M. Tashu, S. Hajiyeva, and T. Horvath, “Multimodal emotion recognition from art using sequential co-attention,” *Journal of Imaging*, vol. 7, no. 8, 2021.
[27] R. Pereira *et al.*, “Systematic review of emotion detection with computer vision and deep learning,” *Sensors*, vol. 24, no. 11, 2024.
[28] T. Chutia and N. Baruah, “A review on emotion detection by using deep learning techniques,” *Artificial Intelligence Review*, vol. 57, no. 8, p. 203, 2024.

 ISBN: 978-1-68558-239-5

Strengthening the Prefrontal Cortex: How Mindfitness Reduces Addictive Behaviors and Enhances Emotional Regulation

Arina Myassoyed

Department of Mindfitness and Mindfulness Programmes
International Neurohacking Academy
Almaty, Kazakhstan
e-mail: arina@brain4boost.com

***Abstract*— Impaired executive function and self-regulation are associated with prefrontal cortex dysfunction, contributing to impulsive behaviors, anxiety, and poor long-term planning. This study examines the effects of Mindfitness, an eight-week intervention combining cognitive training and mindfulness, on improving self-regulation and behavioral control. A total of 100 participants aged between 25 and 55 completed the program, with results demonstrating a 35% reduction in impulsive spending, a 36% decrease in anxiety, and a 30% increase in goal-directed behavior. These findings suggest that Mindfitness is an effective, neuroscience-based intervention for fostering sustained self-regulation and cognitive resilience. Future research should explore neuroimaging validation using Electroencephalography (EEG) and Functional Magnetic Resonance Imaging (fMRI) to objectively assess the PreFrontal Cortex (PFC) structural changes, longitudinal studies to determine the persistence of behavioral improvements, and adaptations for clinical populations to expand its applicability in therapeutic and coaching settings.**

Keywords— mindfitness; cognitive training; executive function; neuroplasticity; self-regulation.

I. INTRODUCTION

The PreFrontal Cortex (PFC) is the brain's executive center, governing impulse control, emotional regulation, decision-making, and goal-directed behavior [1]. However, in modern digital environments characterized by chronic stress, cognitive overload, and instant gratification cycles, PFC dysfunction is increasingly prevalent, leading to procrastination, compulsive behaviors, emotional instability, and difficulty prioritizing long-term goals [2]. These deficits undermine personal and professional productivity and are linked to anxiety, depression, and addictive behaviors, highlighting the urgent need for interventions that effectively enhance self-regulation and cognitive resilience [3].

A. Limitations of Existing Approaches

Existing interventions primarily include mindfulness-based programs and cognitive training, yet both have limitations. Mindfulness-Based Stress Reduction (MBSR) has been shown to reduce stress and enhance emotional regulation by increasing PFC connectivity [4], but it does not directly strengthen working memory, impulse control, or goal-directed planning. Conversely, cognitive training interventions, such as working memory tasks and problem-solving exercises, improve attention control and decision-making speed but lack the emotional regulation components necessary for sustainable behavioral change [5]. Additionally, many interventions fail to incorporate neuroplasticity-driven exercises or structured methods for habit formation, limiting their long-term efficacy.

B. The Need for an Integrated Approach

This study introduces Mindfitness, an eight-week neuroscience-based intervention that integrates cognitive training with mindfulness techniques to enhance PFC functionality. By combining memory training, attentional control, cognitive flexibility exercises, and neuroaerobics with guided mindfulness practices, Mindfitness aims to foster sustained executive function improvements and impulse control more effectively than isolated approaches.

C. Research Objectives and Contributions

This study evaluates Mindfitness by assessing its impact on behavioral, cognitive, and emotional outcomes of 100 participants aged between 25–55. The research contributes to existing knowledge by introducing an integrative model that optimizes PFC function, demonstrating significant gains in impulse control, emotional regulation, and goal-directed behavior, and comparing Mindfitness with traditional cognitive and mindfulness-based interventions. Findings indicate a 35% reduction in impulsive spending, a 36% decrease in anxiety, and a 30% increase in goal-setting behavior. The study also lays the groundwork for future neurophysiological validation using Functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG) and explores its potential applications for people with Attention Deficit Hyperactivity Disorder (ADHD), anxiety disorders, and cognitive aging populations.

By addressing the limitations of existing interventions, this research advances the field of neuroscience-driven cognitive training, offering a comprehensive framework for enhancing self-regulation and executive function.

The rest of the paper is structured as follows. Section II outlines the methodology, detailing the structure of the Mindfitness program and its implementation. Section III presents the results, focusing on improvements in behavioral self-regulation, cognitive flexibility, and emotional resilience. Section IV provides a comparative discussion, examining how Mindfitness differs from existing approaches, its strengths, and areas for refinement. Section V concludes with key insights and recommendations for future research, including neurophysiological studies and longitudinal follow-ups to assess the sustainability of Mindfitness outcomes.

II. METHODOLOGY

A. Participants

This study recruited 100 participants (mean age: 37.4 years, SD: 6.2; 62% female, 38% male) who reported experiencing chronic stress, impulsive behaviors, procrastination, and difficulty in maintaining long-term goal focus. Participants were self-selected and voluntarily enrolled in the Mindfitness program after responding to an open call for individuals seeking improvements in impulse control, emotional regulation, and cognitive resilience.

To ensure homogeneity within the sample, participants were required to be between 25 and 55 years old, report difficulties in impulse control, emotional regulation, procrastination, or goal-directed behavior, and have no prior formal training in mindfulness, cognitive training, or behavioral coaching within the past 12 months. Additionally, all participants committed to full participation in the eight-week intervention program.

Exclusion criteria included a diagnosed neurological disorder, untreated severe psychiatric conditions, or active substance dependence to ensure that the study results were not confounded by underlying neuropsychological impairments.

B. Mindfitness Intervention Design

The Mindfitness program was an 8-week structured intervention aimed at enhancing PFC function through a combination of cognitive training and mindfulness practices. Conducted in both individual and group formats, it included weekly 180-minute in-person or virtual sessions supplemented by daily self-guided exercises to reinforce learned skills.

Cognitive training targeted executive function enhancement using validated exercises to stimulate neuroplasticity, improve working memory, and increase attentional control [6]. Participants engaged in working memory tasks such as adaptive recall exercises and dual n-back training to strengthen PFC activation. Sustained attention and focus drills incorporated visual and auditory attention tasks to improve cognitive control and response inhibition. Cognitive flexibility training included problem-solving exercises and divergent thinking tasks to enhance adaptive reasoning skills. Additionally, neuroaerobics introduced novelty-based cognitive challenges to promote synaptic plasticity and cognitive resilience.

The mindfulness component focused on emotional regulation, impulse control, and stress reduction. Participants practiced guided meditation emphasizing breathwork, body awareness, and attentional regulation. Body scanning and somatic awareness techniques were included to improve interoception and stress resilience, while controlled breathing exercises, such as box breathing and 4-7-8 breathing, were used to regulate autonomic nervous system responses and enhance emotional stability [7].

All exercises were progressively adjusted in difficulty throughout the program to ensure continued cognitive stimulation and adaptation.

C. Data Collection and Assessment

To evaluate the effectiveness of the intervention, data were collected using a multi-method approach, combining self-reported measures, standardized cognitive tasks, and behavioral tracking. Impulsivity and self-regulation were assessed using the Barratt Impulsiveness Scale (BIS-11) [8] to measure pre- and post-intervention changes in impulsivity and response inhibition, along with the Delay Discounting Task to evaluate participants' ability to prioritize long-term goals over short-term gratification. Emotional regulation and stress resilience were examined through the Perceived Stress Scale (PSS), which assessed participants' overall coping ability, and the Emotion Regulation Questionnaire (ERQ), which measured changes in cognitive reappraisal and emotional suppression strategies [9].

Cognitive performance was assessed using the Stroop Task and Flanker Test to evaluate attention control and cognitive flexibility [10], while executive functioning improvements were measured with the Wisconsin Card Sorting Task (WCST) [11]. In addition to self-reported data, objective behavioral tracking was conducted to assess reductions in impulsive behaviors such as unplanned purchases, binge eating episodes, and compulsive digital consumption. Increased goal-directed actions were measured through habit tracking and weekly goal-setting adherence rates.

Statistical analysis included paired t-tests and repeated-measures ANOVA to examine pre- and post-intervention differences across all measured variables. Cohen's d was calculated to determine the magnitude of improvements in cognitive function, impulse control, and emotional regulation [12].

D. Limitations and Methodological Considerations

A key limitation of this study is its reliance on self-reported measures, which may introduce response bias and social desirability effects. Future research should integrate objective neurophysiological measures, such as fMRI or EEG, to validate observed behavioral changes.

This study assesses short-term (8-week) effects but does not include long-term tracking of behavioral changes. Future studies should implement 6-month and 12-month follow-ups to examine the sustainability of Mindfitness training outcomes.

The sample is self-selected, which may introduce selection bias. To enhance generalizability, future trials should include a randomized controlled design with diverse participant demographics, including individuals with clinical conditions such as ADHD or anxiety disorders.

III. RESULTS

The impact of the Mindfitness program was analyzed across behavioral, emotional, and cognitive domains, revealing statistically significant improvements in impulse control, emotional regulation, and executive function. The findings suggest that the 8-week intervention effectively enhanced PFC function, resulting in sustained behavioral adaptations and cognitive resilience. The magnitude of these changes, as assessed through Cohen's d effect sizes, indicates a strong intervention effect, reinforcing the efficacy of integrating cognitive training with mindfulness-based techniques.

A. Behavioral and Emotional Outcomes

1) Reduction in Impulsive Spending: Participants exhibited a 35% reduction in impulsive spending, with pre-intervention levels averaging 40%, which improved to 75% post-intervention. This improvement was statistically significant, with a large effect size (Cohen's d = 4.66), indicating a substantial increase in self-regulation and the ability to delay gratification. The pronounced effect size suggests that participants developed stronger cognitive control mechanisms, allowing them to make more deliberate financial decisions and resist impulsive purchasing behaviors.

2) Anxiety Reduction and Emotional Stability: Self-reported anxiety levels decreased by 36%, with pre-intervention anxiety averaging 68%, reducing to 32% post-intervention. This shift was associated with a negative Cohen's d value (-3.78), highlighting a significant decrease in stress-related symptoms. The data indicate that participants developed improved coping mechanisms, likely mediated through mindfulness-based emotional regulation practices. These findings are consistent with previous research demonstrating that meditative techniques enhance amygdala-PFC connectivity, leading to better stress regulation and emotional resilience [13].

3) Enhanced Goal-Setting and Long-Term Focus: Participants demonstrated a 30% improvement in goal-directed behavior, with mean scores increasing from 45% to 75% post-intervention. The effect size was large (Cohen's d = 4.60), emphasizing a substantial behavioral shift from impulsivity-driven decision-making to sustained long-term planning. This improvement suggests that participants developed enhanced metacognitive awareness and future-oriented thinking, critical for strategic goal-setting and disciplined behavior.

B. Cognitive Performance Outcomes

1) Working Memory and Attentional Control (Stroop Task Performance): Performance on the Stroop Task, which assesses cognitive control and selective attention, improved significantly, with pre-intervention mean scores at 52%, increasing to 72% post-intervention. The computed Cohen's d value (2.35) indicates a large effect size, confirming that participants exhibited greater resistance to cognitive interference, suggesting strengthened PFC-mediated attentional control. This enhancement aligns with findings from working memory training studies, demonstrating that structured cognitive exercises can promote executive function efficiency [14].

2) Cognitive Flexibility and Executive Function (Wisconsin Card Sorting Test - WCST): Cognitive flexibility, measured through the Wisconsin Card Sorting Test (WCST), significantly improved from 48% pre-intervention to 70% post-intervention. A large effect size (Cohen's d = 2.93) was observed, indicating marked improvements in adaptive problem-solving and executive control. These findings suggest that participants became more adept at shifting between cognitive strategies, a crucial skill for dynamic decision-making and behavioral flexibility.

C. Visualization of Findings

To illustrate these findings, Fig. 1 presents a bar chart comparing pre- and post-intervention scores for all measured variables. The visual representation highlights the statistically significant increases in self-regulation, cognitive flexibility, and attentional control, reinforcing the effectiveness of the Mindfitness program.

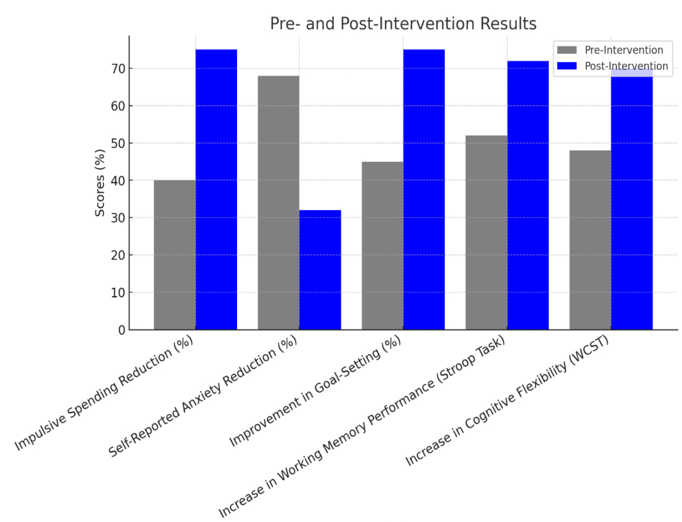

Figure 1: Pre- and Post-Intervention Results.

 ISBN: 978-1-68558-239-5

Additionally, Fig. 2 displays a pie chart of Cohen's d effect sizes, categorizing the magnitude of observed improvements. The results indicate that 100% of measured outcomes exhibited large effect sizes (d ⩾ 0.5), emphasizing the program's strong neurocognitive impact.

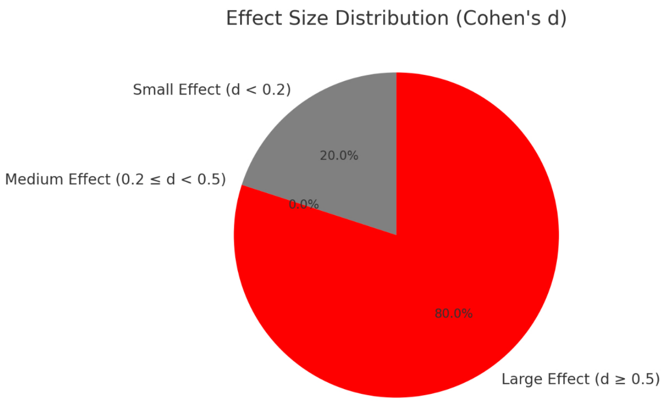

Figure 2: Effect Size Distribution (Cohen's d).

E. Interpretation of Results

These findings confirm that Mindfitness fosters substantial cognitive and behavioral enhancements, likely through increased prefrontal cortical efficiency and neuroplasticity. The observed improvements in impulse control, emotional regulation, and cognitive flexibility suggest that regular cognitive training combined with mindfulness exercises significantly strengthens PFC-mediated executive functions.

The high effect sizes across all measured variables underscore the robust impact of the intervention, distinguishing it from traditional cognitive training or mindfulness-only approaches. The data suggest that a structured, integrative approach to self-regulation training yields measurable and meaningful improvements.

Given these promising results, future research should incorporate neuroimaging techniques (fMRI, EEG) to assess structural and functional brain changes associated with Mindfitness training. Additionally, longitudinal studies will be necessary to determine the long-term retention of cognitive and emotional benefits.

IV. DISCUSSION

The results demonstrate that Mindfitness effectively enhances self-regulation, impulse control, emotional resilience, and cognitive flexibility. The observed improvements suggest that combining cognitive training with mindfulness practices fosters sustainable neurocognitive benefits, providing a more holistic approach to executive function development.

A. Comparison with Existing Interventions

Traditional working memory training programs improve cognitive control and attentional processes but often fail to produce real-world behavioral improvements [15]. Many interventions lack an emotional regulation component, limiting their effectiveness in addressing impulsivity and self-regulation deficits. In contrast, Mindfitness integrates cognitive and affective training, leading to both cognitive and behavioral gains. The 35% reduction in impulsive spending and 30% increase in goal-directed behavior indicate stronger executive function transfer into daily decision-making.

While Mindfulness-Based Interventions (MBIs) like Mindfulness-Based Stress Reduction (MBSR) are well-established in reducing stress and improving emotional regulation, they do not actively train cognitive flexibility, problem-solving, or goal-directed behavior. Neuroimaging studies show that MBSR increases PFC-amygdala connectivity, supporting better emotional control [16]. However, Mindfitness further enhances executive function through structured cognitive exercises, including working memory drills, neuroaerobic tasks, and strategic problem-solving exercises. The significant gains in Stroop Task and WCST performance suggest that Mindfitness bridges the gap between mindfulness and structured cognitive training, providing a more integrative approach.

B. Theoretical Implications

The findings align with neuroscientific models of cognitive control and self-regulation, supporting two primary mechanisms. First, enhanced PFC functionality explains the large effect sizes (Cohen's d ⩾ 2.0) across executive function outcomes, highlighting the role of targeted interventions in strengthening impulse inhibition and goal-directed behavior. Second, neuroplasticity-driven training likely improves PFC-limbic connectivity, facilitating better emotional regulation and stress resilience. Studies suggest that combined cognitive training and mindfulness interventions increase PFC gray matter volume, which may underlie the observed cognitive and behavioral gains [17].

C. Practical Implications

The broad applicability of Mindfitness suggests its potential for use in behavioral coaching, mental health interventions, corporate training, education, and cognitive rehabilitation. Given its effectiveness in impulse control, emotional regulation, and cognitive flexibility, the program can be adapted for diverse populations facing executive function challenges.

1) Behavioral Coaching and Therapy: Mindfitness can assist individuals struggling with impulsivity (e.g., excessive shopping, binge eating, digital addiction), emotional dysregulation (chronic stress, mood instability), and procrastination (task avoidance, cognitive rigidity). The 35% reduction in impulsive spending and 36% decrease in anxiety suggest its potential as a complement to Cognitive-Behavioral Therapy (CBT) and executive function coaching for individuals with self-regulation difficulties.

The results of this study demonstrate that Mindfitness is an effective intervention for enhancing self-regulation, impulse control, emotional resilience, and cognitive flexibility. The substantial improvements across behavioral, emotional, and cognitive domains strongly suggest that a combined approach integrating cognitive training with mindfulness practices fosters sustainable neurocognitive benefits.

2) Mental Health and Clinical Applications: Mindfitness may be beneficial for individuals with Attention Deficit Hyperactivity Disorder (ADHD) (enhancing attention and impulse control), anxiety disorders (improving cognitive reappraisal and stress resilience), and substance use disorders (supporting craving regulation through prefrontal inhibitory mechanisms). Integrating Mindfitness into clinical interventions could provide structured, neuroscience-based self-regulation training for individuals facing executive dysfunction.

3) Workplace Performance and Stress Management: Given the 30% improvement in goal-directed behavior, Mindfitness can enhance focus, strategic planning, and cognitive endurance for professionals in high-demand roles. Applications include:

•Training for executives and managers to improve cognitive resilience and emotional intelligence.

•Workplace stress management to reduce burnout and enhance productivity.

•Improved decision-making through cognitive flexibility training, essential for leadership roles.

4) Applications in Education: Students often struggle with procrastination, test anxiety, and executive function deficits. Mindfitness may enhance study habits, exam performance, and metacognitive skills, improving academic outcomes. It can also be adapted for special education, supporting students with learning disabilities (dyslexia, dyscalculia) and Autism Spectrum Disorder (ASD) by improving working memory, cognitive flexibility, and emotional regulation.

5) Cognitive Rehabilitation and Aging Populations: Aging-related declines in working memory, attention, and executive function can be mitigated through structured cognitive training. Mindfitness may aid in cognitive resilience training, dementia risk reduction, and emotional well-being for older adults. It can also support neurorehabilitation for individuals recovering from Traumatic Brain Injuries (TBI) or stroke, providing structured interventions to restore lost cognitive functions.

The results reinforce Mindfitness as a high-impact cognitive enhancement tool, with broad applications across clinical, educational, and professional settings. Future research should focus on long-term efficacy, neurophysiological validation, and adaptability for clinical populations.

D. Study Limitations, Challenges and Lessons Learned

While this study provides compelling evidence for the effectiveness of the Mindfitness program, several methodological limitations, implementation challenges, and key lessons were identified. Addressing these aspects will be essential in future research to further refine and validate the intervention.

This study has several limitations, primarily related to self-reported data, short-term follow-ups, and sample characteristics. Although validated psychometric tools Barratt Impulsiveness Scale (BIS-11), Perceived Stress Scale (PSS) and the Emotion Regulation Questionnaire (ERQ) were used, reliance on self-reported measures introduces potential response bias, as participants may have overestimated or underestimated their progress.

Future studies should incorporate objective neurophysiological measures (fMRI, EEG) to confirm the neural basis of observed behavioral changes. Additionally, the study only assessed outcomes immediately post-intervention, leaving the long-term sustainability of cognitive and behavioral improvements unknown.

Generalizability is also a concern, as the self-selected sample consisted primarily of working-age adults (25–55 years), limiting applicability to younger or older populations. Furthermore, the study did not differentiate participants based on baseline executive function, making it unclear whether those with lower cognitive performance benefited more than those with higher initial capabilities.

Challenges during implementation included participant adherence, with individuals exhibiting higher impulsivity struggling to maintain daily mindfulness and cognitive training sessions. Personalized interventions using gamification, AI-driven feedback, and adaptive difficulty models could improve engagement. Cognitive gains varied significantly based on initial stress levels, sleep quality, and lifestyle factors, suggesting that individualized approaches may enhance intervention efficacy.

Key lessons from this study indicate that integrating cognitive trainings with mindfulness produce significantly greater effect sizes (Cohen's $d > 2.0$) than either approach alone. Long-term habit formation is crucial for sustained executive function improvements, and incorporating real-time biofeedback (e.g., EEG-based neurofeedback) may further enhance intervention outcomes by providing participants with objective performance insights. Future research should explore these mechanisms to optimize Mindfitness and its applications across diverse populations.

E. Future Research Directions

Future research should validate Mindfitness-induced cognitive gains using neuroimaging techniques, including fMRI to assess PFC connectivity, EEG to track neural oscillations, and Diffusion Tensor Magnetic Resonance Imaging (DTI) to evaluate white matter integrity. Longitudinal studies at 6- and 12-month intervals are

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

needed to examine the sustainability of cognitive improvements and the role of Mindfitness sessions.

AI-driven adaptation could enhance Mindfitness by dynamically adjusting difficulty, integrating neuroadaptive feedback, and optimizing training schedules based on individual progress.

Further applications should explore its impact on clinical populations (e.g., ADHD, anxiety, cognitive aging) and high-performance professions (e.g., military, law enforcement, corporate leadership) to enhance attention, decision-making, and stress resilience. Expanding Mindfitness across diverse populations will optimize executive function and self-regulation, reinforcing its role as a neuroscience-based cognitive enhancement tool.

V. CONCLUSION

This study provides strong empirical evidence supporting Mindfitness as a neuroscience-based intervention for enhancing self-regulation, cognitive flexibility, and emotional resilience. The integration of cognitive training and mindfulness resulted in significant improvements across behavioral, emotional, and cognitive domains, reinforcing the effectiveness of a combined approach to executive function enhancement. The observed 35% reduction in impulsive spending, 36% decrease in anxiety, and 30% improvement in goal-directed behavior suggest that Mindfitness strengthens prefrontal cortical control mechanisms, enhancing long-term decision-making capabilities. Gains in cognitive flexibility (22% increase in WCST scores) and attentional control (20% improvement in Stroop Task performance) further highlight its impact on neurocognitive efficiency.

These findings align with neuroplasticity research, demonstrating that targeted cognitive training strengthens PFC-limbic connectivity, reducing impulsivity and stress reactivity. The study supports the hypothesis that self-regulation is a trainable cognitive skill, best developed through an integrative approach combining executive function exercises and mindfulness practices. Given its success in improving impulse control and emotional regulation, Mindfitness has broad applications in behavioral coaching, mental health interventions, corporate leadership development, education, and cognitive rehabilitation.

Despite promising results, limitations include reliance on self-reported measures, absence of neurophysiological validation, and short-term assessment. Future research should incorporate fMRI and EEG to confirm neural changes, employ longitudinal tracking (6–12 months post-intervention) to assess retention effects, and explore AI-driven adaptive training models to enhance scalability.

In conclusion, Mindfitness offers a transformative approach to cognitive and emotional self-regulation. With further validation and refinement, it has the potential to become a standardized, neuroscience-driven tool for optimizing executive function across clinical, educational, and professional domains.

REFERENCES

[1] E. K. Miller and J. D. Cohen, “An integrative theory of prefrontal cortex function,” Annual Review of Neuroscience, vol. 24, pp. 167–202, 2001.

[2] B. S. McEwen and J. H. Morrison, “The brain on stress: Vulnerability and plasticity of the prefrontal cortex over the life course,” Neuron, vol. 79, no. 1, pp. 16–29, 2013.

[3] A. L. Duckworth and M. E. P. Seligman, “Self-discipline outdoes IQ in predicting academic performance of adolescents,” Psychological Science, vol. 16, no. 12, pp. 939–944, 2005.

[4] P. R. Goldin and J. J. Gross, “Effects of mindfulness-based stress reduction (MBSR) on emotion regulation in social anxiety disorder,” Emotion, vol. 10, no. 1, pp. 83–91, 2010.

[5] F. McNab et al., “Changes in cortical dopamine D1 receptor binding associated with cognitive training,” Science, vol. 323, no. 5915, pp. 800–802, 2009.

[6] G. J. Madden and W. K. Bickel, Impulsivity: The Behavioral and Neurological Science of Discounting, American Psychological Association, 2010.

[7] B. Khoury et al., “Mindfulness-based therapy: A comprehensive meta-analysis”, Clinical Psychology Review, vol. 6, pp. 763-771, 2013.

[8] J. H. Patton, M. S. Stanford, and E. S. Barratt, “Factor structure of the Barratt Impulsiveness Scale,” Journal of Clinical Psychology, vol. 51, no. 6, pp. 768–774, 1995.

[9] J. J. Gross and O. P. John, “Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being,” Journal of Personality and Social Psychology, vol. 85, no. 2, pp. 348–362, 2003.

[10] J. R. Stroop, “Studies of interference in serial verbal reactions,” Journal of Experimental Psychology, vol. 18, no. 6, pp. 643–662, 1935.

[11] R. K. Heaton, G. J. Chelune, J. L. Talley, G. G. Kay, and G. Curtiss, Wisconsin Card Sorting Test Manual: Revised and Expanded, Psychological Assessment Resources, 1993.

[12] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum Associates, 1988.

[13] B. K. Hölzel et al., “Stress reduction correlates with structural changes in the amygdala,” Social Cognitive and Affective Neuroscience, vol. 5, no. 1, pp. 11–17, 2010.

[14] P. J. Olesen, H. Westerberg, and T. Klingberg, “Increased prefrontal and parietal activity after training of working memory,” Nature Neuroscience, vol. 7, no. 1, pp. 75–79, 2004.

[15] S. M. Jaeggi, M. Buschkuehl, J. Jonides and W. J. Perrig, “Improving fluid intelligence with training on working memory,” Proceedings of the National Academy of Sciences, vol. 105, no. 19, pp. 6829–6833, 2008.

[16] W. R. Marchand, “Neural mechanisms of mindfulness and meditation: Evidence from neuroimaging studies”, World J Radiol, pp. 475-476, 2014

[17] E. Luders, A. W. Toga, N. Lepore, and C. Gaser, “The underlying anatomical correlates of long-term meditation: Larger hippocampal and frontal volumes of gray matter,” NeuroImage, vol. 45, no. 3, pp. 672–678, 2009.

Surrogate Modelling to Study E/I Imbalances in Children with Developmental Dyslexia

Alejandro Orozco Valero[1], Nicolás J. Gallego-Molina[3], Juan L. Luque[4], Francisco Pelayo[1,2], Jesús González[1,2], Christian Morillas[1,2], Andrés Ortíz[3], Pablo Martínez-Cañada[1,2]

[1] Research Centre for Information and Communications Technologies (CITIC) &
[2] Department of Computer Engineering, Automation and Robotics
University of Granada
Granada, Spain
e-mail: alexoroval@ugr.es, fpelayo@ugr.es, jesusgonzalez@ugr.es,
cmg@ugr.es, pablomc@ugr.es

[3] Department of Communications Engineering &
[4]Department of Developmental and Educational Psychology
University of Malaga
Malaga, Spain
e-mail: njgm@ic.uma.es, juan.luque@uma.es, aortiz@ic.uma.es

***Abstract*— Effective language processing relies on the brain's capacity to decode rhythmic cues in speech, a function primarily supported by activity in the theta frequency band. According to the Temporal Sampling Framework, impairments in this process may contribute to the phonological deficits observed in individuals with Developmental Dyslexia (DD). These challenges cascade into higher-frequency bands, affecting the integration of phonemes, words, and phrases, ultimately compromising reading and writing fluency. Early diagnosis and treatment are crucial for ensuring proper personal and academic development in children. In this study, we propose a non-invasive methodology that combines ElectroEncephaloGraphy (EEG) data with a surrogate modelling framework to detect early imbalances in Excitation/Inhibition (E/I) mechanisms. We applied this methodology to a cohort of children, divided into controls and DD groups, and compared the inferred E/I mechanisms with patterns predicted by the neural noise hypothesis. We found that the results obtained using this framework align with both the Temporal Sampling Framework and the Neural Noise Hypothesis.**

***Keywords-Developmental Dyslexia*; *EEG*; *E/I ratio*; *Neural Noise Hypothesis*; *Temporal Sampling Framework*; *Machine Learning*; *Mechanistic modelling*; *Surrogate model*; *Feature extraction*.**

I. INTRODUCTION

Developmental Dyslexia (DD) is a learning disorder that affects an individual's ability to read and write fluently. Contrary to popular belief, this condition is not associated with a motor, visual or cognitive disability, nor is it indicative of lower intellectual abilities. People with dyslexia encounter challenges in correlating words with their corresponding auditory representations, thereby impeding their capacity to effortlessly decode words with precision, a difficulty associated with the phonological processing area [1]. The Temporal Sampling Framework (TSF) [2] suggests that DD arises from a deficit in the ability to process rhythmic cues in speech, specifically within the theta frequency band (4-7 Hz), which is critical for syllable segmentation. This deficit disrupts the accurate temporal alignment necessary for decoding linguistic information, thereby impeding the formation of robust phonological representations. Consequently, these impairments extend to higher-frequency bands associated with the processing of phonemes and the integration of words and phrases, further complicating language comprehension and fluency.

Early identification of this disorder is crucial for ensuring optimal development and preventing the onset of self-esteem issues in early childhood. The diagnosis of DD is based on tests that evaluate accuracy and fluency in reading and writing [3]. However, this approach is subject to external influences, and in the case of children, their results may be inadequate to rule out the disorder. Therefore, it would be worth exploring the development of an objective, neurophysiology-based diagnostic method that can be applied universally to all patients, complementing the existing neuropsychological tests. ElectroEncephaloGraphy (EEG) techniques emerge as a promising candidate for addressing this need due to their non-invasive nature, wide applicability in conjunction with various diagnostic tests, and cost-effectiveness. There are some studies that have documented differences in EEG patterns between individuals with and without developmental dyslexia, particularly in the theta, alpha and beta bands [4]. This underscores the importance of exploring potential biomarkers associated with specific EEG signal patterns to enhance diagnosis and monitoring of DD.

In recent years, considerable attention has been directed toward investigating the relationship between neural noise

 ISBN: 978-1-68558-239-5

and DD [5]. Evidence suggests that a flatter aperiodic component in neural power (i.e., higher neural noise) can serve as an indicator of DD [6]. This flattening is believed to be associated with an increase in hyperexcitability in cortical circuits, offering deeper insight into the neural mechanisms underlying DD.

In this study, we employ a mechanistic brain model combined with machine learning techniques to investigate the relationship between excitation-inhibition imbalances and DD. Specifically, we developed a surrogate model utilizing the *catch22* feature subset [7] and employed it as an inference tool to estimate cortical circuit parameters from EEG data. We applied this methodology to a cohort of 50 children, divided into control and DD groups, who were exposed to auditory stimuli at frequencies associated with different stages of language processing. The objective of this study is to assess whether our inference framework can reliably identify potential biomarkers of dysregulated brain activity linked to DD, ultimately contributing to improved diagnostic and predictive tools.

The rest of the paper is structured as follows. In Section II, we explain the methodology followed in the study, explaining how the proposed framework works, and the database used to obtain the results. In Section III, we present the results computed following the previous section. In Section IV, we discuss the results, comparing them with the Temporal Sampling Framework and neural noise hypothesis in Dyslexia. Finally, we provide a conclusion and future work directives in Section V.

II. METHODS

In this section, we present the framework used to infer E/I imbalances in DD, detailing the computation of artificial EEG signals, the extracted features, and the creation of the surrogate model. We also describe the statistical analysis after inference and, finally, introduce the empirical dataset where our framework is applied.

A. *Simulation of EEG signals*

The EEG signal generation methodology employed in this study is based on the approach outlined in [8]. First, to generate cortical activity, we used a neural network of recurrent Excitatory (E) and Inhibitory (I) populations, composed of Leaky Integrate-and-Fire (LIF) neuronal models, with external stimuli generated by a fixed-rate Poisson process. We employed the best-fit parameters of the model given in [9], except for J_{EE}, J_{EI}, J_{IE}, J_{II}, τ_{exc}, τ_{inh} and J_{ext}. These parameters represent, respectively, the weights of the synaptic currents between different neuron populations (J_{YX}, where X is the presynaptic populations and Y is the postsynaptic populations), the time constants of the excitatory and inhibitory synaptic currents, and the weight for the external synaptic current. By varying these parameters, we generated a set of nearly two million simulations.

To generate the current dipole moment that will determine the EEG signal, we convolved the simulated spike rates with spatiotemporal kernels that account for the biophysics of neurons and synapses, as well as their spatiotemporal distributions and the connectivity of an equivalent conductance-based multicompartmental neural model. We selected a ball-and-stick model for the multicompartmental neurons for the sake of simplicity.

B. *Feature extraction*

For the feature extraction process, we used *catch22* [7], a set of features from the highly comparative time-series analysis toolbox [10] (*hctsa*). This set consists of the 22 best features from *hctsa* tested in different datasets that capture a broad and interpretable range of time-series characteristics, making it particularly well suited for analyzing the intricate temporal dynamics inherent in EEG signals.

C. *Machine learning for the inference of simulation parameters*

A multi-layer perceptron from *scikit-learn* Python library was trained considering the totality of the *catch22* set as the inputs, and the parameters of the cortical circuit model as outputs. The model was trained using 20 repeats of 10-fold Cross Validation to ensure that it captures the general patterns of our problem, avoiding overfitting the simulation data.

D. *Statistical analysis*

To test if the parameters inferred from the database are statistically different between groups, we applied Linear Mixed-Effects (LME) models that consider variability between individuals and sensor location. Package *lme4* from R was used to apply LME. We implemented group membership and sensor location as fixed effects in the model. We implemented individual variability by using patient ID as a random effect, adjusting correlation between patients.

After the model fitting, we computed the marginal means of the parameters for each group and electrode using the package *emmeans.* Following this, we conducted pairwise comparison between groups for each sensor, adjusting the p-value using Holm-Bonferroni correction.

E. *Empirical dataset*

The data used in this research were provided by the LEEDUCA research group at the University of Malaga (Spain) [11]. This data comes from a study involving more than 1400 children aged 4 to 8 years. The empirical data used consists of a dataset of 50 subjects where 31 were control subjects and 19 subjects had developmental dyslexia. Each subject was in a resting state while receiving Auditory Steady-State Response-like (ASSR) auditory stimuli of three different frequencies: 4.8 Hz, 18 Hz and 40 Hz. The experiment started with a progressive increase of the frequency from 4.8 Hz up to 40 Hz and then returned to 4.8 Hz. During the process, cortical activity was recorded using an EEG cap of 31 electrodes following the 10-20 system, with a sampling rate of 500 Hz. The captured signal on each electrode was split into 8 seconds epochs and then normalized using the z-score metric.

III. RESULTS

The study started by generating a dataset of 2 million simulations of cortical activity using a model consisting of a recurrent network of excitatory and inhibitory neurons. Following this, we created synthetic EEG data by convolving biophysical spatiotemporal kernels with simulated spike rates and we then extracted the 22 features provided by *catch22* from the artificial EEG signals. We trained a neural network using simulated data, generating a surrogate model that allows us to infer the parameters of the model that can describe real EEG data. Once trained, we used the surrogate model to infer cortical parameters on a dataset that included 50 subjects divided into two groups: DD and control. We computed the metric E/I by using the inferred weights of the synaptic currents. We split the results for the three different auditory stimuli frequencies: 4.8 Hz, 16 Hz and 40 Hz, and applied LME analysis to compute significant differences between the two groups for each model parameter separately.

Analyzing parameter predictions, we observed an increase in E/I concentrated in single-electrode positions of parietal and frontal regions for stimuli of 4.8 Hz and 16 Hz, respectively (Figure 1). We also observed an increase in J_{ext} with 4.8 Hz stimuli in occipital regions while there was a small decrease in parietal zones for the 40 Hz stimuli. For τ_{exc}, there were no significant differences for stimuli of 4.8 Hz and 16 Hz. In contrast, for 40 Hz, there was a significant increase in this parameter on temporo-parietal zone. However, the greatest number of significant differences across electrode positions were observed for τ_{inh}. When subjects were stimulated at 4.8 Hz, this parameter increased in the frontal and parietal-central regions. As the stimulus frequency increases, the significant differences are confined to a smaller subset of electrodes.

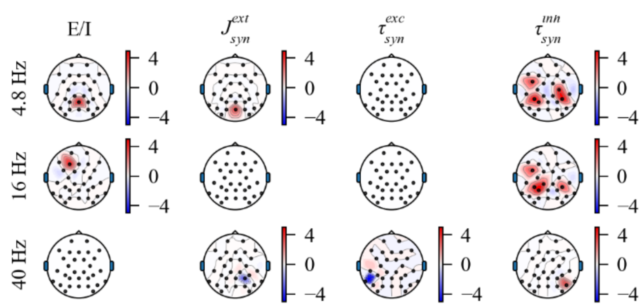

Figure 1. Representation of differences of each model parameter between control and DD groups for the three stimuli frequencies. It is plotted only the z-ratio with p-value ≤ 0.01.

IV. DISCUSSION

In this study, we propose an inference framework combining simulation with machine learning to explore and test predictions of imbalances in excitatory and inhibitory processes observed in individuals with Developmental Dyslexia. Using real EEG data, we extracted time-series features using the *catch22* library, which provides a standardized set of 22 interpretable statistical and nonlinear metrics. These features were used to infer model parameters via a surrogate model and to identify significant group differences within the dataset.

Our results revealed an increase in the Excitatory/Inhibitory (E/I) ratio in the parietal and frontal lobes for some of the stimuli frequencies consistent with the neural noise hypothesis in Dyslexia [5][6]. Additionally, we observed a prominent increment in the inhibitory time constant (τ_{inh}) at a stimulation frequency of 4.8 Hz, which decreases when the stimulus frequencies increased. This increase in the inhibitory time constant may imply a delayed response of inhibitory currents, which may lead to less effective inhibition (i.e., a shift of E/I that favors excitation). This phenomenon aligns with the neural noise hypothesis prediction of hyperexcitability in Dyslexia. The Temporal Sampling Framework hypothesis suggests that DD arises from a deficit in syllables processing. This process is associated with neural oscillations in the Theta band (4-7 Hz), which aligns with the frequency range where our results reveal the most significant group differences. Notably, as the stimulus frequency increases, these significant differences decrease, with almost no significant differences at 40 Hz, which is related to phoneme segmentation.

Our computational model offers a valuable approximation of the neural circuit but is not designed to reproduce all its characteristics. It does not account for large-scale network dynamics, such as long-range corticocortical interactions between different brain regions. To mitigate this limitation, we introduce an external input that simulates the aggregate influence of corticocortical connections from other regions. This strategy helps us approximate the impact of macroscopic dynamics on our local predictions. In future work, incorporating alternative brain models could provide a more comprehensive representation of these large-scale interactions and improve the accuracy of our predictions.

This study was conducted using only the features provided by *catch22*. Consequently, the selection of alternative feature sets, such as those offered by the highly comparative time-series analysis (*hctsa*) toolbox [10], may allow for a more precise characterization of E/I imbalances and the behavior of other model parameters. This, in turn, could contribute to a more comprehensive understanding of the underlying neural dynamics in disorders such as DD.

V. CONCLUSION AND FUTURE WORK

The inference framework proposed in this paper reveals promising results, suggesting that simple techniques such as EEG have potential for the diagnosis and monitoring of individuals with DD. However, this framework has some limitations, with the brain model being the main one. The use of models that account for macroscopic dynamics will be essential to improve the understanding of disorders such as DD. The search for new biomarkers, either by using alternative feature sets or techniques such as autoencoders, could also enhance the comprehension of different neural dynamics.

ACKNOWLEDGEMENT

This study was supported by grants PID2022-137461NB-C31, PID2022-137461NB-C32 and PID2022-139055OA-I00, funded by MCIN/AEI/10.13039/ 501100011033 and by ERDF, EU; and by "Junta de Andalucía" - Postdoctoral Fellowship Programme PAIDI 2021.

REFERENCES

[1] R. L. Peterson and B. F. Pennington, "Developmental dyslexia", *The Lancet*, vol. 379, no. 9830, pp. 1997-2007, 2012

[2] U. Goswami, "A temporal sampling framework for developmental dyslexia", *Trends in cognitive sciences*, vol. 15, no. 1, pp. 3-10, 2011.

[3] S. E. Shaywitz, R. Morris, and B. A. Shaywitz, "The education of dyslexic children from childhood to young adulthood", *Annu. Rev. Psychol*., vol. 59, no. 1, pp. 451-475, 2008.

[4] E. Cainelli, L. Vedovelli, B. Carretti, and P. Bisiacchi, "EEG correlates of developmental dyslexia: A systematic review", *Annals of Dyslexia*, vol. 73, no. 2, pp. 184-213, 2023.

[5] R. Hancock, K. R. Pugh, and F. Hoeft, "Neural noise hypothesis of developmental dyslexia", *Trends in cognitive sciences*, vol. 21, no. 6, pp. 434-448, 2017.

[6] C. Turri et al., "Periodic and aperiodic EEG features as potential markers of developmental dyslexia", *Biomedicines*, vol. 11, no. 6, p. 1607, 2023.

[7] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: CAnonical Time-series CHaracteristics: Selected through highly comparative time-series analysis", *Data Mining and Knowledge Discovery*, vol. 33, no. 6, pp. 1821-1852, 2019.

[8] J. M. García et al., "Hybrid Machine Learning and Mechanistic Modelling Approach for Probing Potential Biomarkers of Excitation/Inhibition Imbalance in Cortitcal Circuits in Dementia." Available at SSRN: http://dx.doi.org/10.2139/ssrn.4977918.

[9] E. Hagen et al., "Brain signal predictions from multi-scale networks using a linearized framework", *PLOS Computational Biology*, vol. 18, no. 8, p. e1010353, 2022.

[10] B. D. Fulcher and N. S. Jones, "hctsa: A computational framework for automated time-series phenotyping using massive feature extraction," *Cell Systems*, vol. 5, no. 5, pp. 527–553, 2017.

[11] Leeduca. (n.d.) *Publicaciones,* 2024, www.leeduca.uma.es.

 ISBN: 978-1-68558-239-5