

A Multimodal Approach for Emotion Recognition in Piano Performances

Alice Sironi

dept. of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
alice.sironi@polito.it

Paolo Belluco

LWT³ SRL
Milan, Italy
paolo.belluco@lwt3.com

Massimiliano Zanoni

dept. of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
massimiliano.zanoni@polito.it

Cristina Rottondi

dept. of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
cristina.rottandi@polito.it

Abstract—This study presents a multimodal Music Emotion Recognition (MER) framework designed to track the continuous emotional states experienced by performers during live piano executions. Within the context of the Internet of Sounds (IoS) and Internet of Musical Things (IoMusT), real-time affect recognition enables novel interactive applications. The proposed architecture fuses physiological embeddings generated by a dedicated LSTM-based surface electromyography (sEMG) encoder with deep acoustic representations extracted from SCHuBERT, a pre-trained audio emotion model. Experimental results on a dataset recorded under controlled conditions mimicking live concerts demonstrate consistent improvements across all evaluated pieces, achieving up to 0.863 validation and 0.831 test accuracy—a 1–2% gain over the EMG-only baseline. These findings highlight the value of combining bodily and auditory cues to capture performer expressivity, laying the groundwork for real-time multimodal interaction in live performance environments.

Index Terms—Music Emotion Recognition, Piano Performance, Multimodal Emotion Recognition.

I. INTRODUCTION

Throughout history, artistic expression has evolved alongside technological and societal transformations, continuously seeking new languages capable of reflecting emerging forms of perception, communication, and participation. Today, advances in artificial intelligence, sensing technologies, immersive media, and networked digital ecosystems are opening new possibilities for artistic creation and experience [19], while contemporary digital arts provide a privileged context for exploring novel forms of human–machine interaction, collaboration, and co-creation [10].

Within this evolving landscape—and particularly in the emerging fields of the Internet of Sounds (IoS) and Internet of Musical Things (IoMusT) [25]—musical performances can be conceived as interconnected ecosystems where performers, audiences, sensing devices, and intelligent computational services interact to create enriched and adaptive experiences. Enabling such experiences requires systems capable of understanding and responding to the expressive content of a performance in real time. For instance, real-time affective under-

standing could be repurposed to wireless control external stage peripherals during a performance (e.g., dynamic stage lighting, visual effects, or audience haptic devices). Consequently, the analysis and modeling of musical expressivity—and, in particular, of the emotions conveyed by performers—have become key challenges for the development of adaptive, audience-centered, and immersive musical applications.

Our work situates itself within the field of Music Emotion Recognition (MER), a research area at the intersection of music information retrieval, affective computing, and machine learning that aims to identify and characterize emotional content expressed through music and musical performance. Crucially, while MER often aims to recognize emotions perceived by a listener, this study focuses specifically on classifying the emotional state actually experienced and felt by the performer during the act of playing. While traditional MER approaches often focus on assigning emotional labels to entire musical excerpts, live performance scenarios require a more dynamic perspective, capable of capturing how emotional expression evolves throughout an execution. For this reason, our work is closely related to Music Emotion Variation Detection (MEVD), which focuses on tracking temporal changes in emotional expression during a performance. By modeling the evolution of the performer’s emotional and expressive state over time, MEVD provides a foundation for intelligent systems capable of adapting their behavior in response to ongoing musical events, thus enabling interactive and immersive audience experiences.

Traditionally, MER systems have focused almost exclusively on the auditory characteristics of music—such as tempo, melody, and harmony. However, emotional communication in live music emerges from the interplay of acoustic cues, timing variations, dynamics, body movements, muscular activity, and performer–instrument interaction. While listeners primarily perceive the resulting sound, a significant portion of expressive intent is embodied in gestures and physiological responses that are only partially reflected in the audio signal. Capturing such

complexity therefore requires a multimodal perspective. However, though multimodal approaches to emotion recognition have been widely addressed in the literature [20], they have rarely been explored in music live performances.

In a previous study, we investigated real-time MER from audio streams, demonstrating the feasibility of low-latency affective analysis based solely on acoustic information [27]. While promising, audio-based approaches provide only a partial representation of musical expressivity, as emotional intent is also reflected in performers' physiological and embodied responses. Building upon these observations, this paper explores a multimodal MEVD framework based on the fusion of embeddings extracted from acoustic and electromyographic (EMG) signals. The proposed approach combines modality-specific embeddings into a shared representation designed to capture complementary aspects of emotional expression during piano performance. Methodologically, both modality-specific encoders remain frozen during multimodal training, isolating the contribution of the fusion head to evaluate whether combining EMG-based representations of performers' physiological activity with high-level audio embeddings extracted from a pre-trained MER model provides a richer description of performer expressivity than single-modality approaches. The resulting multimodal representation is used to characterize emotional expression during live piano performance.

Our proposed architecture demonstrates consistent improvements in classification performance, achieving validation accuracies up to 0.863 and test accuracies up to 0.831 on the participant-selected repertoire, corresponding to a 1–2% gain over the EMG-only baseline across all evaluated pieces. These results highlight the effectiveness of multimodal fusion and provide deeper insight into the relationship between expressive intention and physiological response during musical performance.

II. RELATED WORK

MER is a research field at the intersection of Music Information Retrieval, affective computing, and machine learning, aiming to automatically identify emotions conveyed by or experienced through music [11] [14]. Early MER systems mainly relied on handcrafted audio features and conventional machine learning techniques, while recent advances in deep learning have enabled the automatic extraction of emotional representations directly from audio signals through architectures such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks [15], [16] [21]. More recently, research efforts have also addressed temporal modeling of emotions in music, giving rise to task such as MEVD, which focuses on capturing the evolution of emotional content over time within a musical piece [1].

Human emotions are inherently multimodal in nature, as they emerge from the interaction of behavioral, physiological, and contextual signals. As a result, multimodal learning has become a central paradigm in affective computing, where combining complementary sources of information has consistently been shown to improve the robustness and accuracy of emotion

recognition systems compared to unimodal approaches [20]. In this context, recent studies have explored a wide range of multimodal fusion strategies integrating audio, facial expressions, speech, body movements, and physiological signals to model complex emotional dynamics in a variety of domains [5] [6].

Despite these advantages, multimodal approaches remain relatively underexplored in MER. Most existing MER and MEVD systems still rely predominantly on audio recordings or symbolic musical representations, with only a limited number of studies investigating the integration of additional modalities related to the musical context or the performer [13]. Among the few exceptions, recent works have explored multimodal emotion recognition of playing musicians by combining acoustic signals with central and peripheral physiological measurements such as EEG and ECG [26].

In parallel, a substantial work has investigated expressive musical performance, focusing on how musicians communicate artistic and emotional intentions through performance behaviors. In particular, piano performance has been widely studied as a paradigmatic case for expressive music research, given the fine-grained control performers have over timing, dynamics, articulation, and physical gesture. Previous studies have shown that expressive intent is reflected in pianists' body movements and posture, and motion capture technologies have been extensively used to analyze kinematic patterns associated with musical expression [23] [24]. Similarly, physiological measurements such as surface electromyography (sEMG) have been employed to investigate muscular activity during performance [17], providing insight into performers' internal states that are not directly observable from audio signals [7] [28] [8]. Although these works are generally not framed as emotion recognition studies, they consistently demonstrate that performance-related modalities encode rich and structured information related to expressive intent.

Previous studies have explored multimodal affective computing and expressive music performance using combinations of audio, motion, or physiological signals [26], but the integration of these modalities remains limited, particularly with respect to the joint modeling of acoustic and muscular activity during piano performance. Moreover, existing approaches have largely focused on emotion recognition from the listener's perspective or on performance analysis, while comparatively less attention has been devoted to investigating the relationship between performers' physiological responses and the expressive content conveyed through music. Motivated by these considerations, this study investigates the complementary contribution of acoustic features and sEMG signals for the analysis of emotional expression in piano performance.

III. BACKGROUND NOTIONS

A. Representation of Emotional Spaces

A central component of MER and MEVD research is the definition of the emotion representation to adopt. This representation directly influences annotation strategies, dataset construction, analysis algorithm design, and evaluation proto-

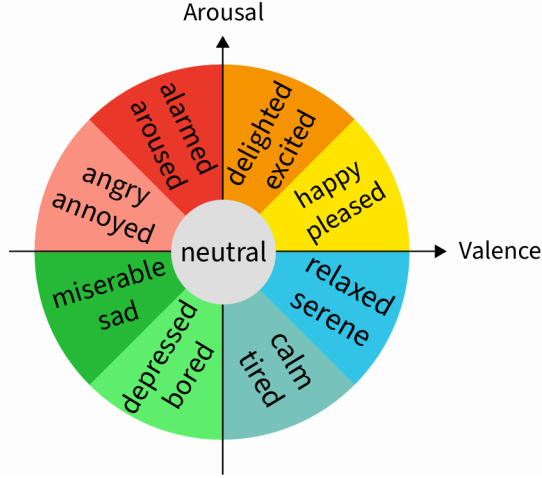


Fig. 1: Emotional classes for the classification task

cols. Two main paradigms dominate the literature: *categorical models* and *dimensional models*.

Categorical models conceptualize emotions as discrete, distinct states. A foundational reference in this domain is Ekman’s model of basic emotions [9], which identifies six primary emotional categories (anger, disgust, fear, happiness, sadness, and surprise). In Music Emotion Recognition, categorical representations remain intuitive and well suited for classification tasks; however, they may impose rigid boundaries on experiences that are inherently continuous and overlapping.

Dimensional models, in contrast, represent emotions within a continuous affective space. The most influential framework is Russell’s Circumplex Model of Affect [22], which organizes emotions along two orthogonal axes: valence (pleasure–displeasure) and arousal (activation–deactivation). Within this representation, emotions are not isolated entities but positions in a continuous space, where nearby emotions convey similar affective meanings [3]. This formulation is particularly suitable when emotional expression cannot be adequately described by a finite set of discrete, linguistically defined categories, as it naturally captures intermediate states and gradual variations in emotional expression.

To combine the intuitive interpretability of dimensional spaces with the formal requirements of classification, we adopt a discretized representation derived from the continuous Circumplex Model. Following the approach of EMOPIA [?], which partitions the valence–arousal plane into four quadrants, we increase spatial granularity by bisecting each quadrant into two sub-regions, yielding eight distinct emotion classes. (Fig. 1). Additionally, a neutral class was incorporated into the proposed schema. This state, which signifies an absence of strong emotional intensity, is conceptualized as a small circular area near the plane’s origin, characterized by low values in both the valence and arousal dimensions. This representation preserves the dimensional structure of the original model while remaining suitable for annotation and classification.

It is worth noting that while the annotation layout is geometrically grounded in the continuous valence–arousal space, the emotion recognition task in this paper is formulated strictly as a categorical classification problem across the predefined spatial regions. Grounding the discrete categories within this underlying continuous space serves two main purposes. First, during data collection, visualizing categories within the two-dimensional plane provided annotators with intuitive semantic relationships and spatial proximity between emotions. Second, this spatial layout establishes a unified annotation framework that naturally supports future extensions toward continuous affect tracking and regression-based modeling.

B. Surface Electromyography (sEMG)

sEMG is a non-invasive technique used to measure the electrical activity generated by skeletal muscles during contraction. Unlike intramuscular electromyography, which requires needle electrodes inserted directly into the muscle tissue, sEMG records muscle activation through electrodes placed on the skin surface.

Muscle contraction originates from the activation of motor units (MUs), where a motor unit consists of a motor neuron and the muscle fibers it innervates. When an action potential reaches the neuromuscular junction, the muscle fibers belonging to the activated motor unit generate electrical impulses known as Motor Unit Action Potentials (MUAPs). Since several motor units are simultaneously recruited during voluntary contractions, the signal detected by a surface electrode corresponds to the superposition of numerous MUAPs propagating beneath the electrode. Consequently, the recorded sEMG signal can be interpreted as the spatial and temporal summation of motor unit activities (Fig. 2), and represented as:

$$x_{EMG}(t) = \sum_{i=1}^M \text{MUAP}_i(t - \tau_i) + n(t), \quad (1)$$

where M denotes the number of active motor units, τ_i represents the activation instant of the i -th motor unit, and $n(t)$ models measurement noise and acquisition artifacts.

From a spectral perspective, the useful information carried by sEMG is mainly concentrated between approximately 20 and 450–500 Hz [18], whereas frequencies below this range are typically dominated by motion artifacts and electrode displacement, and higher frequencies mainly contain electronic noise.

sEMG recordings are commonly acquired using bipolar electrodes. In this configuration, two electrodes are positioned along the muscle fibers while a third reference electrode is placed over an electrically neutral anatomical location (for example clavicles). The recorded signal corresponds to the differential voltage between the two active electrodes, which significantly attenuates common-mode interference, improves the signal-to-noise ratio, and reduces electromagnetic disturbances [8]. This differential acquisition configuration has

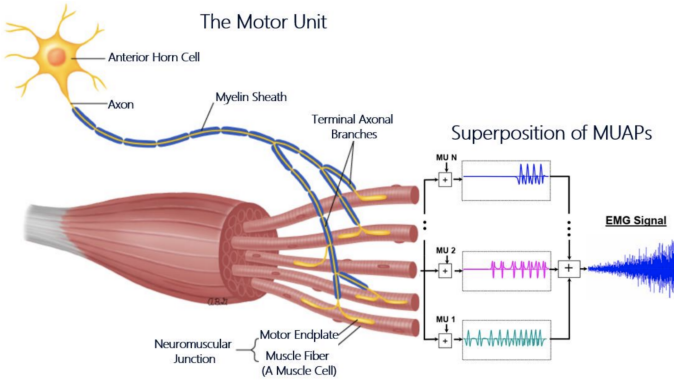


Fig. 2: Single MU and EMG signal as a combination of MUAPs

become the standard solution for physiological monitoring and biomedical signal analysis [2].

Beyond such applications, sEMG has recently attracted increasing attention in affective computing [28]. Emotional states are often accompanied by involuntary muscular responses, including variations in facial expression, neck tension, and shoulder activation. Consequently, sEMG provides direct access to physiological manifestations of emotions. In piano performance, these characteristics are particularly relevant. While the acoustic signal captures the expressive content of the performance, EMG reflects the muscular activity underlying its production, including facial activation, muscle tension, and posture. As these physiological responses often vary with emotional expression, EMG provides complementary information beyond the acoustic domain.

C. Audio encoding through SCHuBERT

The acoustic modality is represented through embeddings extracted using *SCHuBERT*, a real-time end-to-end Piano Music Emotion Recognition (PMER) architecture previously developed by the authors [27]. *SCHuBERT* directly processes the raw audio waveform and learns a high-level latent representation through self-supervised representation learning.

Built upon a pre-trained DistilHuBERT backbone [4] (a lightweight version of HuBERT [12], i.e., a self-supervised neural network for speech representation learning), *SCHuBERT* adapts the model to the specific task of PMER through a fine-tuning process. This adaptation allows the latent representations originally learned from speech to become specialized for expressive piano performances while preserving the general acoustic knowledge acquired during large-scale self-supervised pre-training.

Given an input audio segment $\mathbf{x}_a \in \mathbb{R}^{L_a}$, *SCHuBERT* feeds the waveform directly into DistilHuBERT, producing the sequence of contextual embeddings:

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{T_a}\}, \quad \mathbf{e}_i \in \mathbb{R}^{768}. \quad (2)$$

The contextual embeddings generated at the last Transformer layer are first aggregated through temporal mean pooling,

$$\bar{\mathbf{e}} = \frac{1}{T_a} \sum_{i=1}^{T_a} \mathbf{e}_i. \quad (3)$$

In the original *SCHuBERT* architecture, this pooled representation is then processed by a lightweight classification head specifically introduced for PMER. The classification head consists of a fully-connected layer that learns a task-specific projection of the generic DistilHuBERT representation, followed by a dropout layer for regularization and a final Softmax classifier that predicts the emotional quadrant in the Valence–Arousal space. The entire architecture is fine-tuned on the EMOPIA dataset [13], allowing the dense layer to learn latent representations specifically optimized for piano music emotion recognition.

Formally, the audio embedding generated by *SCHuBERT* can be expressed as:

$$\mathbf{z}_{\text{audio}} = f_{\text{Dense}}(\bar{\mathbf{e}}), \quad \mathbf{z}_{\text{audio}} \in \mathbb{R}^{768}, \quad (4)$$

where $f_{\text{Dense}}(\cdot)$ denotes the linear transformation learned during supervised fine-tuning.

The resulting embedding is subsequently processed by a dropout layer and finally mapped to the four emotion classes through a Softmax classifier:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_c \mathbf{z}_{\text{audio}} + \mathbf{b}_c). \quad (5)$$

IV. PROPOSED MULTIMODAL FRAMEWORK

As illustrated in Fig. 3, the proposed architecture consists of two modality-specific encoding branches followed by an embedding fusion stage and a final emotion classifier. The acoustic branch employs the *SCHuBERT* encoder presented in the previous section to extract 768-dimensional audio embeddings, while the physiological branch uses the proposed sEMG signal encoder (EMGER) to generate compact representations from multichannel EMG recordings. The embeddings produced by the two branches are fused and fed to a multilayer perceptron (MLP), which predicts the target emotional category.

A. *SCHuBERT*: the audio encoder

The audio encoder is directly inherited from *SCHuBERT*. Given an audio segment $\mathbf{x}_a$, the encoder generates the task-specific embedding

$$\mathbf{z}_{\text{audio}} = f_{\text{SCHuBERT}}(\mathbf{x}_a), \quad \mathbf{z}_{\text{audio}} \in \mathbb{R}^{768}, \quad (6)$$

where $f_{\text{SCHuBERT}}(\cdot)$ denotes the fine-tuned *SCHuBERT* model described in Section III-C. In the proposed multimodal framework, the final Softmax classification layer is removed, and the 768-dimensional representation produced by the last dense layer is directly used as the audio embedding.

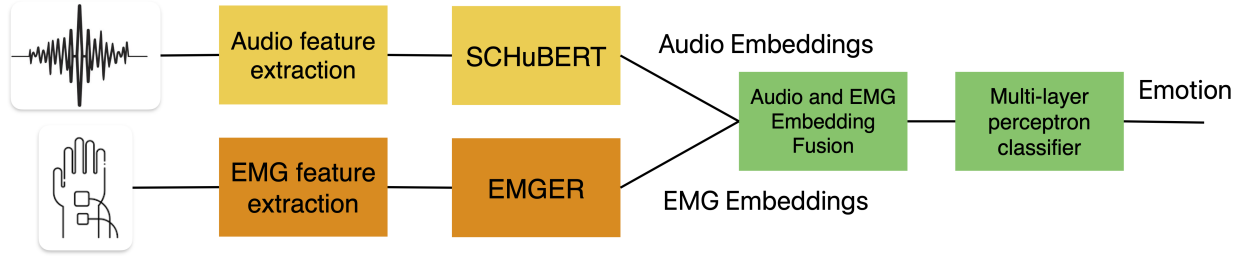


Fig. 3: Overview of the proposed multimodal emotion recognition framework. The acoustic branch employs the SCHuBERT encoder to extract task-specific audio embeddings from raw piano audio, while the physiological branch exploits the proposed EMGER encoder to generate embeddings from multichannel EMG signals. The two representations are subsequently fused and processed by a multilayer perceptron for emotion classification.

B. EMGER: Electromyography signal encoder

To complement the expressive information extracted from the acoustic modality, we introduce *EMGER* (ElectroMyoGraphy Embedding Generator), a neural architecture specifically designed to encode sEMG signals into compact latent representations suitable for multimodal emotion recognition. Unlike conventional EMG classification systems that directly operate on raw physiological signals, EMGER combines lightweight signal processing with deep sequential learning to generate semantically meaningful embeddings describing the muscular activity of the performer. The model exploits eight bipolar sEMG channels.

1) *Signal preprocessing*: Raw EMG recordings are first preprocessed in order to remove acquisition artifacts and standardize the signals before feature extraction. Raw EMG signals are sampled at 750 Hz: $x_m(t)$, $m = 1, \dots, 6$, and subjected to a multistage preprocessing pipeline to reduce noise and artifacts. A fifth-order Butterworth band-pass filter (20–490 Hz) was applied to remove low-frequency motion artifacts and high-frequency noise. A 50 Hz notch filter suppressed power-line interference and its harmonics, while median filtering attenuated transient spikes caused by sensor instability. Finally, peak normalization scaled each channel to the range $[-1, 1]$, ensuring numerical stability and comparability across recordings.

2) *Time-domain representation*: Instead of directly feeding raw EMG samples into the neural network, EMGER represents muscle activity through computationally efficient time-domain descriptors. Let $W_{i,m} = [x_m(iH), \dots, x_m(iH + L_w - 1)]$ be the i -th sliding window of length L_w and hop size H for the m -th EMG channel.

For each window and for each channel, five classical EMG descriptors are computed:

- **Variance (VAR)** quantifies the variability of the EMG signal around its mean within the analysis window. It reflects fluctuations in muscle activation and is often associated with changes in contraction intensity and recruitment of motor units.

$$\text{VAR} = \frac{1}{L_w} \sum_{n=1}^{L_w} (w_n^{(i,m)} - \bar{w}^{(i,m)})^2, \quad (7)$$

- **Integrated EMG (iEMG)** represents the cumulative muscle activity over time, providing an estimate of total muscular effort within the window.

$$\text{iEMG} = \sum_{n=1}^{L_w} |w_n^{(i,m)}|, \quad (8)$$

- **Mean Absolute Value (MAV)** computes the average signal amplitude and serves as a robust indicator of overall muscle activation level, less sensitive to outliers than raw amplitude measures.

$$\text{MAV} = \frac{1}{L_w} \sum_{n=1}^{L_w} |w_n^{(i,m)}|, \quad (9)$$

- **Simple Square Integral (SSI)** emphasizes higher-amplitude signal components by squaring the EMG values, making it particularly sensitive to strong contractions associated with high-intensity emotional states.

$$\text{SSI} = \sum_{n=1}^{L_w} (w_n^{(i,m)})^2, \quad (10)$$

- **Root Mean Square (RMS)** provides an estimate of the effective power of the EMG signal by combining information about both amplitude and energy. It is one of the most widely used EMG descriptors because it is strongly correlated with the level of muscle contraction while remaining relatively robust to signal polarity.

$$\text{RMS} = \sqrt{\frac{1}{L_w} \sum_{n=1}^{L_w} (w_n^{(i,m)})^2}. \quad (11)$$

These descriptors summarize complementary characteristics of muscle activation. Integrated EMG and MAV quantify the overall activation level, VAR measures signal variability, SSI estimates muscular energy, while RMS provides a robust estimate of signal power. Together, these features preserve the temporal evolution of muscular activity while significantly reducing the dimensionality of the original signals.

Since five descriptors are extracted from each of the eight EMG channels, every temporal instant is represented by $8 \times 5 = 40$ features.

Dense Layer
Input shape: (None,19,40) - Output shape: (None, 19, 128)
LSTM Layer
Input shape: (None,19,128) - Output shape: (None, 19, 256)
LSTM Layer
Input shape: (None,19,256) - Output shape: (None, 256)
Batch Normalisation Layer
Input shape: (None,256) - Output shape: (None, 256)
Dropout Layer
Input shape: (None,256) - Output shape: (None, 256)
Dense Layer
Input shape: (None,256) - Output shape: (64)

Fig. 4: EMGER Encoder Schema

Sliding the analysis window across the entire EMG segment produces the feature matrix

$$\mathbf{F}_{\text{EMG}} = [\phi_1, \phi_2, \dots, \phi_K] \in \mathbb{R}^{K \times 40}, \quad (12)$$

where ϕ_i denotes the 40-dimensional feature vector, and K denotes the number of temporal windows. Consequently, the physiological signal is converted into a multivariate temporal sequence suitable for recurrent neural processing.

3) *EMGER architecture*: The architecture of EMGER is illustrated in Fig. 4.

The input feature matrix $\mathbf{F}_{\text{EMG}} \in \mathbb{R}^{K \times 40}$ is first projected into a higher-dimensional latent space through a fully-connected layer

$$\mathbf{H}_{\text{EMG}}^{(0)} = \text{ReLU}(\mathbf{F}_{\text{EMG}} \mathbf{W}_0 + \mathbf{b}_0), \quad (13)$$

producing a sequence of 128-dimensional feature vectors.

Temporal dependencies are then modeled through two stacked LSTM layers.

The first recurrent layer generates

$$\mathbf{H}_{\text{EMG}}^{(1)} = \text{LSTM}_1(\mathbf{H}_{\text{EMG}}^{(0)}) \in \mathbb{R}^{K \times 256}, \quad (14)$$

preserving the complete temporal evolution of the sequence.

The second recurrent layer summarizes the entire sequence into its final hidden state

$$\mathbf{h}_{\text{EMG}} = \text{LSTM}_2(\mathbf{H}_{\text{EMG}}^{(1)}) \in \mathbb{R}^{256}, \quad (15)$$

which represents the temporal dynamics of muscle activation over the whole analysis window.

Batch Normalization and Dropout are subsequently employed to improve optimization stability and reduce overfitting.

Finally, a dense projection layer maps the recurrent representation into a compact latent space

$$\mathbf{z}_{\text{EMG}} = \text{ReLU}(\mathbf{h}_{\text{EMG}} \mathbf{W}_e + \mathbf{b}_e), \quad \mathbf{z}_{\text{EMG}} \in \mathbb{R}^{64}. \quad (16)$$

This 64-dimensional vector constitutes the EMG embedding generated by EMGER.

In the original standalone architecture, the embedding is processed by a final Softmax classifier

$$\hat{y} = \text{Softmax}(\mathbf{W}_c \mathbf{z}_{\text{EMG}} + \mathbf{b}_c), \quad (17)$$

to predict the emotional category.

However, similarly to SCHuBERT, the classifier is discarded in the proposed multimodal framework. Instead, only the latent representation $\mathbf{z}_{\text{EMG}}$ is retained and employed as the physiological embedding describing the muscular activity associated with each performance segment. These embeddings provide a compact representation of performer physiology while preserving temporal dynamics that are highly informative for emotion recognition.

C. Multimodal Fusion Architecture

The final stage integrates the audio and physiological representations into a unified multimodal embedding. Let $\mathbf{z}_{\text{audio}}$ denote the 768-dimensional audio embedding for a given segment, and $\mathbf{z}_{\text{EMG}}$ denote the corresponding 64-dimensional EMG embedding. These vectors are concatenated to form a joint representation capturing both acoustic intent and embodied response,

$$\mathbf{z}_{\text{fusion}} = [\mathbf{z}_{\text{audio}}; \mathbf{z}_{\text{EMG}}] \in \mathbb{R}^{832}. \quad (18)$$

The concatenated vector is passed to a two-layer perceptron incorporating Batch Normalization and Dropout regularization to improve convergence stability and mitigate overfitting. The final dense layer applies a Softmax activation function to produce probability estimates across eight emotion classes:

$$\hat{y} = \text{Softmax}(f_{\text{MLP}}(\mathbf{z}_{\text{fusion}})). \quad (19)$$

V. DATA ACQUISITION PROTOCOL

A. Musical Pieces

The experimental protocol involved eight professional pianists and required them to perform four musical pieces each, thus resulting a total of 32 music pieces. Three of these pieces are common to all pianists and were selected from the standard piano repertoire because they are technically straightforward. This choice allows performers to focus primarily on emotional expression rather than technical accuracy. Furthermore, the selected pieces exhibit distinct expressive characteristics, enabling the observation of different emotional responses during performance:

- *R. Schumann, Kinderszenen, Op. 15, No. 1*. It conveys a gentle sense of nostalgia and introspection, often associated with memories of past experiences.
- *W.A. Mozart, Piano Sonata No. 16 in C Major: Allegro*. It is characterized by clarity, lightness, and playful energy. These qualities tend to evoke feelings of joy and positive engagement.
- *Schubert, Moments Musicaux, Op. 94, No. 3*. It expresses a sense of serenity, calmness, and acceptance. This latter

emotion can be viewed positively, almost comparable to happiness and pleasantness.

Additionally, pianists were asked to perform a fourth piece of their own choice. Each piece has an average duration of approximately 3-5 minutes.

B. Data Acquisition setup

In the following we describe the acquisition protocols of musical and sEMG data adopted in our study.

1) *Music data acquisition:* A digital piano (Roland FP-30X digital piano) was used to capture both audio and MIDI data during the performance. MIDI data was recorded directly from the piano, whereas audio was acquired using a Focusrite Scarlett 4i4 Gen 1 audio interface as float32 format, with a sampling rate set at 44.1 kHz. MIDI messages were saved in two formats: a single-track MIDI file and a CSV file that encapsulates the details of each message such as timestamps together with the other MIDI parameters.

2) *sEMG data acquisition:* sEMG signals were acquired using a dedicated electronic setup consisting of surface electrodes connected to an LWT³¹ eight-channel EMG acquisition system, which digitized the signals and transmitted them via USB to a computer. Data acquisition, visualization, and storage were performed using *RawPower*, the proprietary software developed by LWT³. The recorded signals were stored in CSV format, with each column corresponding to an acquired channel and associated timestamps. The acquisition sampling rate was set to 750 Hz.

The electrode placement targeted four muscles involved in body posture and facial expression during emotional performance, selected according to the work by Barbero [2] on EMG-based emotion recognition and refined through consultations with professional pianists. The monitored muscles were the *Zygomaticus Major*, *Sternocleidomastoid*, *Upper Trapezius*, and *Erector Spinae*. Each muscle was instrumented bilaterally (left and right sides of the body) using a bipolar electrode configuration, with two electrodes placed on each muscle following standard sEMG acquisition guidelines. Consequently, the acquisition setup comprised 16 electrodes in total (4 muscles $\times$ 2 body sides $\times$ 2 electrodes), corresponding to the eight differential channels provided by the acquisition board, each channel measuring the potential difference between the two electrodes placed on a single muscle.

Additionally, two reference electrodes were positioned on the clavicles to provide a common reference for the measurements. These electrodes were used exclusively for signal referencing and were not recorded as acquisition channels.

To ensure signal stability and minimize motion artifacts during piano performance, specialized garments and mounting accessories designed by LWT³ were employed. In particular, a dedicated vest securely housed the EMG acquisition board while facilitating stable electrode placement and reducing cable movement, thereby improving recording quality and performer comfort throughout the experimental sessions.

3) *Data synchronization:* Prior to beginning their performance, the pianists were instructed to remain still for a few seconds. This precaution facilitates a clean and precise cut of the recordings, removing any unnecessary movements made before the start of execution. Subsequently, the pianists were asked to play a single note while simultaneously performing a deliberate, unnatural gesture, enabling temporal synchronization between the audio signal recorded from the keyboard and the EMG data. For instance, the pianist may clench their teeth at the exact moment of the key press, and then start playing the musical piece.

C. Musicians Preparation

Before recordings, participants were informed about the whole procedure and provided their informed consent. After preparation, participants were given some time to get used to the equipment. A verification phase followed, during which the quality of the acquired signals was checked. This included verifying that EMG signals captured by each channel were stable, without any interference or artifact causes by misplaced electrodes.

During the performance, both data modalities were synchronously recorded. The protocol required each piece to be performed multiple times (at least two) to observe possible changes due to familiarity or fatigue. Subsequently, the recording selected for inclusion in the dataset was identified through a two-stage evaluation process. First, the pianists were asked to provide subjective feedback on the performance, reporting whether they felt emotionally engaged and fully concentrated during the execution. This self-assessment was then complemented by a technical evaluation performed by the researchers, who inspected the acquired data to verify signal quality, the absence of significant artifacts, and the occurrence of as few performance errors as possible. The recording that best satisfied both the subjective and technical criteria was ultimately selected for the investigation of the performer's emotional engagement.

D. Data Annotation

For annotation purposes, performances were annotated directly by the performers themselves. This design choice was adopted to ensure that the annotations reflected the performers' own internal emotional experiences during the execution (felt emotion), rather than emotions inferred by external observers (perceived emotion).

Due to the experimental workflow, the annotation phase took place asynchronously in the days following the recording sessions. Following data acquisition, researchers first performed technical validation and track selection to upload the optimal performances onto a dedicated web-based platform. Once the platform was populated, participants were granted secure access using credentials to access a personalized environment displaying only their own selected recordings, ensuring controlled and uninfluenced labeling. Self-annotations were completed within a few days post-performance, depending on each pianist's availability.

¹<https://lwt3.com>

Although retrospective self-reporting may inherently introduce memory decay, this limitation is significantly mitigated by the musicians’ deep familiarity with the repertoire. Since all selected pieces had been performed extensively by the pianists across diverse contexts prior to the study, the performers possessed an acute awareness of their precise emotional states associated with each musical section. Furthermore, allowing pianists to review their performances via the interactive web player facilitated strong contextual recall, balancing accurate self-assessment with the preservation of ecological validity during the live recordings.

The web platform—implemented with an HTML/JavaScript front end, a Java EE Servlet back end, and a MySQL database—provided a dedicated annotation interface featuring an audio player with an interactive timeline. Through this interface, participants could review their recordings and assign, modify, or remove emotion labels at precise temporal locations. The available labels corresponded to the nine emotional classes defined in the adopted emotional space (Section III). To facilitate the task, the interface displayed a graphical representation of the continuous emotional space. Although the model utilizes discrete categorical labels, their metric arrangement provides an intuitive visualization of emotional proximity and semantics, aiding annotators in accurately capturing subtle affective transitions compared to conventional categorical lists.

Upon completion of the annotation phase, all labels were exported in JSON format structured hierarchically, associating each pianist with their recordings and timestamped emotion annotations (expressed in milliseconds).

VI. NUMERICAL ASSESSMENT

To ensure a fair comparison between the proposed multimodal (MM) framework and the EMGER baseline, the same modality-specific encoders are employed throughout the experimental pipeline. The EMGER architecture is first trained independently on the sEMG data following the procedure described in Section IV-B. The resulting network is directly used to evaluate the standalone EMGER model, whose test performance serves as the unimodal baseline. Subsequently, the final Softmax classification layer is removed, and the remaining part of the network is retained as the EMG encoder, generating the 64-dimensional physiological embedding z_{EMG} . An analogous procedure is adopted for the acoustic modality. The SCHuBERT model, previously fine-tuned for PMER, is used as the audio encoder by removing its final classification layer and retaining the 768-dimensional embedding z_{audio} introduced in Section IV-A. During the training of the MM architecture, both modality-specific encoders remain frozen, while only the MM fusion module and the final MLP classifier are optimized. This training protocol ensures that the comparison between the standalone EMGER model and the proposed MM framework isolates the contribution of MM fusion, so that any performance improvement can be attributed to the integration of complementary physiological and acoustic embeddings rather than to additional optimization of the individual encoders. Furthermore, because SCHuBERT

was originally evaluated on a distinct dataset and task [27], it is employed here strictly as a frozen, high-level acoustic feature extractor.

A. Training Configuration

Data preprocessing and feature extraction followed the procedure described in Section V for both the standalone EMGER model and the proposed MM framework.

Given an audio recording and the corresponding multichannel sEMG signals, the two modalities were synchronously segmented using a sliding window of duration $L_s = 3000$ ms and hop size $H_s = 500$ ms, generating paired audio and physiological segments $(x_a^{(i)}, x_{\text{EMG}}^{(i)})$, with $i = 1, \dots, N$.² For each physiological segment $x_{\text{EMG}}^{(i)}$, the feature extraction procedure described in Section IV-B was applied using analysis windows of length $L_w = 300$ ms and hop size $H_w = 150$ ms, producing the temporal feature matrix $F_{\text{EMG}}^{(i)} \in \mathbb{R}^{K \times 40}$, where K denotes the number of overlapping sub-windows within the segment. This hierarchical windowing strategy captures short-term muscle activation dynamics while improving feature robustness by aggregating statistical descriptors instead of directly processing the raw sEMG signals.

To ensure reliable supervision, a segment was retained only if it was entirely contained within a single annotated emotional interval. Consequently, each retained training sample $(x_a^{(i)}, F_{\text{EMG}}^{(i)}, y^{(i)})$ was associated with a unique emotion label $y^{(i)}$, while segments spanning multiple annotated intervals were discarded. This conservative labeling strategy reduces annotation ambiguity and improves the robustness of the learning process.

The segmentation procedure was applied independently to every recording, yielding a total of $N = 15\,793$ valid samples across all pianists. Specifically, the dataset comprises 2,127 segments from the Schumann performances, 3,799 from Schubert, 4,458 from Mozart, and 5,409 from the participants’ selected repertoire. The different numbers of samples reflect the duration of the corresponding recordings while preserving all valid emotion-labeled segments.

The segmentation procedure was first applied independently to each musical piece, producing four disjoint datasets corresponding to the Schumann, Schubert, Mozart, and participant-selected repertoires. Each dataset was then independently partitioned into training (70%), validation (15%), and test (15%) subsets using a two-stage random split, in which the test set was first extracted and the validation set was subsequently obtained from the remaining samples. Finally, the training, validation, and test partitions obtained from the four repertoires were merged, yielding the final training, validation, and test sets used for model development. Consequently, each subset contains samples from all musical pieces while preserving the original proportion of samples within each repertoire.

Both the standalone EMGER model and the proposed MM architecture were trained using the same AdamW optimizer

²Preliminary experiments considered a shorter window ($L_s = 500$ ms), which proved insufficient to capture stable muscle activation patterns and resulted in lower validation accuracy.

TABLE I: Comparison between EMG-based model and MM System performances

	Model	Train Acc	Val Acc	Test Acc
Schumann	EMGER system	0.858	0.810	0.761
	MM system	0.864	0.820	0.779
Mozart	EMGER system	0.875	0.815	0.823
	MM system	0.886	0.817	0.831
Schubert	EMGER system	0.832	0.795	0.772
	MM system	0.841	0.803	0.779
Selected piece	EMGER system	0.887	0.851	0.829
	MM system	0.898	0.863	0.831

with an initial learning rate of 10^{-4} and a batch size of 32. The learning rate was reduced by a factor of 0.5 whenever the validation loss did not improve for 10 consecutive epochs, while early stopping was triggered after 20 epochs without improvement. The model parameters corresponding to the minimum validation loss were retained for the final evaluation.

B. Results and Interpretation

Table I summarizes validation and test accuracies per piece for both EMGER and MM models, computed along the entirety of each musical piece.

Across all musical contexts, MM fusion consistently improved validation accuracy by approximately 1–2%. The participant-selected repertoire exhibited the most substantial improvement, with validation accuracy rising from 0.851 to 0.863 and test accuracy from 0.829 to 0.831.

Training accuracy trends further illustrate the models’ behavior. For example, in the participant-selected piece, the EMG-only model achieved 0.887 training accuracy, while the MM model reached 0.898. Despite similar training performance, validation accuracy improved more significantly in the MM case, indicating better generalization.

Convergence dynamics also differed between models. As shown in Fig. 5, the MM system typically reached optimal validation performance within approximately 40 epochs, whereas EMG-only models required between 60 and 130 epochs depending on piece duration and emotional complexity. This suggests that audio embeddings provide complementary contextual constraints that accelerate optimization and reduce ambiguity in physiological signal interpretation.

Validation curves further indicated improved stability in the MM configuration. Oscillatory behavior observed in EMG-only training was reduced when audio embeddings were incorporated. The generalization gap between training and validation accuracy was also slightly smaller in the MM system, reflecting improved robustness against overfitting.

Performance variability across pieces reflects emotional structure. Schumann and Schubert works, dominated by low-arousal emotional states, exhibited lower absolute accuracy compared to the more emotionally dynamic participant-selected repertoire. Mozart’s Allegro achieved relatively high baseline performance due to clearer high-arousal cues.

Importantly, MM gains were observed consistently across all pieces, including those with strong class imbalance. This demonstrates that fusion does not merely compensate for

specific weaknesses but provides structurally complementary information.

Overall, these results validate the central premise of this work: musical emotion emerges from the interaction between acoustic expressivity and embodied physiological response. By integrating EMG-derived muscle activation patterns with deep audio embeddings, the proposed framework achieves higher predictive accuracy, faster convergence, improved stability, and enhanced generalization. These findings strongly motivate future extensions incorporating motion capture and facial expression analysis, potentially enabling real-time MM emotion tracking in live concert scenarios.

VII. CONCLUSION

This study demonstrated the effectiveness of a multimodal framework for MER, EMG and audio data to capture both physiological and acoustic dimensions of musical expressivity. By combining an LSTM-based EMG model with a pre-trained MER audio model, the proposed system achieved improved accuracy and robustness compared to unimodal approaches, confirming the value of multimodal integration in emotion analysis.

The experiments, conducted on a controlled dataset mimicking live concert conditions, suggest that multimodal learning provides a richer and more reliable representation of performers’ emotional states. Although emotional subjectivity and the retrospective nature of self-reporting remain inherent challenges, the results highlight the scientific relevance of combining bodily and auditory cues in musical emotion analysis.

Future work should address current limitations by expanding the dataset with additional participants and longer performances, incorporating synchronized video to reduce memory recall biases during annotation, and conducting formal statistical significance testing alongside isolated audio-only benchmarks. In line with the Internet of Sounds and Internet of Musical Things paradigms, these developments lay the foundation for low-latency, real-time emotion tracking systems capable of driving interactive stage environments during live performances.

REFERENCES

- [1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392, 2017.
- [2] Marco Barbero, Roberto Merletti, and Alberto Rainoldi. *Atlas of muscle innervation zones: understanding surface electromyography and its applications*. Springer Science Business Media, 2012.
- [3] M. Buccoli, M. Zanoni, G. Fazekas, A. Sarti, and M. Sandler. A higher-dimensional expansion of affective norms for english terms for music tagging. pages 316–322, 2016.
- [4] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091, 2022.
- [5] Changfeng Chen and Qiang Li. A multimodal music emotion classification method based on multifeature combined network classifier. *Mathematical Problems in Engineering*, 2020(1):4606027, 2020.
- [6] Wenwen Chen et al. A novel long short-term memory network model for multimodal music emotion analysis in affective computing. *Journal of Applied Science and Engineering*, 26(3):367–376, 2022.

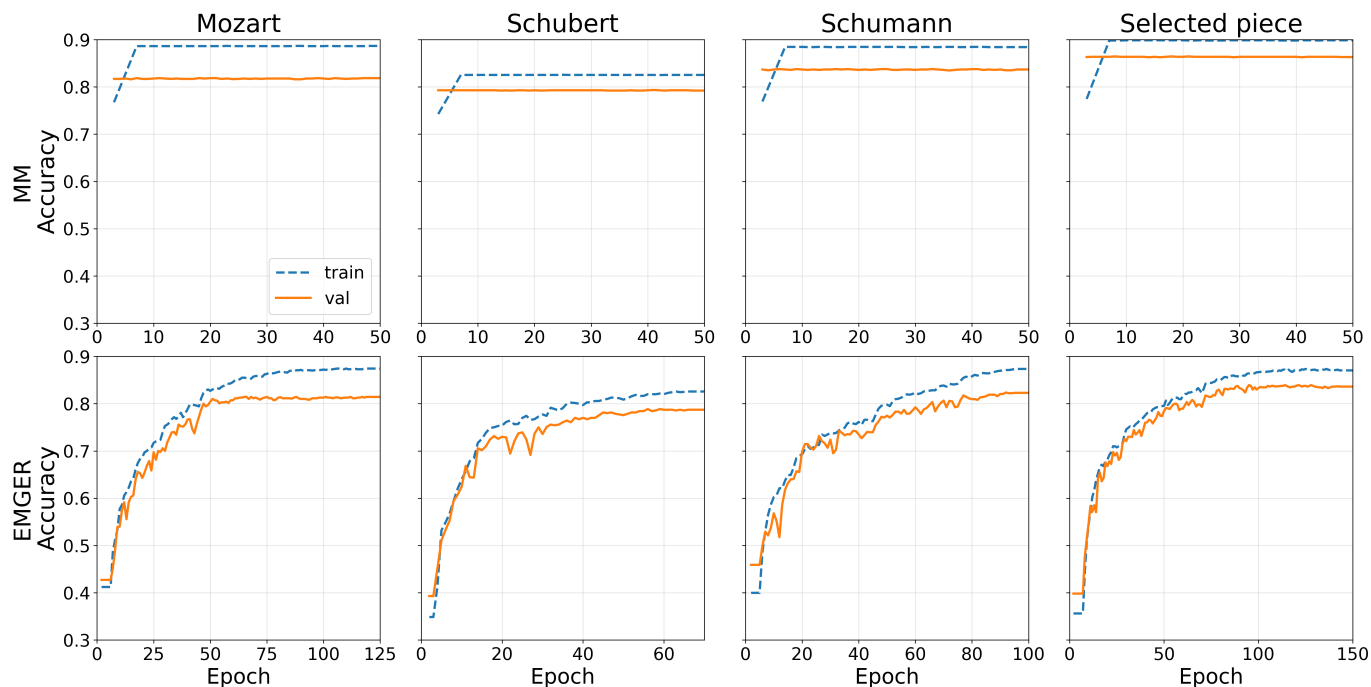


Fig. 5: Training and validation accuracy curves obtained for the four experimental settings. The top row reports the MM models, while the bottom row reports the EMGER models.

- [7] Rubana H. Chowdhury, Mamun B.I. Reaz, Mohd Alauddin Bin Mohd Ali, Ashrif A.A. Bakar, Kalaivani Chellappan, and Tae G. Chang. Surface electromyography signal processing and classification techniques. *Sensors*, 13(9):12431–12466, 2013.
- [8] Edward A. Clancy, Evelyn L. Morin, Gelareh Hajian, and Roberto Merletti. Surface electromyogram amplitude estimation: Best practices. *Journal of Electromyography and Kinesiology*, 72, 2023.
- [9] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [10] Jennifer Haase and Sebastian Pokutta. Chapter 16 - human-ai cocreativity: Exploring synergies across levels of creative collaboration. In Matthew J. Worwood and James C. Kaufman, editors, *Generative Artificial Intelligence and Creativity*, Explorations in Creativity Research, pages 205–221. Academic Press, 2026.
- [11] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):166335, 2022.
- [12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.
- [13] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, pages 318–325, 2021.
- [14] Xingguo Jiang, Yuchao Zhang, Guojun Lin, and Ling Yu. Music emotion recognition based on deep learning: A review. *IEEE Access*, 12:157716–157745, 2024.
- [15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [16] Ting Li. Music emotion recognition using deep convolutional neural networks. *Journal of Computational Methods in Science and Engineering*, 24(4-5):3063–3078, 2024.
- [17] Davide Lionetti, Luca Turchet, Massimiliano Zanoni, and Paolo Belluco. Muscle-guided guitar pedalboard: Exploring interaction strategies through surface electromyography and deep learning. In S M Astrid Bin and Courtney N. Reed, editors, *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 241–251, Utrecht, Netherlands, September 2024.
- [18] Roberto Merletti and Silvia Muceli. Tutorial. surface emg detection in space and time: Best practices. *Journal of Electromyography and Kinesiology*, 49:102363, 2019.
- [19] Nikolaos Partarakis and Xenophon Zabulis. A review of immersive technologies, knowledge representation, and ai for human-centered digital experiences. *Electronics*, 13(2), 2024.
- [20] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.
- [21] Yi Qiu, Yu Lin, and Yun Lin. Music audio emotion regression using the fusion of convolutional neural networks and bidirectional long short-term memory models. *PeerJ Computer Science*, 11:e3086, 2025.
- [22] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.
- [23] Shubham Sharma, Shubhankar Verma, Mohit Kumar, and Lavanya Sharma. Use of motion capture in 3d animation: motion capture systems, challenges, and recent trends. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 289–294. IEEE, 2019.
- [24] Marc R Thompson, Juan Ignacio Mendoza, Geoff Luck, and Jonna K Vuoskoski. Relationships between audio and movement features, and perceived emotions in musical performance. *Music & Science*, 6, 2023.
- [25] Luca Turchet, Mathieu Lagrange, Cristina Rottondi, György Fazekas, Nils Peters, Jan Østergaard, Frederic Font, Tom Bäckström, and Carlo Fischione. The internet of sounds: Convergent trends, insights, and future directions. *IEEE Internet of Things Journal*, 10(13):11264–11292, 2023.
- [26] Luca Turchet, Barry O’Sullivan, Rupert Ortner, and Christoph Guger. Emotion recognition of playing musicians from eeg, ecg, and acoustic signals. *IEEE Transactions on Human-Machine Systems*, 54(5):619–629, 2024.
- [27] Massimiliano Zanoni, Riccardo Rossi, Alice Sironi, Paolo Belluco, and Cristina Rottondi. Schubert: a real-time end-to-end model for piano music emotion recognition. In *6th International Conference on Audio for Virtual and Augmented Reality and Immersive Games*. AES, 2026.
- [28] Mingde Zheng, Michael S. Crouch, and Michael S. Eggleston. Surface electromyography as a natural human-machine interface: A review. *IEEE Sensors Journal*, 22(10):9198–9214, 2022.