UNIVERSIDAD
NACIONAL
DE COLOMBIA

Brain Music: Sistema compositivo, gráfico y sonoro creado a partir del comportamiento frecuencial de las señales cerebrales

Brain Music: Generative system for symbolic music creation from affective neural responses

Hernán Darío Pérez Nastar

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería eléctrica, electrónica y computación
Manizales, Colombia
21 - Nov - 2023

Brain Music: Sistema compositivo, gráfico y sonoro creado a partir del comportamiento frecuencial de las señales cerebrales

Hernán Darío Pérez Nastar

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magister en Ingeniería de automatizcíon industrial

Director:
Ph.D. Germán Castellanos Dominguez
Co-Director:
Ph.D. Andres Marino Alvarez Meza

Línea de Investigación:
Investigación en Aprendizaje Profundo y señales Biológicas
Grupo de Investigación:
Grupo de Control y procesamiento digital de señales

Universidad Nacional de Colombia
Facutad de Ingeniería y arqitectura, Departamento ingeniería electrica y electrónica
Manizales, Colombia
2023

Acknowledgements

I extend my heartfelt gratitude to my supervisor, Cesar Germán Castellanos, for his invaluable insights and unwavering support throughout this work. His patience and guidance have been crucial during these years, and I am truly grateful for their contributions. Additionally, I would like to thank my co-supervisor, Andres Marino Alvarez, whose constant help and direction have played a pivotal role in bringing this thesis to fruition.

I am also indebted to the Signal Processing and Recognition Research Group at Universidad Nacional de Colombia Sede Manizales for their generous support during my studies. Their assistance has been instrumental in my academic journey.

Lastly, I wish to express my deep appreciation to my parents and all the individuals who stood by me during this challenging process. Their encouragement and belief in me during difficult times made all the difference.

I am truly grateful for the financial support provided by the program "PROTOTIPO FUNCIONAL DE LENGUA ELECTRÓNICA PARA IDENTIFICACIÓN DE SABORES EN CACAO FINO DE ORIGEN COLOMBIANO - Convoctoria 2020-0740", funded by Minciencias, which enabled me to pursue my master studies.

Resumen

Esta tesis de maestría presenta una metodología de aprendizaje profundo multimodal innovadora que fusiona un modelo de clasificación de emociones con un generador musical, con el propósito de crear música a partir de señales de electroencefalografía, profundizando así en la interconexión entre emociones y música. Los resultados alcanzan tres objetivos específicos:
Primero, ya que el rendimiento de los sistemas interfaz cerebro-computadora varía considerablemente entre diferentes sujetos, se introduce un enfoque basado en la transferencia de conocimiento entre sujetos para mejorar el rendimiento de individuos con dificultades en sistemas de interfaz cerebro-computadora basados en el paradigma de imaginación motora. Este enfoque combina datos de EEG etiquetados con datos estructurados, como cuestionarios psicológicos, mediante un método de "Kernel Matching CKA". Utilizamos una red neuronal profunda (Deep&Wide) para la clasificación de la imaginación motora. Los resultados destacan su potencial para mejorar las habilidades motoras en interfaces cerebro-computadora.
Segundo, proponemos una técnica innovadora llamada "Labeled Correlation Alignment"(LCA) para sonificar respuestas neurales a estímulos representados en datos no estructurados, como música afectiva. Esto genera características musicales basadas en la actividad cerebral inducida por las emociones. LCA aborda la variabilidad entre sujetos y dentro de sujetos mediante el análisis de correlación, lo que permite la creación de envolventes acústicos y la distinción entre diferente información sonora. Esto convierte a LCA en una herramienta prometedora para interpretar la actividad neuronal y su reacción a estímulos auditivos.
Finalmente, en otro capítulo, desarrollamos una metodología de aprendizaje profundo de extremo a extremo para generar contenido musical MIDI (datos simbólicos) a partir de señales de actividad cerebral inducidas por música con etiquetas afectivas. Esta metodología abarca el preprocesamiento de datos, el entrenamiento de modelos de extracción de características y un proceso de emparejamiento de características mediante Deep Centered Kernel Alignment, lo que permite la generación de música a partir de señales EEG.
En conjunto, estos logros representan avances significativos en la comprensión de la relación entre emociones y música, así como en la aplicación de la inteligencia artificial en la generación musical a partir de señales cerebrales. Ofrecen nuevas perspectivas y herramientas para la creación musical y la investigación en neurociencia emocional. Para llevar a cabo nuestros experimentos, utilizamos bases de datos públicas como **GigaScience, Affective Music Listening y Deap Dataset**.

Palabras clave: aprendizaje profundo, señales EEG, clasificación de emociones, generación de música, interfaz cerebro-computadora (BCI), aprendizaje multimodal, generación de música simbólica, piano roll, inteligencia artificial

Abstract

This master's thesis presents an innovative multimodal deep learning methodology that combines

an emotion classification model with a music generator, aimed at creating music from electroencephalography (EEG) signals, thus delving into the interplay between emotions and music. The results achieve three specific objectives:

First, since the performance of brain-computer interface systems varies significantly among different subjects, an approach based on knowledge transfer among subjects is introduced to enhance the performance of individuals facing challenges in motor imagery-based brain-computer interface systems. This approach combines labeled EEG data with structured information, such as psychological questionnaires, through a "Kernel Matching CKA"method. We employ a deep neural network (Deep&Wide) for motor imagery classification. The results underscore its potential to enhance motor skills in brain-computer interfaces.

Second, we propose an innovative technique called "Labeled Correlation Alignment"(LCA) to sonify neural responses to stimuli represented in unstructured data, such as affective music. This generates musical features based on emotion-induced brain activity. LCA addresses variability among subjects and within subjects through correlation analysis, enabling the creation of acoustic envelopes and the distinction of different sound information. This makes LCA a promising tool for interpreting neural activity and its response to auditory stimuli.

Finally, in another chapter, we develop an end-to-end deep learning methodology for generating MIDI music content (symbolic data) from EEG signals induced by affectively labeled music. This methodology encompasses data preprocessing, feature extraction model training, and a feature matching process using Deep Centered Kernel Alignment, enabling music generation from EEG signals.

Together, these achievements represent significant advances in understanding the relationship between emotions and music, as well as in the application of artificial intelligence in musical generation from brain signals. They offer new perspectives and tools for musical creation and research in emotional neuroscience. To conduct our experiments, we utilized public databases such as **GigaScience, Affective Music Listening and Deap Dataset**.

Keywords: Deep Learning, EEG Signals, Emotion Classification, Music Generation, Brain-Computer Interface (BCI), Multimodal Learning, Symbolic Music Generation, Piano Roll, Artificial Intelligence

Contenido

1 Preliminaries

1.1. Motivation

In recent years, the field of Machine Learning (ML) has witnessed remarkable advancements in multimodal tasks. Neural network architectures are now being applied to tasks that extend beyond single modalities. For example, language and vision are integrated into tasks such as visual question answering [10], commonsense reasoning [170], dialogue [40], and phrase grounding [121]. Audio signal processing also has advanced in speech recognition and visual speech synthesis [115]. In addition to the uses listed above, multimodality can be used in brain-computer interface (BCI) systems. BCI systems use electrical signals from the brain to control external devices, such as wheelchairs, prosthetics, and computers [117]. Multimodality can improve the accuracy and reliability of BCI systems by combining signals from different sources, such as Electroencephalography(EEG), Electromyography(EMG), and Functional Magnetic Resonance Imaging(fMRI). This can allow people with disabilities to control external devices more naturally and intuitively. [54]

Emotion recognition (or detection) stands as a significant scientific problem in Affective Computing, an emerging research field proposed by Picart [120], aiming to empower computer systems with the capacity to accurately process, recognize, and comprehend emotional information expressed by humans for natural human-computer interactions (HCI) [119]. This paradigm plays a crucial role in both Artificial Intelligence and Ambient Intelligence [88], attracting researchers from diverse disciplines, such as Computer Science, Electronic Engineering, Human Factors Engineering, Psychology, Neuroscience, and Medical Science, among others, to collaborate on understanding and advancing the capabilities of Affective Computing. Concurrently, another influential paradigm in the field of BCI is Motor imagery (MI). It involves the cognitive process of mentally generating quasi-perceptual experiences without external stimuli [81]. In practice, MI has been used as a therapy that contributes to children's motor learning and improves the motor skills of children with motor problems [69, 16], evaluates screen-time and cognitive development [144], and aids in attentional focus and rehabilitation [15, 140, 134]. BCI systems frequently employ electroencephalography (EEG) to decode MI-based brain signals due to its non-invasiveness, portability, relatively low cost, and high temporal resolution [135]. By merging these two paradigms, researchers from various domains can further explore the potential of multimodal approaches and deep learning techniques to foster advancements in both Affective Computing and BCI research, leading to novel applications in artificial intelligence and human-computer interaction research [1].

On the other hand, music synthesis, known for its capacity to create original audio signals by embedding representations and extracting informative properties from diverse and intricate data, has found extensive applications. These applications range from fostering artistic innovation to producing adaptable and copyright-free music for use in games and videos [157]. In the realm of acoustic representations derived from a variety of audio sources, this serves as the fundamental basis for music generation [17].

In the domain of music, a related area involves working with symbolic music representations, such as MIDI (Musical Instrument Digital Interface). MIDI representations encapsulate critical musical details, including tempo, pitch, dynamics, and duration [103]. These details offer a wealth of data for in-depth analysis of musical content. While audio and symbolic music generation tasks are distinct, it's worth noting that the underlying architectures of deep learning and encoding techniques share remarkable similarities [21].

However, some approaches to automatic musical composition, such as rule-based methods [147], address the complexities of musical perception and involve the segregation of complex compositional structures such as melody, harmony, rhythm, timbre, and even styles. Taking advantage of improved perception capabilities regarding unstructured data (Multimedia: Images, audio and video)[79], machine learning (ML) models fed with raw data (without any prepossessing) in the time domain have shown great promise in sound generation, where architectures are tightly coupled to audio representations. [20]. Nonetheless, when learning musical features from arbitrary corpora, the challenge lies in adapting ideas and patterns borrowed from diverse contexts to achieve a specific objective. This style of learning poses several issues for ML architectures, including capturing short and long-term music structures, conducting low-level and high-level analysis such as onset/offset detection, rhythm estimation, harmonic analysis, instrument detection, structural segmentation, genre, and mood classification, developing models with inherent reasoning to reduce training data requirements, and promoting transparent and objective evaluation methodologies [62]. Through the integration of music generation and emotion recognition paradigms, researchers can harness the potential of 'multimodal' approaches, which involve the simultaneous use of multiple types of data or sensory information. By doing so and employing advanced deep learning techniques, they open up exciting opportunities for the development of interactive applications that are emotionally enriched. These applications span diverse fields, including virtual reality, entertainment, and assistive technology, and they are poised to provide novel, immersive experiences.

In the fields mentioned above, data representation varies as follows:

1.) EEG: The EEG signal directly represents brain activity and is a valuable tool in studying human brain physiological phenomena. Its primary characteristics include being typically noisy and susceptible to environmental interference, as it often mixes with other signals like EOG, ECG, and EMG, along with various artifacts and noises [148]. EEG signals can be categorized as spontaneous or evoked, with spontaneous EEG being the rhythmic potential fluctuation generated by the nervous system without external stimuli and evoked potentials referring to measurable changes in the cerebral cortex caused by external sensory stimulation. EEG signals are highly nonlinear

due to human tissue adaptation and physiological regulation. Additionally, these signals are unstable, influenced by external environmental factors, and exhibit strong non-stationarity, leading researchers to employ statistical analytic approaches to discover and recognize their features. In cognition, the most relevant frequency range for EEG signals is 0.5-30 Hz, which researchers often decompose into five sub-bands corresponding to distinct cognitive functions [64] 2.) Audio: Audio signals have been extensively studied, and their representations depend on specific applications. In Music, raw audio can be dealt with, but due to its high density (up to 44.1 kHz), subsampling or feature extraction is often employed to represent audio more manageable way [136]. Music Information Retrieval, investigates audio features that can be useful in diverse musical applications. 3.) Symbolic Music: MIDI is one of the most common symbolic music representations, popular for connecting musical technology devices such as synthesizers, Digital Audio Workstations (DAWs), and Virtual Studio Technology (VSTs). For machine learning applications, the representation of Piano Roll is also prevalent, consisting of a matrix where columns represent time and rows represent pitch [71].

In the local context, the Signal Processing and Digital Signal Group (GCPDS) at the National University of Colombia has been actively engaged in the development of machine learning algorithms focusing on brain-computer interfaces, computer-assisted diagnosis systems for neurological disorders, and neuro-feedback systems, among others. These research projects have been conducted with various regional stakeholders, including the Transmedia Research Center of the University of Caldas. Joint research and development initiatives, such as the project titled "Brain Music: Prototipo de interfaz interactiva para generación de piezas musicales basado respuestas eléctricas cerebrales y técnicas de composición atonal" have paved the way for the current investigation. Additionally, GCPDS has collaborated with other partners on various projects, fostering a dynamic research environment:

- "Prototipo de interfaz cerebro-computador de bajo costo para la detección de patrones relevantes de actividad eléctrica cerebral relacionados con TDAH"(2021-now)
- "Herramienta de apoyo a la predicción de los efectos de anestésicos locales vía neuroaxial epidural a partir de termografía por infrarrojo"(2020-now)
- "Desarrollo de un sistema automático de análisis de volumetría cerebral como apoyo en la evaluación clínica de recién nacidos con asfixia perinatal"(2019-now)
- Çaracterización morfológica de estructuras cerebrales por técnicas de imagen para el tratamiento mediante implantación quirúrgica de neuroestimuladores en la enfermedad de Parkinson"
  (2019-now)
- "Herramienta de apoyo al diagnóstico del TDAH en niños a partir de múltiples características de actividad eléctrica cerebral desde registros EEG"(2019-now)

1.2. Problem statement

Dealing with EEG signals, that measure the brain's electrical activity captured from the scalp, is challenging due to their non-stationarity, low signal-to-noise ratio, and complexity, which necessitate extensive preprocessing and feature extraction approaches for accurate analysis. Consequently, Deep Learning has gained popularity as it has shown great promise in leveraging the characteristics of EEG signals, allowing it to learn relevant features from raw data [108] autonomously. Similarly, applications related to music information retrieval face difficulties in handling audio signals due to their complex structures and the diverse characteristics required for different tasks [38].

Regarding Deep Learning, the primary objectives in Motor Imagery and Emotion recognition are to achieve accurate predictions while maintaining model interpretability. Conversely, music content generation focuses on generating high-quality data that preserves the training data's structures and exhibits a certain level of creativity [61].

A multimodal approach helps address general challenges in EEG such as: low interpratility, inter-subject variability (the models are not generalizable to different subjects), among others... [42] and music generation tasks, specifically the problem of few data, as there is no extensive multimodal EEG and Music database. For this reason, finding the right databases, such as GigaScience, BCMI-MIdAS and DEAP, is crucial to allow the correlation of these two fields of research. During the development of this work, the following specific difficulties were encountered: 1.) Multimodality of EEG with structured data and transfer learning. 2) Labeled alignment for sonification based on affective neural responses. 3.) Symbolic music generation from affective neural EEG signals.

1.2.1. Problem 1: EEG + Structured Data

In practice, MI capability can be assessed to determine the extent to which a user engages in a mental representation of movements, while on the other hand the performance of a machine learning model in MI is determined by its ability to effectively predict the mental state of the user (if they are thinking about the movement of a hand, the tongue, a foot...). This EEG paradigm should collect information from users primarily through self-report questionnaires [96]. Using information from these questionnaires is believed to be useful in helping prediction models improve their performance. However, very little evidence shows a secure correlation between classification accuracy and questionnaire scores. Several reasons may explain this respect [167, 128]: weak and ambiguous self-interpretation in understanding questionnaire instructions, laboratory paradigms restricted to a narrow class of motor activity, time limitations that guarantee consistent mental states, difficulty in learning characteristics of subjects with illiteracy (refers to subjects who cannot learn this task despite completing many sessions) BCI, among others. Therefore, although psychological assessment and questionnaires are probably the most accepted and validated methods in medical contexts [149], their inclusion in automated prediction of BCI skills remains very rare due to their questionable reliability and reproducibility. [33].

To improve the predictive utility, the joint analysis of different imaging modalities is achieved that

can explain the relationships discovered between the anatomical, functional and electrophysiological properties of the brain [91, 34]. However, in addition to the problems that may arise from questionnaire implementation, multimodal analysis research efforts pose a challenging problem in terms of combining categorical data with imaging measurements, facing the following restrictions [50, 37]: Different spatial and temporal sampling rates, non-instantaneous and non-linear coupling, low signal-to-noise ratio, lack of interpretable results, optimal combination of individual modalities still undetermined, and effective dimensionality reduction to improve discriminability of multiple features extracted views [164].

Another improving approach to BCI skills is to perform several training sessions where participants learn how to modulate their sensorimotor rhythms appropriately, relying on the spatial specificity of MI-induced brain plasticity [180]. However, collecting extensive data is time-consuming and mentally exhausting during a prolonged recording session, deteriorating the measurement quality.

1.2.2. Problem 2: EEG + unstructured Data in affective paradigm

Sound synthesis is often based on statistical distributions inferred from training acoustic data or complementary multimedia information sources depending on different applications, for example: speech synthesis (emulating a speaker's voice) [143], processing multimodal audiovisual and multi-instrumental configuration [48], text and symbolic transcriptions [106]. Specifically, different complementary data are provided according to the mentioned applications to train the deep learning models and architectures [109]. For music synthesis applications, the most common deep learning method requires a large amount of audio data to be added to the input set (a large training set), which is taken from vast online music repositories. Nevertheless, conditioning strategies for low-level music synthesis may include non-acoustic data used to create audible sounds (also known as sonification [49]), such as speech, images, text, and videos. Moreover, sonification can be used for more unique sources, such as non-empty objects containing fluids [162], mode vibrations of protein and amino acid building blocks [169], and the silent nature of flames [98]. Other sources are biosignals captured from the human body, including electromyography [100] and electrocardiographic data [127]. Even so, electroencephalography (EEG) signals reflect emotions more accurately in real time than other peripheral neurophysiological data. It also offers more reliable data acquisition hardware with increasing affordability. For example, EEG-based affective brain-computer interfaces have attracted interest in developing music creation systems [93]. However, the estimation accuracy of induced affective states using EEG signals might need to be improved for applying conditioning to ML architectures [139]. Often, the modeling of emotions needs to be more consistent and is strongly context-dependent [160], not to mention that the brain processes involved in the induction and mediation of affective states by emotionally evocative stimuli are poorly understood due to the difficulty [7].

Various feature sets such as mel frequency cepstral coefficients (MFCCs) [56], Log-mel spectrum, Mel filter bank [43], Constant-Q spectrum are used for feature extraction from audit data [124], inspired by the human audit system and physiological findings or also modern approaches in which

features are extracted by neural networks and raw input data (raw Audio) is processed. These features offer a broad set of possibilities for automatic descriptions of musical signals [123], taking advantage of the ability to extract acoustic descriptors over a wide dynamic range. Along with acoustic characteristics, when it comes to tasks related to music, embeddings and symbolic representations are found for machine learning [90], one of the most used symbolic representations is the MIDI format. Regarding obtaining EEG parameters, there are several limitations. Firstly, the mechanisms that evoke emotions are related to sound perception and are incredibly subjective (information focus, cultural impact, musical structure orientation) [72, 63], however studies such as citeloui2021neuroscience, katthi2021deep show us that there is a correlation. On the other hand, the EEG often contains significant artifacts unrelated to the presented stimulus and caused by other cognitive tasks or reference noise [163]. Because of this, there are no standard methods for extracting features from EEG data within machine learning frameworks dedicated to EEG sonification in recent years [92].
Another issue is integrating data from multiple heterogeneous sensors into a low-dimensional representation, learning the joint temporally modulated dependencies from both modalities (audio and EEG) that are assumed to be mutually correlated [45], study like [**?**] shows several experiments using EEG recordings from subjects listening to speech and music stimuli. In these experiments, they found that the deep models improve the Pearson correlation significantly over the linear methods (average absolute improvements of 7.4 % in speech tasks and 29.3 % in music tasks) they also analyzed the impact of several model parameters on the stimulus-response correlation. Feature reduction and selection are conducted as a first step to handle the large dimensionality of the extracted characteristics and increase their interpretability [171].

1.2.3. Multimodal: EEG + Symbolic Music and generative model

The field of computer-Based Music Systems is not new at all, however there is some ambiguity in what it encompasses [22]. It can include, from the first generation of a computer melody in 1957 by Newman Guttman to digital sound synthesis in 1983 by the Yamaha DX7. However, the greatest ambiguity is that it can have two different objectives: to design and construct autonomous music-making systems [28] or to design and construct computer-based environments to assist human musicians [114]. In recent years, the application of deep learning and machine learning techniques in the generation of musical content has gained prominence. These methods offer a distinct advantage in terms of generality compared to traditional, handcrafted models like grammar-based [142] or rule-based music generation systems [51]. Deep learning models can be trained on diverse musical corpora, allowing them to learn and adapt to various musical genres.
The versatility of machine learning-based generation systems is especially apparent when large-scale musical datasets become available [53]. They can automatically learn the intricate nuances of different musical styles from such datasets and generate new musical content. This capability is particularly useful in scenarios where the complexity of the desired application surpasses the boundaries of analytical formulations or manual design. [65]

Symbolic models, such as rule-based systems and grammars, are utilized in music to represent harmony, melody, rhythm, structure, and form. In contrast, sub-symbolic models, like machine learning algorithms embeddings, are employed to learn musical characteristics from a collection of musical pieces automatically (Sub-symbolic models use numerical values, vectors, or continuous data representations. They do not rely on explicit symbols or rules to represent information). These models offer a generative and interactive approach to aid musicians in crafting new music by leveraging their enhanced ïntelligent"memory, encompassing associative, inductive, and generative capabilities. This has become achievable due to the increasing availability of music in various forms, like sound, scores, and MIDI files, which computers can process. However, integrating sub-symbolic techniques, such as deep learning, with symbolic techniques, including constraints and reasoning, still needs to be solved. Researchers are actively exploring ways to unite the strengths of these approaches to develop more sophisticated and comprehensive music generation systems, enhancing the creative potential for musicians [22].

Articles like [67, 175, 105] are examples among many that One of the approaches in Music Generation Systems are those related to musical generation and emotions. However, the complexity of human emotions, particularly in the context of music [24, 161], has led researchers to propose various definitions for the fundamental types of emotions. For instance, in 1972, Ekman identified six basic emotions based on the analysis of facial expressions [95]. In 1980, Russell introduced the emotional circumplex Model, **1-1** to map common emotions and investigate their correlations [129], In Russel's circumplex Model, emotions are represented in a two-dimensional space where the axes correspond to the levels of arousal and valence. Arousal, on one axis, reflects the intensity of emotions, spanning from .^{ex}cited"to çalm."Valence, on the perpendicular axis, signifies the evaluation of emotions, ranging from "depressed"to "serene."

For instance, consider "happy.^{a}nd .^{a}ngry.^{em}otions; they can both exhibit a high level of arousal, but they are associated with opposite valence values, resulting in distinct emotional outcomes. This model also supports the concept of an emotion palette, which includes fundamental emotions like happiness, sadness, anger, and fear [116]. This two-dimensional plane model has been extensively utilized to analyze and quantify emotions, leading to its application in diverse fields to explore the interplay of emotions in different contexts. In 2007, Gomez et al. examined the relationship between two-dimensional emotional plans and musical characteristics and found that in distinguishing between negative and positive valence, the key discriminators were mode, harmonic complexity, and rhythmic articulation. On the other hand, when distinguishing between high and low arousal, the most influential factors were tempo, accentuation and rhythmic articulation. Interestingly, tempo, stress, and rhythmic articulation were also the features that showed the strongest correlations with physiological measures [58].

Affective computing, a subset of artificial intelligence, revolves around detecting, processing, interpreting, and emulating human emotions [158]. The burgeoning development of portable, non-invasive human sensor technologies, including brain-computer interfaces (BCI), has sparked significant interest among scholars across various disciplines in emotion recognition. The widespread availability of electronic devices has recently led to increased engagement in social media, online

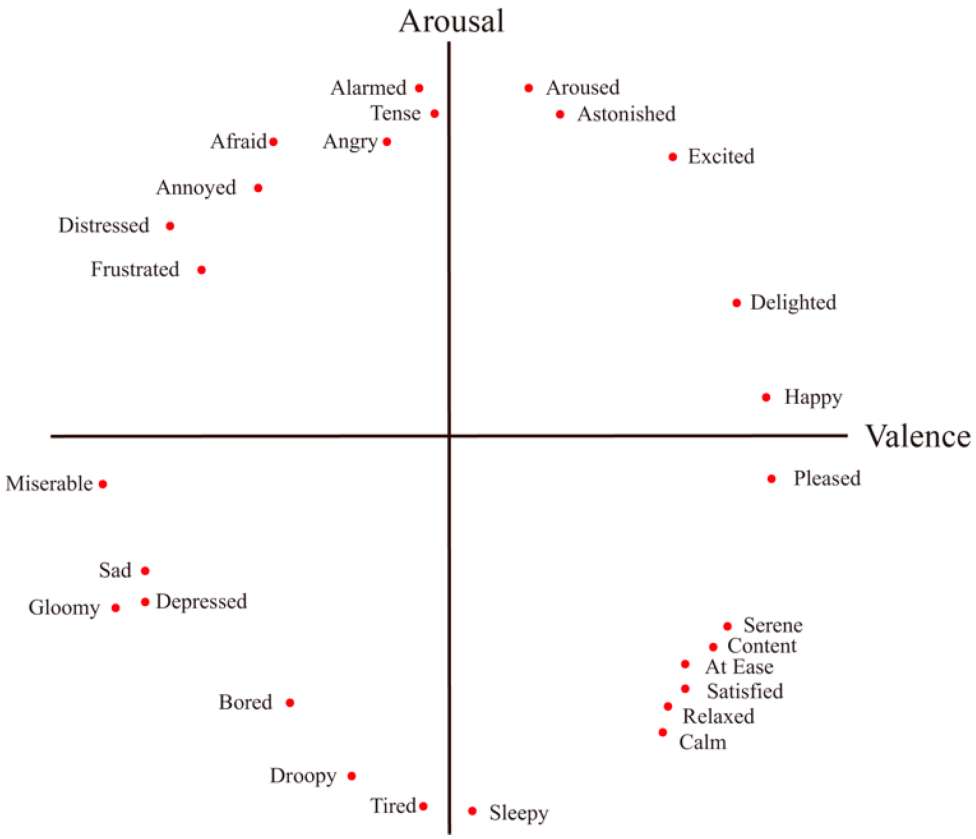


Figura 1-1: Circumplex model proposed by Russel

gaming, e-commerce, and other digital activities [64]. However, most modern human-computer interaction (HCI) systems need to grasp and comprehend emotional data, needing more emotional intelligence and the ability to recognize human emotions to inform decision-making and action. BCIs, as portable non-invasive sensor technologies, capture brain signals and harness them as inputs for systems aiming to humanize HCI [60]. Notably, EEG signals generated by the central nervous system exhibit rapid responses to emotional changes compared to other peripheral neural signals and have demonstrated their importance in emotional recognition [89].

Finding a suitable representation to match EEG signals and music holds great utility [35], as EEG-based emotion recognition research has gained significant popularity across multiple disciplines in recent decades [3]. Although the available scientific data on emotional states and their structure remain limited, researchers have established a strong correlation between EEG activity and music-induced emotions. Certain music can alter neural activity and trigger emotional responses in individuals. Vuilleumier and Trost [153] have demonstrated that emotion recognition in music occurs rapidly. They suggest that these emotions come to the forefront through a combination of activation in emotional and motivational brain systems (including reward pathways), which confer music with its valence. In addition, several other brain regions beyond emotional systems, encompassing those related to motor functions, attention, and memory, also exert their influence. As the

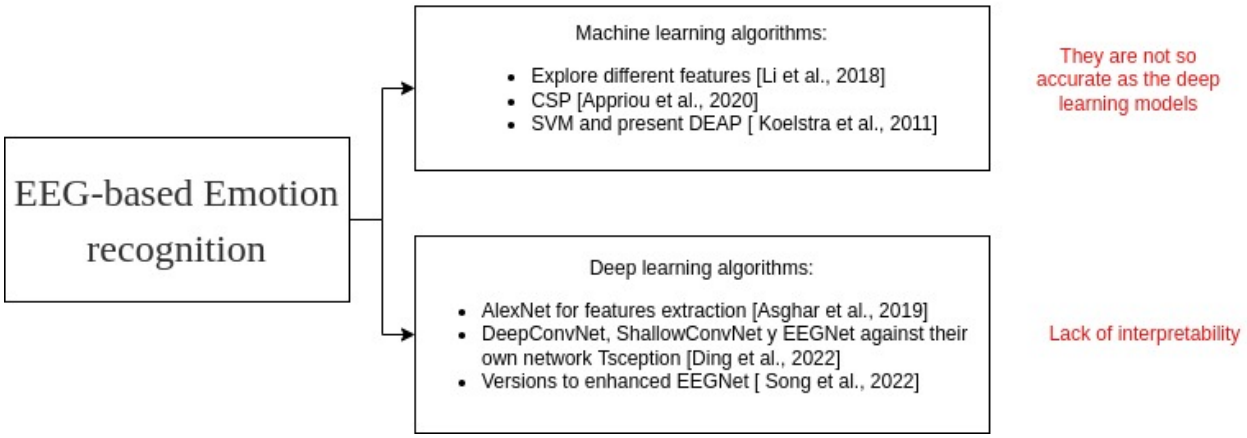
discussion unfolds, they delve into the neural underpinnings responsible for orchestrating the synchronization of cognitive and motor processes by music, elucidating how these mechanisms relate to one's affective experience.
Moreover, Nordstrom and Laukka [111] discovered that emotion recognition accuracy improved as gate duration increased, reaching stability after a certain point. For various emotions, above-chance accuracy was achieved with very short stimuli, suggesting that emotion recognition is a rapid process based on low-level physical features, indicating that emotional responses are multidimensional. This understanding emphasizes the complexity of emotional experiences tied to music. Functional magnetic resonance imaging (fMRI) studies conducted by Bodner and Shaw [19] further highlight the influence of music on brain activity. Listening to Mozart's music resulted in the expected temporal lobe activation and frontal lobe engagement, accompanied by significant $\alpha - wave$ changes. This phenomenon may be attributed to the highly organized nature of Mozart's music, with its regular repetition of melody aligning with the rhythmic cycle of brain electricity, impacting the human body.
Furthermore, the reliability of EEG-based emotion recognition is emphasized by its ability to capture electrical signals produced by neurons, which humans cannot intentionally control. In contrast, other methods based on body movement, posture, voice, and expression can be deliberately manipulated, leading to potential inaccuracies in emotion recognition [3]. With its inherent capability to capture genuine physiological signals, the EEG approach offers a more dependable means of understanding and interpreting emotional responses to music.
Research Question:
How can a **multimodal methodology** be employed to generate symbolic music compositions by leveraging **EEG-based emotion recognition** as a foundational element to ensure emotionally resonant music?

1.3. Literature review

1.3.1. EEG Emotion Recognition review:

The task of recognizing emotions based on EEG signals has been approached through various methodologies1.3.1, including the utilization of artificial intelligence (AI) and machine learning techniques [151]. Presented below is a concise overview of notable advancements in this field[36]: One approach that has been explored involves the application of machine learning algorithms to classify emotions by utilizing features extracted from EEG signals. In a study by Li et al. [87], different EEG characteristics were examined to enhance the classification of emotions among distinct subjects. Another conventional method employed is Common Spatial Patterns (CSP), as discussed by Appriou et al. [11], which aims to classify cognitive and affective states based on EEG signals. Notably, Koelstra et al. [78] contributed significantly by introducing the DEAP database, this dataset offers a multi-modal approach for studying human affective states. It includes electroencephalogram (EEG) and peripheral physiological data from 32 participants who watched 40

one-minute music video excerpts. Participants provided ratings on various emotional dimensions, including arousal, valence, liking, disliking, dominance, and familiarity. In addition, for 22 of these participants, frontal face video data was also recorded during the experiment, they achieved high classification accuracy using Support Vector Machines (SVM) and traditional features.
Deep Learning techniques have gained prominence in emotion recognition utilizing EEG signals. For instance, Asgar et al. [12] employed an extensive neural network, such as AlexNet, to extract features and classify emotions using SVM, they achieves better classification accuracy compared to recently reported work when validated on SJTU SEED and DEAP datasets. Alternatively, Ding et al. [47] developed a network inspired by established EEG processing networks, including DeepConvNet, ShallowConvNet, and EEGNet, and evaluated its performance in emotion recognition datasets DEAP and MAHNOB-HCI. Moreover, Song et al. [138] proposed an enhanced version of EEGNet, termed LSDD-EEGNet, a novel framework for depression detection. They combine CNN and LSTM for feature extraction and utilize a domain discriminator to align data representation spaces. Their approach outperforms traditional machine learning methods and deep learning models, particularly in subject-independent evaluation. This suggests that LSDD-EEGNet holds promise as a robust method for detecting depression.



Review Emotion recognition

1.3.2. Transfer Learning:

To harness the advantages of transfer learning in the examination of EEG signals, it is imperative to employ tactics that align with individual variances and diminish data prerequisites. These tactics facilitate fine-tuning the model specifically for the intended subject [173]. For example, in the study conducted by Kant et al. [73], they employ pre-existing models like VGG16 and Alex-net as a foundation for model adaptation. This technique reduces the amount of training data required for the MI classification task. In this particular scenario, the EEG signals are transformed into equivalent image representations using continuous wavelet transform, which are then subjected to deep network training. Similarly, Zhang et al. propose five distinct approaches to adapt an EEG-BCI system based on deep convolutional neural networks for MI decoding in their research [172].

Each approach fine-tunes a pre-existing model that has undergone extensive training, adapting it to enhance the performance assessment for a specific subject of interest. More recently, researchers have explored methodologies centered on weighted instances [159] and domain adaptation [178]. In the context of weighted instances, transfer learning is employed to select source domain data that closely resembles the target domain, aiding in the training of the classification model for the target domain. In the second scenario, deep transfer learning techniques are expanded to encompass multi-subject training in EEG, where the objective is to align the feature distributions from individual feature extractors in an MI-based BCI system using maximum-mean discrepancy. Nonetheless, there exist two primary constraints that must be addressed in order to extract sets of shared features among subjects with analogous distributions: the availability of small-scale datasets and the notable signal variances across subjects [172]. Achieving sufficient consistency in the feature space and probability distribution of the training and test data, while averting negative transfer effects, remains a formidable challenge. This encompasses issues such as effective feature extraction from multimodal data capable of distinguishing between MI tasks, selection of transferable entities and transferability evaluations, and the assignment of appropriate weights for transfer learning [145].

1.3.3. Feature Alignment Strategies:

Regarding the association between music stimuli and evoked neural responses, two distinct approaches are utilized for music generation assessment: a) A regression-based approach that directly predicts a real-valued correlation between the coupled sets. b) A recognition-based approach that couples the feature modalities through a standard set of categorical labels, indirectly evaluating the relationship by analyzing the contribution learned by each training feature assemblage to classifier performance [113].

Up to this point, numerous correlation-based techniques that analyze music using EEG data have been documented. These include Canonical Correlation Analysis (CCA), which transforms two sets in a way that maximizes their estimated correlation [25]. There have also been advancements in CCA-variant techniques [181], Multifractal Detrended Cross-Correlation Analysis [122], and coupled Nonnegative Tensor Decomposition [131]. Additionally, several machine learning (ML) approaches have emerged. For instance, deep CCA is capable of inferring the optimal feature mapping [8], and architectures based on Convolutional Neural Networks (CNNs) have been developed to calculate space similarity [27], among other methods. However, the effectiveness of these aforementioned feature alignment strategies is compromised when the training data contains noise or exhibits high variability [66]. EEG recordings often have a poor signal-to-noise ratio due to weak signals mixed with intrinsic noise that has a much higher amplitude than that produced by biological sources. As a result, there is intra-subject and inter-subject variability. Consequently, feature extraction and alignment strategies necessitate multiple repetitions across numerous runs and trials. However, stimulus-response paradigms typically have limited auditory datasets per individual due to participant fatigue, which presents a challenge when attempting to enhance feature alignment

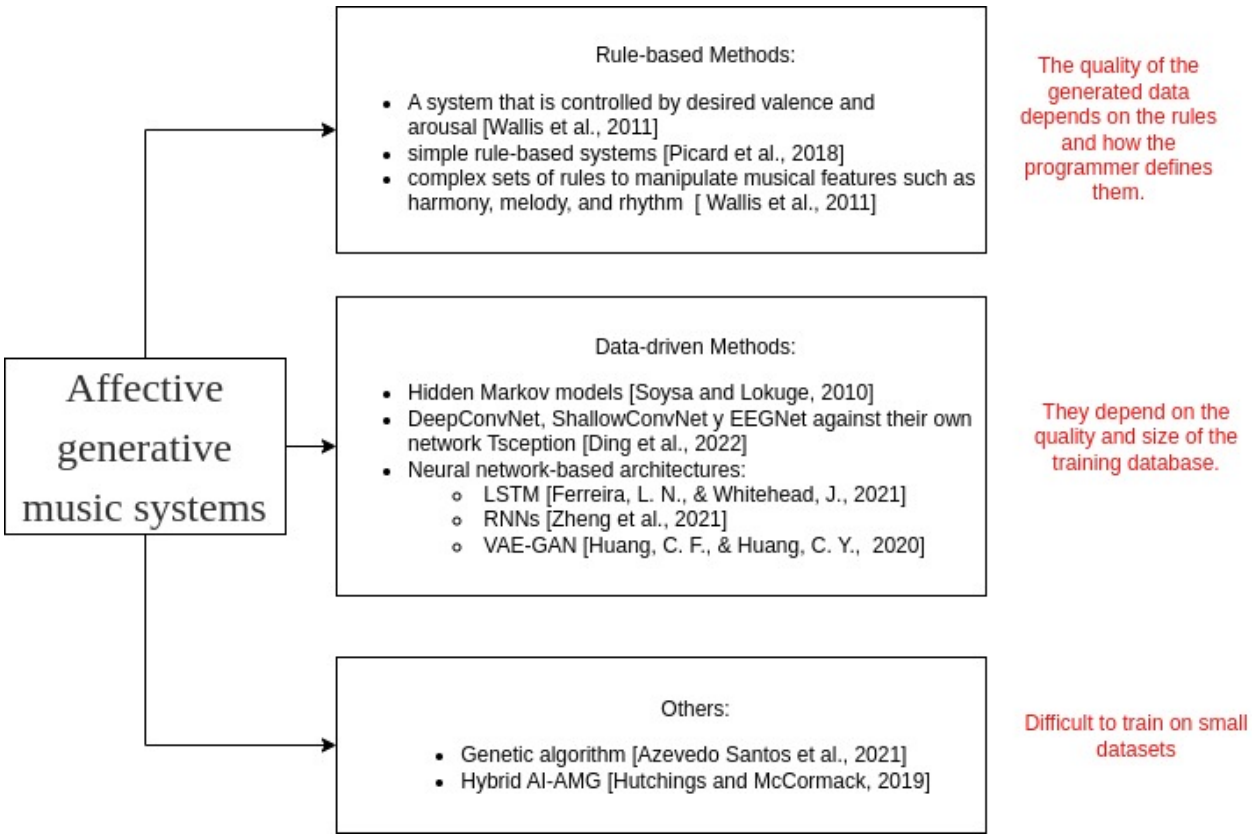
strategies for measuring the similarity between elicited audio stimuli and evoked EEG responses.

1.3.4. Affective generative music systems:

Affective generative music systems have been explored using various methods and techniques to create music that elicits specific emotional responses1.3.4. Dash and Agres [42] provide a comprehensive review of AI-based affective music generation systems, discussing different methods and challenges in this domain.
Rule-based Methods: Rule-based methods establish relationships between musical features and emotion dimensions. Wallis et al. [154] presented a rule-based generative music system controlled by desired valence and arousal, using an emotion map to depict the feature-to-emotion category of rules. Picard et al. [118] developed simple rule-based systems that select the tempo and mode of affective compositions based on target emotion information. More complex rules have also been employed to manipulate musical features such as harmony, melody, and rhythm.
Data-driven Methods: Data-driven methods utilize music databases to train models for generating affective music. Hidden Markov models (HMMs) have been employed to select chord progressions for composing affective music [141]. Neural network-based architectures have also been utilized, including Long Short-Term Memory (LSTM) networks [52], Recurrent Neural Networks (RNNs) [177], and Variational Auto Encoder-Generative Adversarial Network (VAE-GAN) [65]. These models capture the emotional context and generate music accordingly.
Other Approaches: Genetic algorithms have been used to pose affective music generation as an optimization problem, aiming to find optimal sets of musical events or generative musical rules[13]. Hybrid AI-AMG systems combine neural network architectures with rule-based methods, leveraging context and emotion information to manipulate musical features such as harmony, melody, and rhythm, often applied to pre-composed melodic themes to create affective music [68].

1.3.5. Multimodal neural networks in music and EEG:

Researchers have been exploring the potential of multimodal neural networks in various domains. For instance, Vishesh et al. [152] developed a novel approach called DeepTunes, which utilizes deep learning techniques to generate music based on facial expressions. They leveraged facial emotion data to synthesize music that resonates with different emotional states, for the facial recognition model, a Convolutional Neural Network (CNN) has been implemented, GPT-2 has been used to generate lyrics and stacked LSTM networks have been constructed to generate music, the model produced music for each emotion class, and the resulting piano compositions were assessed through a questionnaire. The audience provided feedback based on the emotions they experienced when listening to the generated music. Similarly, He et al. [59] proposed a multimodal multitask neural network for motor imagery classification using EEG and fNIRS signals. Their model demonstrated promising results in classifying motor imagery tasks by integrating information from both EEG and fNIRS modalities. Moreover, Miyamoto et al. [102] claimed to focus on music gene-



Affective generative music systems [42]

ration using EEG data. They introduced an online EEG-based emotion prediction system designed to forecast users' affective states and generate music accordingly. It's worth noting that the specific details of their validation process are not clearly specified in publicly available information.

1.4. Aims:

1.4.1. General aims

This master's thesis aims to design and develop a novel deep-learning methodology that integrates an emotion classification model with a musical generation model. The proposed model aims to generate musical content directly from EEG signals, to enhance the understanding of the relationship between emotions and music.

1.4.2. specific aims

- Introduces a deep learning model for improving the poor-performing individuals in MI-based BCI systems, using the advantages of the multimodality between Electroencephalography data and structured data.

- Proposes a method for sonifying neural responses to labeled affective music listening using auditory and electroencephalographic features that are maximally congruent with the brain activity signal, the nonsymbolic data (audio), and the label set.

- Develop a deep learning model to generate symbolic music (MIDI) content directly from brain activity signals elicited by emotions and evaluate the music quality of the generation.

1.4.3. Outline and contributions

- Chapter 2 propose a parameter-based approach for cross-subject transfer learning to improve the performance of individuals with inefficiency in motor imagery-based brain-computer interface (BCI) systems. The approach involves using kernel embedding to pool data from labeled EEG measurements and psychological questionnaires. A Deep&Wide neural network for MI classification is implemented to pre-train the network from the source domain. The layer parameters are then transferred to initialize the target network through a fine-tuning procedure, recomputing the accuracy. Data fusion combines categorical and real-valued features through stepwise kernel matching via Gaussian embedding. To evaluate the approach, paired source-target sets are selected based on inefficiency-based clustering by subjects, exploring two strategies for choosing the best-performing subjects from the source space: single-subject and multiple-subjects. Validation results for discriminant MI tasks demonstrate that the introduced Deep&Wide neural network shows competitive accuracy even after including questionnaire data, showcasing its potential for improving BCI motor skills.

- In the 3th chapter, we propose an innovative approach called Labeled Correlation Alignment (LCA) to sonify neural responses from affective music listening data, aiming to generate low-level music based on brain activity elicited by emotions. To address inter/intra-subject variability, a combination of PhaseLockingValue and Gaussian Functional Connectivity is employed. The two-step LCA approach involves Centered Kernel Alignment and canonical correlation analysis to select multimodal representations with higher relationships between auditory and neural features. LCA enables a backward transformation to estimate the matching contribution of each extracted brain neural feature set, facilitating a physiological explanation of the generated music. Validation results demonstrate the effectiveness of the LCA approach in creating acoustic envelopes and distinguishing between acoustic and acoustic outputs, showcasing its potential for sound synthesis applications in various artistic and multimedia contexts.

- In this chapter4, we propose an end-to-end deep learning methodology for generating symbolic music content based on brain activity signals elicited by emotions. The methodology comprises three main stages: data preprocessing of the input signals and training of two deep learning models, namely EEGNet and Autoencoder PianoRoll, for feature extraction from EEG and symbolic music, respectively. Subsequently, a Pearson-based feature-matching pro-

cess was conducted to combine the extracted characteristics, ultimately enabling the generation of content through the electroencephalogram signals.

1.4.4. thesis structure

This thesis is structured as follows: In Chapter 2, we introduce the "Deep&Wide neural network: EEG and Psychological Questionnaires,"where we explore how transfer learning enhances performance in the Motor Imagery protocol. Particularly in this chapter, we establish the connection between EEG signals and Psychological Questionnaires, representing this work's initial foray into multimodality. Here, we relate unstructured data (EEG) with tabular data (Questionnaires). In Chapter 3, titled "Sonified through Labeled Correlation Alignment,"we continue with the multimodal approach by linking brain activity with audio features. The goal is to sonify neural responses to labeled affective music listening using auditory and electroencephalographic features strongly congruent with the label set. In Chapter 4, "Symbolic music (Piano Roll) and EEG Alignment,"we delve into multimodal end-to-end deep learning methodology, where both symbolic audio and EEG features are calculated using neural networks, followed by their matching and music generation. Ultimately, in Chapter 5, we offer our conclusions, delineate potential directions for future research, and enumerate the academic outputs linked to this thesis.

2 Deep&Wide neural network: EEG and Psychological Questionnaires

This results where published in the paper "Deep and wide transfer learning with kernel matching for pooling data from electroencephalography and psychological questionnaires-[33]

2.1. Materials and Methods

2.1.1. Dataset: GigaScience

This dataset [1] contains EEG data recorded during a Brain-Computer Interface (BCI) experimental paradigm of Motor Imagery (MI) movements from fifty-two subjects (although only data from fifty subjects are available). The data were acquired using a 10-placement C-electrode system with 64 channels and a sampling rate of 512 Hz. Each subject performed 100 trials lasting 7 seconds while imagining either left or right-hand movements. The MI paradigm started with a fixation cross presented on a black screen for 2 seconds. Next, a cue instruction appeared randomly on the screen for 3 seconds, prompting the subjects to imagine moving their fingers from the forefinger to the little finger, touching each to their thumb. A blank screen was then shown for a break period, randomly between 4,1 and 4,8 s. During one testing session, MI tasks were repeated 20 times.
Subjective answers to a psychological and physiological questionnaire were collected by *GigaScience* a to investigate performance variations and strategies for subject-to-subject transfer in response to intersubject variability. Subjects filled out the questionnaire three times during the MI paradigm timeline: before the experiment began (answering 15 questions), after each run within the experiment (answering ten questions), and at the end of the experiment (answering four questions).

2.1.2. Representation of EEG data using 2D features

We construct a single matrix from the EEG database obtained through a C-channel montage. This matrix represents the n-th trial as $\{\mathrm{X}_n \in \mathbb{R}^{C\times T}, \boldsymbol{\lambda}_n \in \{0,1\}^{\Lambda}\}_{n=1}^{N}$, where T denotes the number of time points sampled at a rate of F_s. In addition to the EEG data, we generate a one-hot output vector $\boldsymbol{\lambda}_n$ with $\Lambda \in \mathbb{N}$ labels. The proposed transfer learning model is evaluated on a trial basis for discriminating MI tasks. This involves extracting feature sets per trial $\{\hat{\mathrm{X}}_n^r \in \mathbb{R}^C\}_{r=1}^{R}$, where we incorporate

[1] publicly available at http://gigadb.org/dataset/100295

two EEG-based approaches simultaneously ($R = 2$): Continuous Wavelet Transform (CWT) and Common Spatial Patterns (CSP), as suggested by [31] for Deep&Wide learning frameworks. Subsequently, the extracted multi-channel features are transformed into a two-dimensional topographic interpolation $\mathbb{R}^C \rightarrow \mathbb{R}^{W\times H}$ to maintain their spatial interpretation. This mapping converts each extracted trial feature set into a two-dimensional circular view. As a result, we obtain the labeled 2D data $\{\mathrm{Y}_n^z \in \mathbb{R}^{W\times H}, \boldsymbol{\lambda}_n : n \in N\}$, where Y_n^z represents a single-trial bi-domain *t-f* feature array called *topogram*, extracted from each z-th set. It is important to note that the triplet $z=\{r, \Delta_t, \Delta_f$ (with $z \in Z$) indexes a topogram estimated for each included domain principle $r \in R$ at the time-segment $\Delta t \in T$, and within the frequency-band $\Delta f \in F$. Furthermore, we determine the local spatial patterns of relationship within the input topographic set using a square-shaped layer kernel arrangement $\{\mathrm{K}_{i,l}^z \in \mathbb{R}^{P\times P}\}^{I_l,Z}$, where P represents the kernel size. Consequently, the number of kernels varies at each layer $i \in I_l$, and the 2D-convolutional operation is performed in a stepwise manner over the input topogram, Y^z, according to the following procedure:

$$\hat{\mathrm{Y}}_L^z = \left(\boldsymbol{\varphi}_L^z \circ \cdots \circ \boldsymbol{\varphi}_1^z\right)(\mathrm{Y}^z), \tag{2-1}$$

where $\boldsymbol{\varphi}_l^z(\hat{\mathrm{Y}}_{l\text{-}1}^z) = \gamma_l(\mathrm{K}_{i,l}^z \otimes \hat{\mathrm{Y}}_{l\text{-}1}^z + \boldsymbol{B}_{i,l}^z)$ is the convolutional layer, followed by a non-linear activation function $\gamma_l : \mathbb{R}^{W_l^z \times H_l^z} \rightarrow \mathbb{R}^{W_l^z \times H_l^z}$, $\hat{\mathrm{Y}}_l^z \in \mathbb{R}^{W_l^z \times H_l^z}$ is the resulting 2D feature map of l-th layer (adjusting $\hat{\mathrm{Y}}_0^z = \mathrm{Y}^z$), and the arrangement $\boldsymbol{B}_{i,l}^z \in \mathbb{R}^{W_l^z \times H_l^z}$ denotes the bias matrix. Notations $\circ$ and $\otimes$ stand, respectively, for function composition and convolution operator [7].

2.1.3. Using a 2D feature representation

We utilize a Multilayer Perceptron (MLP) Neural Network as a deep learning-based classifier function $\boldsymbol{\varphi} : \mathbb{R}^{W\times H} \mapsto \Lambda$. This function predicts the label probability vector $\tilde{\boldsymbol{v}} \in \{0,1\}^\Lambda$ in the following manner [32]:

$$\tilde{\boldsymbol{v}} = \boldsymbol{\varphi}\left(\boldsymbol{u}_0, \boldsymbol{\Theta}; \boldsymbol{\phi}_D^z \circ \cdots \circ \boldsymbol{\phi}_1^z\right), \tag{2-2a}$$

$$\text{s.t.:} \quad \boldsymbol{\Theta}_0^* = \arg \min_{\mathrm{K}_{i,l}^z, \mathrm{A}_d, \mathrm{B}_{i,l}^z, \boldsymbol{\alpha}_d} \{\mathscr{L}(\tilde{\boldsymbol{v}}_n, \boldsymbol{\lambda}_n | \boldsymbol{\Theta}) ; \forall n \in N\} \tag{2-2b}$$

where $\phi_d(\boldsymbol{u}_{d-1}) = \eta_d(\mathrm{A}_d \boldsymbol{u}_{d-1} + \boldsymbol{\alpha}_d)$ is the fully-connected layer ruled by the non-linear activation function: $\eta_d : \mathbb{R}^{P'_d} \rightarrow \mathbb{R}^{P'_d}$, $P'_d \in \mathbb{N}$ is the number of hidden units at d-th layer, $d = \{0, \ldots, D\}$ ($d=0$ is the initial concatenation before the classification layer), $\mathrm{A}_d \in \mathbb{R}^{P'_d \times P'_{d-1}}$ is the weighting matrix containing the connection weights between the preceding neurons and the hidden units P' of layer d, $\boldsymbol{\alpha}_d \in \mathbb{R}^{P'_d}$ is the bias vector, and $\boldsymbol{u}_d \in \mathbb{R}^{P'_d}$ is hidden layer vector holds the extracted spatial information encoded by the resulting 2D feature maps in the Q domain [7].

For computation at each layer, the hidden layer vector is iteratively updated by the rule $\boldsymbol{u}_d = \phi_d(\boldsymbol{u}_{d-1})$, for which the initial state vector is flattened by concatenating all matrix rows across z and I_l domains as $\boldsymbol{u}_0 = [\mathrm{vec}(\hat{\boldsymbol{Y}}_L^z) : \forall z \in Z]$. The input vector $\boldsymbol{u}_0$ sizes $G = W'H'Z\sum_{l\in L} I_l$, holding $W' < W, H < H'$.

Besides, the optimizing estimation framework of label adjustment minimizes the training parameter set $\Theta_0=\{\mathrm{K}_{i,l}^{z},\mathrm{A}_{d},\boldsymbol{b}_{i,l}^{z},\boldsymbol{\alpha}_{d}\}$, fixing the loss function $\mathscr{L}$: $\mathbb{R}^{\Lambda}\times\mathbb{R}^{\Lambda}\rightarrow\mathbb{R}$ to calculate the gradients employed to update the weights and bias of the proposed Deep&Wide neural network through a certain number of training epochs. The solution is implemented by a mini-batch based gradient descend procedure equipped with automatic differentiation and back-propagation [94].

2.1.4. Transfer learning with added questionnaire data

In EEG analysis applying on Deep Learning for improving the classifier performance, transfer learning is a typical approach to adjust a pre-trained neural network model equipped with the label probability vector $\tilde{\boldsymbol{v}}$, aiming to provide a close domain distance measurement $\delta(\cdot,\cdot)\mathbb{R}^{+}$, lower than a given value $\varepsilon\in\mathbb{R}^{+}$, between the paired domains to approximate the source $\mathscr{Y}^{(s)}$ to the target $\mathscr{Y}^{(t)}$ [155, 33], as follows:

$$\delta(\mathscr{Y}^{(s)}(\mathrm{Y},\mathrm{S}),\mathscr{Y}^{(t)}(\mathrm{Y},\boldsymbol{S})|\tilde{\boldsymbol{v}})\leq\varepsilon \tag{2-3}$$

$$\text{s.t.:}\quad \Theta^{*}=\{\mathrm{K}_{i,l}^{q*},\mathrm{B}_{i,l}^{q*}\} \tag{2-4}$$

In this proposal, we suggest utilizing transfer learning to acquire a target prediction function that is improved by incorporating the personality assessments (questionnaire data matrix, S). Additionally, we employ the stepwise multi-space kernel embedding technique to optimize the network parameters in Ecuación (2-2b). Moreover, we select the paired source-target sets based on the clustering of subjects using an inefficiency-based approach for interpretation.
Hence, in order to combine the categorical data, S, with the real-valued feature map set obtained from EEG, $\boldsymbol{Y}$, we calculate the tensor product space between the respective kernel-matching representations, $\kappa_{\hat{\mathrm{U}}}$ and κ_{S}, as proposed in [137]:

$$\bar{\kappa}=\kappa_{\hat{\mathrm{U}}}\circ\kappa_{\mathrm{S}},\quad \bar{\kappa}\in\mathbb{R}^{J\times J} \tag{2-5}$$

Where $J=\sum_{m=1}^{M}N_m$ (N_m holds the trials for m-th subject), $\kappa_{\mathrm{S}}\in\mathbb{R}^{J\times J}$ is the kernel matrix directly extracted from the questionnaire data $\mathrm{S}\in\mathbb{R}^{J\times N_Q}$ (N_Q is the questionnaire vector lenght), $\kappa_{\hat{\mathrm{U}}}\in\mathbb{R}^{J\times J}$ is the estimated kernel topographic matrix from the projected version $\hat{\mathrm{U}}=\mathrm{U}\boldsymbol{\Upsilon}^{*}$, with $\hat{\mathrm{U}}\in\mathbb{R}^{J\times G'}$ (holding that $G'<G$), $\mathrm{U}\in\mathbb{R}^{J\times G}$ $\mathrm{U}\in\mathbb{R}^{J\times G}$ is the initial data matrix build by concatenating across the trial and subject sets all flattened vectors $\boldsymbol{u}_0^{*}$, which are computed adjusting the optimized parameters $\Theta^{*}=\{\mathrm{K}_{i,l}^{q*},\mathrm{B}_{i,l}^{q*}\}$, and $\boldsymbol{\Upsilon}^{*}\in\mathbb{R}^{G\times G'}$ is the projection matrix presented to maximize the similitude between both estimated kernel embeddings derived from the labeled EEG measurements of MI responses, namely, one from the one-hot label vectors, $\kappa_{\mathrm{V}}\in\mathbb{R}^{J\times J}$, and another from the topographic features, $\kappa_{\mathrm{U}}\in\mathbb{R}^{J\times J}$.
In particular, we match both estimated kernel embeddings through the centered kernel alignment (CKA), as detailed in [6, 33]:

$$\boldsymbol{\Upsilon}^* = \arg\max_{\boldsymbol{\Upsilon}} \mathrm{CKA}\left(\kappa_{\mathrm{U}}, \kappa_{\mathrm{V}}\right) \tag{2-6}$$

where the kernel κ_{V} is obtained from the matrix of predicted label probabilities $\mathrm{V}\in\mathbb{R}^{J\times\Lambda}$ build by concatenating across the trial and subject sets all label probability vectors $\tilde{\boldsymbol{v}}_{mn}$.

2.2. Experimental set-up

The proposed Deep&Wide neural network model for transfer learning, aimed at enhancing the classification of Motor Imagery (MI) responses, involves the following key stages (as depicted in Figura **2-1**):

- Preprocessing and Spatial Filtering: EEG signals are preprocessed and spatially filtered. Subsequently, 2D features are extracted from the input topogram set using a convolutional network.
- MLP Classification: The extracted 2D feature maps are fed into a Multi-Layer Perceptron (MLP) for classification.
- Cross-Subject Transfer Learning: This stage involves stepwise multi-space kernel embedding of real-valued and categorical variables. It facilitates knowledge transfer between different subjects to improve the model's performance. The selection of paired source-target sets is guided by inefficiency-based clustering of subjects, considering their impact on BCI motor skills.

Despite these steps, the classifier's performance may decline due to the presence of irrelevant or redundant features in the extracted representation sets. To address this issue and reduce data complexity, a widely-used unsupervised feature extractor, Kernel Principal Component Analysis (KPCA), is employed. KPCA helps in obtaining a representation of data points' global structure [168].

2.2.1. Preprocessing

For preprocessing, raw EEG channels were filtered $X_c^n \in \mathbb{R}^T$ within the frequency range of within [8-30] Hz using a five-order Butterworth band-pass filter. Subsequently, a bi-domain short-time feature extraction technique, including Continuous Wavelet Transform (CWT) and Common Spatial Pattern (CSP), was applied as previously done in [30]. The CWT feature extraction yields a time-frequency map representing the amplitudes of individual frequencies. At the same time, the CSP aims to enhance class separability by transforming the multi-channel EEG dataset into a lower-dimensional latent source space. The sliding short-time window length parameter $\tau\in\mathbb{R}^+$ was set to 2 seconds with a step size of 1 second, which allows the extraction of 5 EEG segments N_τ=5, as

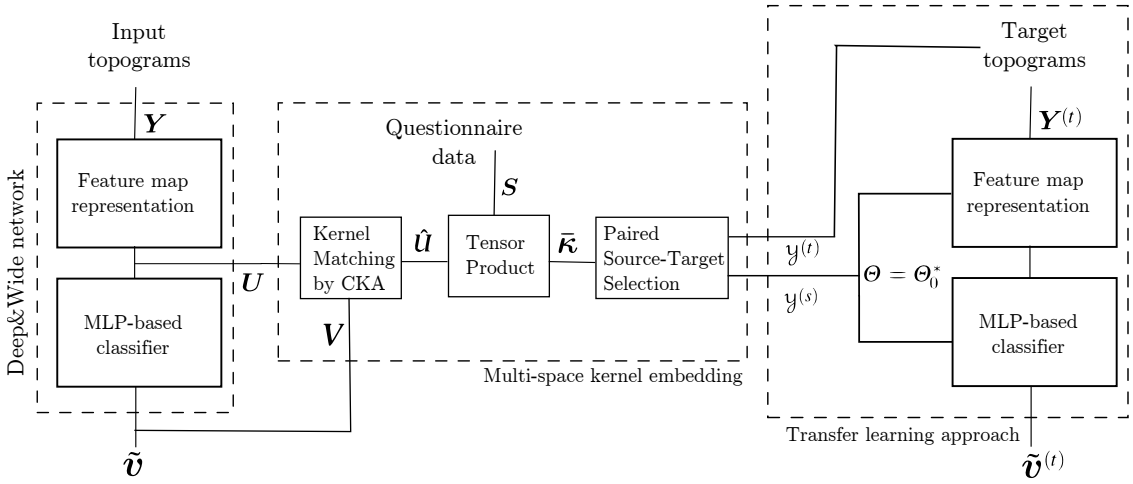


Figura 2-1: Guideline of the proposed transfer learning approach, including Stepwise Kernel Matching to combine data from Electroencephalography and Psychological Questionnaires.

performed in [150]. The spectral range of interest was split into $\Delta f \in \{\mu \in [8\text{-}12],\ \beta \in [12\text{-}30]\}$ Hz rhythms, commonly associated with electrical brain activities during MI tasks [97]. The CWT feature set was computed using the Complex Morlet function with a fixed scaling value of 32. The number of CSP components was also set to 3Λ(where $\Lambda \in \mathbb{N}$ is the number of MI tasks) using a regularized sample covariance estimation.

2.2.2. MLP classifier

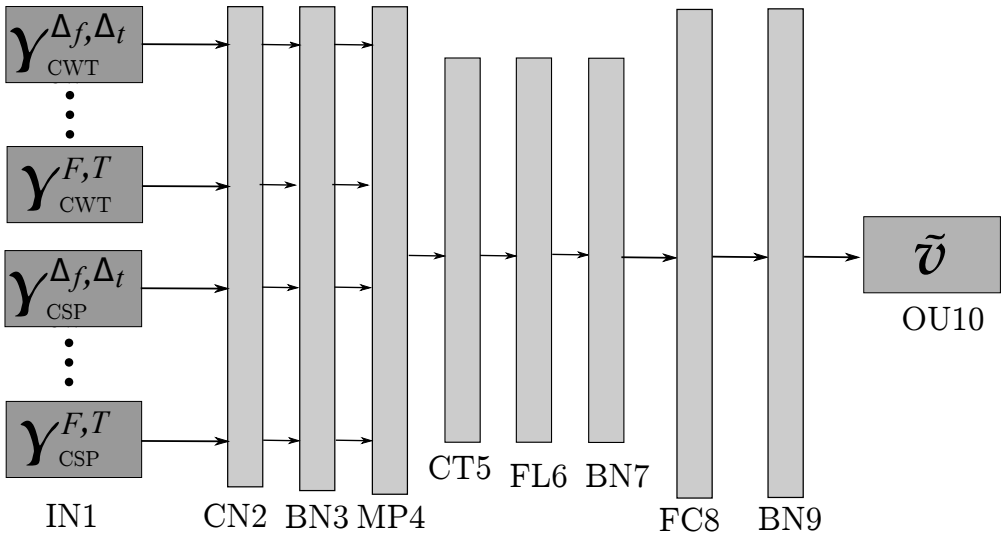


Figura 2-2: Scheme of the proposed Deep&Wide neural network architecture to support MI discrimination. [33]

We extract 2D feature maps from the input topogram set in this phase using a convolutional network. These extracted 2D features are then used to feed the MLP-based classifier, which undergoes parameter tuning as indicated in Cuadro **2-1**. The implementation employs the Adam algorithm with fixed parameters, including a learning rate of 1×10^{-3}, 200 training epochs, and a batch size of 256 samples. The chosen loss function is the mean squared error.
The Deep&Wide neural network framework is implemented using Python code with the TensorFlow toolbox and Keras API to accelerate the learning process. This implementation enables training to utilize multiple GPU devices at Google Colaboratory.

Tabla 2-1: Detailed *Deep&Wide* architecture of transfer learning. Layer FC8 accomplishes the regularization procedure using the *Elastic-Net* configuration, while layers FC8 and OU10 apply a kernel constraint adjusted to *max_norm(1.)*. Notation $O{=}RN_{\Delta}N_{\tau}$, N_{Δ} denotes the number of filter banks, P' – the number of hidden units (neurons), C – the number of classes and I_L stands for the amount of kernel filters at layer L. Notation $||\cdot||$ stands for concatenation operator.

Layer	*Assignment*	*Output dimension*	*Activation*	*Mode*
IN1	Input	$\|\|40\times40\|\|$		
CN2	Convolution	$\|\|40\times40\times2\|\|$	ReLu	*Padding* = SAME
				Size = 3×3
				Stride = 1×1
BN3	Batch-normalization	$\|\|40\times40\times2\|\|$		
MP4	Max-pooling	$\|\|20\times20\times2\|\|$		*Size* = 2×2
				Stride = 1×1
CT5	Concatenation	$\|\|20\times20\times O\cdot I_L\|\|$		
FL6	Flatten	$20\cdot20\cdot O\cdot I_L$		
BN7	Batch-normalization	$20\cdot20\cdot O\cdot I_L$		
FC8	Fully-connected	$\|\|P'\times1\|\|$	ReLu	Elastic-Net
				max_norm(1.)
BN9	Batch-normalization	$\|\|P'\times1\|\|$		
OU10	Output	$\|\|C\times1\|\|$	Softmax	*max_norm(1.)*

Figura **2-2** presents the accuracy results achieved by the MLP-based classifier when fed solely with the 2D feature set extracted earlier, evaluated on the tested subject set. The obtained accuracy values lead to performance evaluation, indicating that the classifier's effectiveness may not be sufficient for brain-computer interface (BCI) systems, as discussed in [130]. Specifically, the tested subject set is clustered into three distinct groups based on their BCI skills:
i) A group of individuals exhibiting the highest accuracy with very low variability in neural responses (colored in green).
ii) A group achieving superior classifier performance but with some response fluctuations (colored in yellow).

iii) A group displaying modest performance along with high unevenness in responses (colored in red).

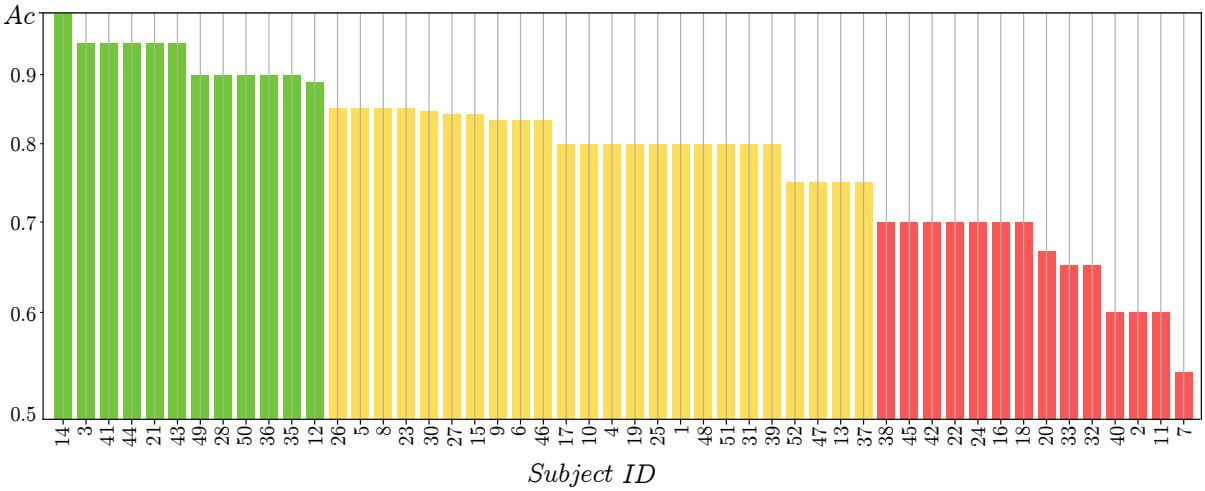


Figura 2-3: Partitions of individuals clustered by the MLP-based accuracy. Each subject performance is painted by his estimated BCI inefficiency partition: Group *I* (green), Group *II* (yellow), and Group *III* (red).

2.2.3. Kernel matching

Algoritmo 1 outlines the procedures for validating the proposed transfer learning approach with multi-space kernel embedding. The Gaussian kernel is employed to represent the available data due to its capability for universal approximation and mathematical tractability. The length scale hyperparameter $\sigma \in \mathbb{R}^{+}$, which governs the variance of the data, is adjusted based on its median estimate. The subsequent steps (3 and 4) involve pairwise kernel matching, initially between the sets of EEG measurements U and label probabilities V. The CKA matching estimator is applied to the concatenated EEG features and predicted label probabilities to achieve alignment across the entire subject set. Empirically, the parameter G' is set to 50, considering the number of subjects in this experiment. In the second matching, all available categorical information from the psychological and physiological evaluations is encoded using CKA, and the resulting feature set is projected onto a common matrix space representation through the kernel/tensor product. It is worth noting that the projected data $\hat{\mathrm{U}}$ obtained from CKA is also embedded. Moreover, we perform dimensionality reduction on the feature sets generated after the stepwise matching using Kernel Principal Component Analysis (KPCA) to evaluate their representational ability.

Furthermore, we estimate the subject similarity matrix to calculate the domain distance between the source-target pairs selected from different clusters of BCI inefficiency. The neighboring similarity matrix $\bar{\Delta}\hat{\xi}$ is introduced, and its pairwise metric elements are computed from the matrices $\hat{\xi} \equiv \hat{\kappa}, \hat{\kappa}$KPCA, as described equation Ecuación (2-7). Here, $\mathrm{cov}(\cdot,\cdot)$ and $\mathrm{seq}(\Delta(m,\forall m')) \in \mathbb{R}^{M}$ correspond to the covariance operator and the sequence composed of all elements of row m, ranked

Algorithm 1 Validation procedure of the proposed approach for transfer learning with stepwise, multi-space kernel matching. † Dimensionality reduction is an optional procedure performed for comparison purposes [33].

Input data: EEG measurement U, predicted label probabilities V, questionary data S, $\forall m \in M$

1: INITIAL PARAMETER SET ESTIMATION Θ_0^*: Compute the baseline MLP-based accuracy from U and V by optimizing $\Theta = \{\mathrm{K}_{i,l}^q, \mathrm{A}_d, \mathrm{b}_{i,l}^q, \alpha_d\}$

2: **for** $\forall m \in M, n \in N$ **do**

3: KERNEL MATCHING between EEG measurement U and labels V,

- Compute Kernel embedding of input data $\xi = \{\kappa_\mathrm{U}, \kappa_\mathrm{S}, \kappa_\mathrm{V}\}$: $\kappa_\xi = \mathscr{N}_\xi(\mu_\xi, \sigma_\xi)$
- Compute Center Kernel Alignment between both spaces:$\mathrm{CKA}(\kappa_\mathrm{U}, \kappa_\mathrm{V})$

4: KERNEL MATCHING on supervised EEG representation for the categorical data

- Compute Kernel embedding of projected data $\hat{\mathrm{U}}$ using $\kappa_{\hat{\mathrm{U}}} = \mathscr{N}_{\hat{\mathrm{U}}}(\mu_{\hat{\mathrm{U}}}, \sigma_{\hat{\mathrm{U}}})$
- Compute tensor product, including the categorical data $\bar{\kappa} = \kappa_{\hat{\mathrm{U}}} \circ \kappa_\mathrm{S}$, $\bar{\kappa} \in \mathbb{R}^{J \times J}$, $J = NM$

5: **end for**

DIMENSIONALITY REDUCTION† by Kernel Principal Components: $\bar{\kappa}_{\mathrm{KPCA}} \in \mathbb{R}^{J \times J}$

6: TRANSFER LEARNING OF PAIRED SOURCE-TARGET SUBJECTS: $\mathscr{Y}^{(s)}$ and $\mathscr{Y}^{(t)}$

- Perform matrix reshaping $\mathbb{R}^{J \times J} \mapsto \mathbb{R}^{M \times J}$: $\hat{\xi} = \{\hat{\bar{\kappa}}, \hat{\kappa}_{KPCA}\}$
- Compute the neighboring similarity matrix of individuals: $\bar{\Delta}, \bar{\Delta}_{KPCA}$
- Compute the intra-subject distance matrix through the domain distance measurement: $\bar{\delta}_{\hat{\xi}}(m) \in \mathbb{R}^+, \forall m \in M$
- Select paired subjects for each transfer learning strategy evaluated:
- *a*) one-source versus one-target, *b*) multiple-source versus one-target
- Recompute the MLP-based accuracy of targets, initializing the parameter set as $\Theta = \Theta_0^*$, fixing P' parameter according to source subject.

7: **Output data**: Accuracy gain achieved by each individual target, according to the selection transfer learning strategy evaluated.

in descending order of the achieved MLP-based accuracy. We were applying covariance to the ranked row vectors of $\bar{\Delta}_{\hat{\xi}}$ is motivated by preserving the similarity information between neighboring subjects.

$$\bar{\Delta}_{\hat{\xi}}(m,m') = \mathrm{cov}(\mathrm{seq}(\Delta_{\hat{\xi}}(m,\forall m')), \mathrm{seq}(\Delta_{\hat{\xi}}(m',\forall m))), \quad \bar{\Delta}_{\hat{\xi}}(m,m') \in \bar{\mathbf{\Delta}}_{\hat{\xi}} \in \mathbb{R}^{M\times M} \tag{2-7}$$
$$\Delta_{\hat{\xi}}(m,m') = \sum_{\forall j \in J} |\hat{\xi}(m,j) - \hat{\xi}(m',j)|^2, \quad \Delta_{\hat{\xi}}(m,m') \in \mathbf{\Delta}_{\hat{\xi}} \in \mathbb{R}^{M\times M}$$

[33]

Figure **2-4** presents the similarity matrix obtained through the tensor product $\mathbf{\Delta}\hat{\xi}$ (left column), revealing some of the relationships between clustered subjects, depending on the evaluated questionnaires. For instance, the dataset Q_1 exhibits two distinct groups, while Q_4 shows three partitions. However, Q_2 and Q_3 do not accurately cluster the individuals. Subsequently, after performing dimensionality reduction using KPCA, the proximity assessments $\bar{\Delta}KPCA$ tend to strengthen the neighboring associations, leading to more well-defined clusters of subjects with distinct feature representations, as shown in the middle column for each questionnaire.

Under the assumption that the closest the association between the paired source-target couples, the more effective their cross-subject transfer learning is executed, we estimate the marginal distance $\bar{\delta}_{\hat{\xi}}(m)\in\mathbb{R}^{+}$ from either version $\mathbf{\Delta}_{\hat{\xi}}, \bar{\gtrless}_{KPCA}$ by averaging the neighboring similarity of each subject over the entire set, where the notation $F : \forall\zeta$ stands for the expectation operator computed across the whole set $\{\zeta\}$, as follows:

$$\bar{\delta}_{\hat{\xi}}(m) = \mathbb{E}\left\{ |\bar{\Delta}_{\hat{\xi}}(m,m')| : \forall m' \in M \right\}, \tag{2-8}$$

In the right column, the values of marginal distances $\bar{\delta}_{\hat{\xi}}(m)$ illustrate that each individual is influenced differently by the stepwise multi-space kernel matching of electroencephalography to psychological questionnaires Q_i. These findings align with the subject cluster properties analyzed earlier. Specifically, Q_1 and Q_4, with more distinct partitions, yield feature representations that are more evenly distributed within the subject set, while Q_2 and Q_3 provide irregular representations. Furthermore, dimensionality reduction enhances the representation of cases in Q_1 and Q_2, while using KPCA tends to diminish the overall similarity level among individuals.

2.2.4. Transfer Learning

The next step involves pairing the learned representation from a *source* subject to a specific *target* subject. Based on the subject partitions according to their BCI skills obtained in Apartado 2.2.2, we choose the candidate sources (i.e., the source space $\mathscr{Y}^{(s)}(,)$) from the best-performing subjects (Group *I*), while the target space $\mathscr{Y}^{(t)}(,)$ consists of the worst-performing participants (Group *III*). We explore two strategies for selecting subjects from the source space (Group *I*) [33]:

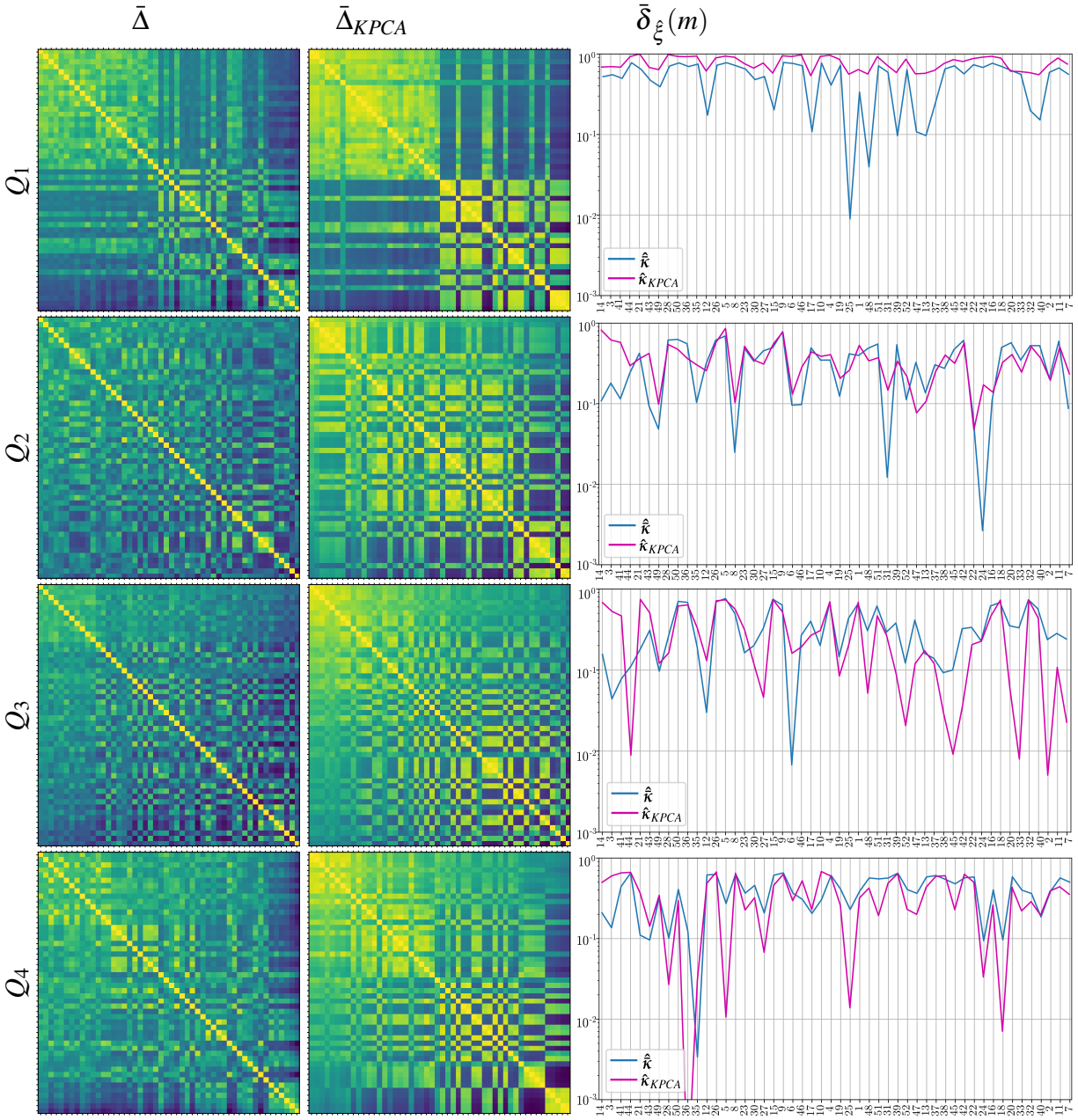


Figura 2-4: Similarity matrix performed by the tensor product and computed domain marginal values $\bar{\delta}_{\hat{\xi}}(m)$. The subjects are ranked in decreasing order of accuracy.

a) *Single source – single-target*, where we select the subject in Group *I* that achieves the highest value of the domain distance measurement in Ecuación (2-9), computed as follows:

$$\max_{\forall m \in \text{Group } I} \bar{\delta}_{\hat{\xi}}(m; Q_i) \tag{2-9}$$

Once the *source-target* pairs are chosen, the pre-trained weights are calculated from each designed *source* subject to initialize the Deep&Wide neural network. This approach avoids

introducing a zero-valued starting iterate and facilitates better convergence of the training algorithm. It is worth noting that the condition in Ecuación (2-9) depends on Q_i, resulting in distinct selected sources for each questionnaire data.

b) *Multiple sources – single-target*, where we choose the subjects in Group I with the four highest domain distance values. In this case, the Deep&Wide initialization procedure utilizes the pre-trained weights estimated from the concatenation of the source topograms.

Figura **2-5** provides detailed classification performance achieved using the proposed transfer learning approach for each strategy of selecting candidate sources. The radar diagram includes all target subjects as axes. For comparison, the graphical representation also shows the MLP-based accuracy (colored in black) as a reference for assessing the performance gain due to the applied transfer learning approach. The blue line represents the accuracy achieved by the features extracted by the tensor product. In contrast, the magenta line corresponds to the accuracy obtained using KPCA, i.e., $\hat{\hat{\kappa}}, \hat{\kappa}_{KPCA}$, respectively.
The odd columns (first and third) show the diagrams for the *Single source – Single-target* strategy, while the even columns are for the *Multiple sources – Single-target* approach. In all cases of questionnaire data Q_i, the stepwise transfer learning, multi-space kernel matching results in an average increase in the baseline classifier performance for subjects belonging to Group *III*, who exhibit modest accuracy and high unevenness of responses. However, there are still some aspects that require further clarification. The *Single source – Single-target* strategy achieves a lower accuracy gain than the latter approach, but it benefits a more significant number of subjects. On the other hand, the *Multiple sources – Single-target* strategy reduces the number of poor-performing subjects that show improvement. However, some subjects demonstrate substantial accuracy gains, such as subject #45 (up to 25 %).
The contribution of categorical data to classifier performance is another aspect to address. The first pair radars in the bottom row (labeled as EEG) show the accuracy improvement achieved by the features extracted from EEG measurements after CKA alignment ($\mathrm{CKA}(\kappa_{\mathbb{U}}, \kappa_{\mathbb{V}})$), which underperforms the transfer learning when questionnaires are added. Regarding the additional dimensionality reduction, its impact on accuracy (shown in magenta) is strongly influenced by the specific fused data Q_i. While Q_1 and Q_2 benefit from the KPCA procedure, Q_4 experiences a decrease in performance. This is evident in the two bottom radars (third and fourth) that depict the averaged effect of transfer learning across the data Q_i, demonstrating that the classifier performance of almost every target individual can be improved by the proposed transfer learning approach, regardless of the strategy for selecting candidate sources. However, some subjects (#38 and #20) do not show positive impacts.
Lastly, the topographic maps in Figura **2-6** visually interpret the proposed transfer learning, reconstructed from the learned network weights according to the algorithm presented in [32]. We compare the estimated approaches, assuming that the discriminatory ability is proportional to the reconstructed weight value. Topograms of a Single-source strategy constructed from both bandwidths ($/beta$ and $/mu$) within different intervals of neural response are shown in the top row.

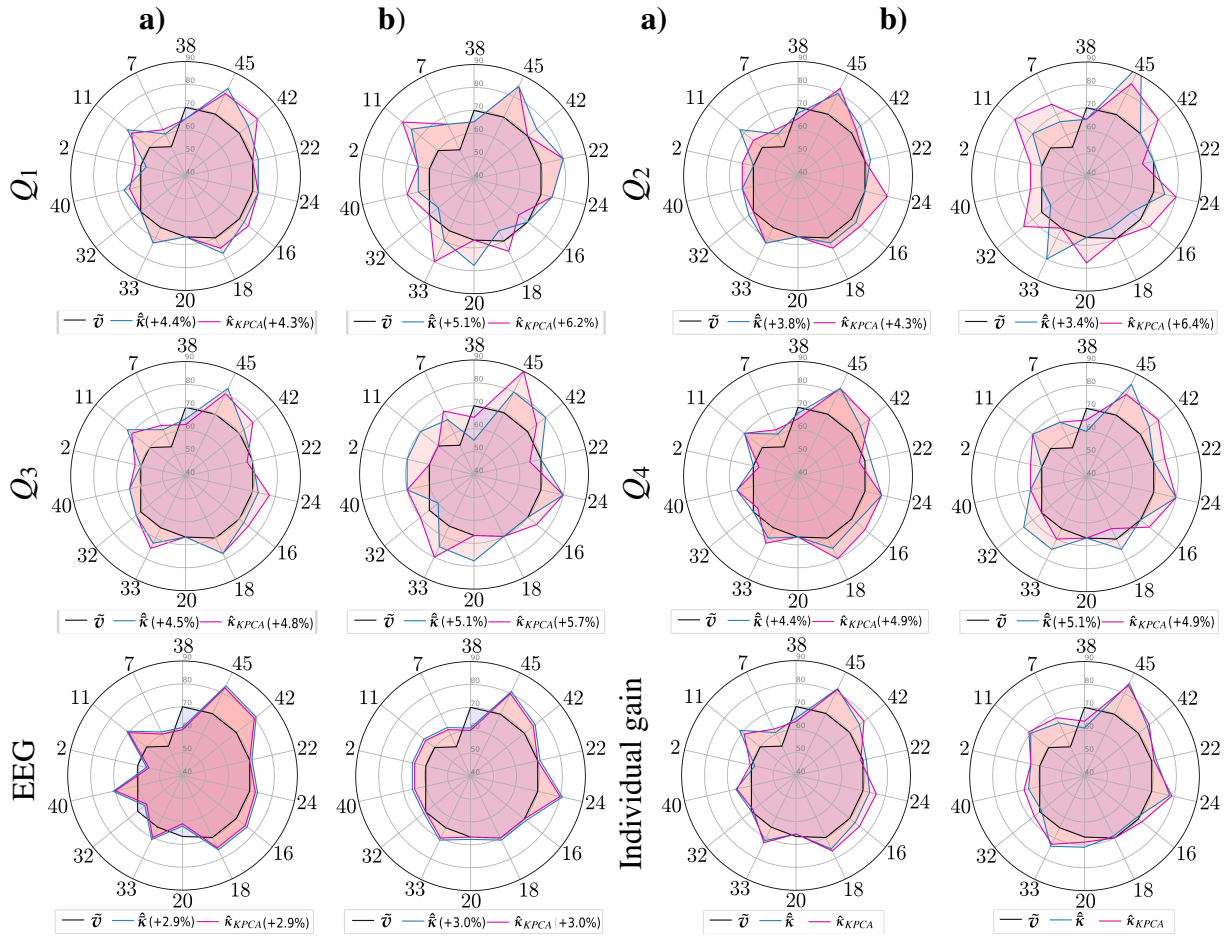


Figura 2-5: Achieved accuracy by validated strategies of selecting source subjects from Group I. a) *Single source – single-target*, b) *Multiple sources – single-target*. Individual gain reports the average accuracy per subject of questionnaire data Q_i and EEG.

The selected source (subject #3) exhibits a weight set with a spatial distribution related to the sensorimotor area, accurately focusing neural responses within the MI segment. Next to Subject #3, we present the target's topograms that help the most from the transfer learning, showing weights with a slightly blurred spatial distribution. The Single-source transfer learning approach reduces the weight variability, as observed in the adjacent topograms. However, the source's effectiveness in reducing variability is limited in the case of the low-skilled target #38, which presents many contributing weights spread across the entire scalp area. Furthermore, the weights appear inside the two intervals (before the cue-onset and ending segment), during which the responses provoked by MI tasks are believed to vanish. As a result, the single-source strategy yields a negative accuracy gain for Target #38 (it drops from 70 % to 65 %). This pattern can also be seen in the second row, which displays the multi-source topograms for the most beneficial target (Target #11) and the least successful target (Target #22). However, including numerous sources leads to weights with a sparse distribution, as observed in the topograms of the chosen subjects (Subjects #3, 14,

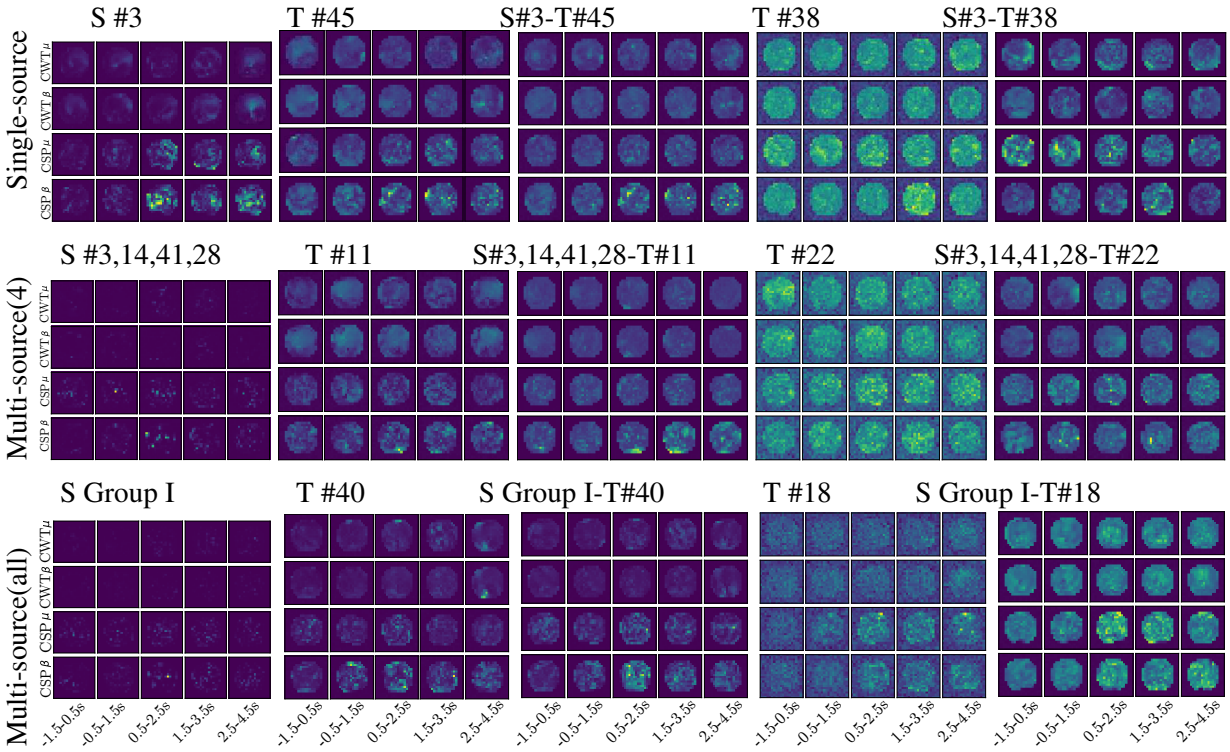


Figura 2-6: Topographic maps of representative subjects with and without transfer learning using just feature map information, presenting the learned weights with meaningful activity reconstructed within both bandwidths (β and μ) across the whole signal length, N_τ.

41, 28). This effect may explain the small number of targets the multi-source strategy improves. To explain this point, the bottom row displays the corresponding spatial distribution performed by the multi-source strategy when including the whole subject set of Group I, resulting in weak and scattered weights. Moreover, compared with the first two rows, the all-subjects source approach in the bottom row leads to the worst performance averaged across the target subject set.

2.3. Discussion

The evaluation highlights the following aspects:

Evaluated NN framework: We utilize the Deep&Wide learning framework with 2D feature maps supporting the MLP-based classifier. Comparison in Table 1 (**2-2**) shows that our proposed transfer learning method outperforms several recently published approaches regarding bi-class accuracy on the GigaScience database.

Challenges and future directions: Despite achieving high accuracy, three aspects require further development. Firstly, we present the bi-domain extraction (CWT and CSP) to address substantial intra-subject variability across trial patterns. However, more compact feature representations need exploration to improve their combination with categorical data, such as using connectivity me-

Tabla 2-2: Comparison of Bi-class accuracy achieved by state-of-the-art approaches in *GigaScience*. The best value is marked in bold. Notation * denotes Deep&Wide framework results with transfer learning (TL). CSP+FLDA: Common spatial patterns and Fisher linear discriminant analysis, LSTM+Optical: Long-short term memory network and optical predictor, SFCSP: Sparse filter-bank CSP, DCJNN: Deep CSP neural network with joint distribution adaptation, MINE+EEGnet: Mutual information neural estimation, MSNN: Multi-scale Neural Network.

Approach	A_c	*Interpretability*
CSP+FLDA [26]	67.60	–
LSTM+Optical [80]	68.2±9.0	–
SFBCSP [174]	72.60	–
DCJNN [176]	76.50	✓
MINE+EEGnet [70]	76.6±12.48	✓
MSNN [76]	81.0±12.00	✓
Proposal	79.5±10.80	✓
Proposal+TL*	**82.6± 8.40**	✓

trics as in [99]. Secondly, developing neural network architectures to capture temporal dynamics and maintain local structures of the time series associated with elicited MI responses, as carried out in [55]. Lastly, the approach takes longer training hours for multi-source cases than single-source, which is considerably faster in network training. However, the model exhibits fast test set classification once fully trained (Table 2, **2-3**), and parameter tuning details are provided.

Tabla 2-3: Achieved training and validation time for each subject dataset.

Approach	*Time per fold*	*Time per training epoch*
Proposal (Single-source)	$\sim 984s$	$< 1s$
Proposal (Multi-source (4))	$\sim 1663s$	$< 1s$
Proposal (Multi-source (all))	$\sim 3176s$	$\sim 1s$
Proposal+TL	$\sim 341s$	$< 1s$

Multi-space kernel matching: We implement stepwise kernel matching via Gaussian embedding to address the challenges of combining categorical and real-valued features. The similarity matrices obtained reveal relationships with BCI inefficiency clusters of subjects. While the evaluated questionnaire data influence the association, this result is essential given previous reports stating no statistically significant differences between questionnaire scores and EEG-based performance [84, 33]. To improve predicting MI performance based on subjective criteria, two main issues need addressing: using more appropriate kernel embedding for categorical scores [23] and dimensionality reduction approaches, such as *t*-Distributed Stochastic Neighbor Embedding [9].
Cross-subject transfer learning: Our transfer learning infers a target prediction function from the

embedded kernel spaces, selecting paired source-target sets based on Inefficiency-based clustering by subjects. This approach increases the baseline classifier accuracy of the worst-performing subjects. However, the way source selection is performed impacts classifier performance. The Multiple-source – Single-target strategy yields more significant accuracy improvements than Single-source – Single-target, but the number of benefited targets decreases. This result suggests the need for future exploration of more effective transfer learning of BCI inefficiency, aiming to combine the source domain with each target space as much as possible. This task is also essential to improve the similarity metric proposed for comparing ordered vectors of different BCI inefficiency clusters 2-7.

2.4. Summary

This chapter proposes a Deep&Wide neural network for motor imagery (MI) classification. The network is first pre-trained on data from the source domain. The layer parameters are then transferred to initialize the target network, which is fine-tuned to recompute the Multilayer Perceptron (MLP)-based accuracy.
We implement stepwise kernel matching via Gaussian embedding to perform data fusion, combining categorical with real-valued features. Paired source-target sets are selected based on inefficiency-based clustering by subjects to evaluate their influence on BCI motor skills for evaluation purposes. We explore two strategies for choosing the best-performing subjects in the source space: single-subject and multiple-subjects. The validation results achieved for discriminant MI tasks demonstrate that the introduced Deep&Wide neural network presents a competitive accuracy performance, even after including questionnaire data [33].

3 Sonification through Labeled Correlation Alignment

3.1. Materials and Methods

3.1.1. Dataset: EEG data investigating neural correlates of music-induced emotion (BCMI-MIdAS)

This dataset [1] was collected from a total of N_S=31 participants. The experimental setup involved six runs, capturing neural responses from the brain, divided into two segments. During the baseline resting recordings, participants remained seated and focused on the screen. Signal acquisition for each subject was performed from 19 channels based on the 10-20 electrode placement system Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2. Each recording lasted 15 seconds and was sampled at 1000 Hz.

The study utilized music stimuli to investigate how music influences emotions. The dataset included 110 excerpts from various scores, covering a wide range of emotional responses, as described in [39].

Notably, the auditory data were labeled based on the two-dimensional arousal-valence plane. Affective states were characterized by the combination of arousal (activated-deactivated) and valence (pleasure-displeasure), resulting in four distinct labeled partitions (N_L=4) [126]: High Arousal Positive Valence states (HAPV), High Arousal Negative Valence (HANV), Low Arousal Negative Valence (LANV), and Low Arousal Positive Valence (LAPV).

3.1.2. Extraction of (Audio)Stimulus-(EEG)Responses

A piecewise stationary analysis accounts for the non-stationarity behavior inherent to training data when characterizing the eliciting acoustic stimuli ($\mathscr{Y} \in \mathbb{R}$) and brain neural responses ($\mathscr{X} \in \mathbb{R}$). Thus, both feature sets ($X \in \mathscr{X}, Y \in \mathscr{Y}$) are extracted from M_τ overlapping segments framed by a smooth-time weighting window lasting $\tau_m \leq T$, with $m \in M_\tau$, where $T \in \mathbb{R}$ is the recording length [7]. Specifically, a set of time-windowed neural response features, $\mathscr{X} \to X$, is extracted from the EEG electrode montage using two functional connectivity metrics (FC), Phase Locking Value (PLV)

[1] publicly available at https://openneuro.org/datasets/ds002721/versions/1.0.2

and Gaussian FC (GFC), estimated on a trial-by-trial basis, respectively as [57, 7]:

$$\Delta\phi_V(x_m^c, x_m^{c'}) = \mathbb{E}\left\{ |\exp(j(\phi_m^c(t) - \phi_m^{c'}(t)))| : \forall t \in \tau_m \right\} \tag{3-1a}$$

$$\Delta\phi_G(x_m^c, x_m^{c'}; \sigma_\phi) = \exp\left(\frac{-\|x_m^c - x_m^{c'}\|_2^2}{2\sigma_\phi^2} \right) \tag{3-1b}$$

where x_m^c and $x_m^{c'}$ are the real-valued EEG vectors captured at instant $m \in M_\tau$ from the corresponding electrodes $c, c' \in N_C$; $\phi_m^c(t)$ and $\phi_m^{c'}(t)$ are the corresponding instantaneous phases $\phi_m^c(t)$ and $\phi_m^{c'}(t)$, with $c \neq c'$, N_C is the number of testing montage channels $\{\boldsymbol{x}_m^c \in [x_m^c : m \in M]\} \in \mathscr{X}$, and $\sigma_\phi \in \mathbb{R}^+$ a length scale hyperparameter. Notations $\|\cdot\|_2$ and $\mathbb{E}\{: \forall v\}$ stand for ℓ_2-norm and expectation operator computed across a variable v, respectively [7].
In parallel, a set of time-windowed acoustic features, $\mathscr{Y} \to Y$, is extracted under the music assessment and music listening paradigms [104]: Zero-Crossing Rate, Zero-Crossing Rate, High/Low Energy Ratio, Spectral Entropy, Spectral Spread, Spectral Roll-off, Spectral Flatness, Roughness, RMS energy, Broadband Spectral Flux, and Spectral flux for ten octave-wide sub-bands. The extracted acoustic features' descriptions are detailed in [110, 77]. Also, the feature set is completed by the short-time auditory envelopes extracted as in [74, 7].

3.1.3. Two-step Labeled Correlation Alignment between Audio and EEG Features

The proposed feature alignment procedure between eliciting audio-stimuli and aroused EEG responses consists of two steps: Firstly, the similarity of each feature space to the label set is assessed using Centered Kernel Alignment. This space allows selecting the extracted representations that match the closest. After selecting the labeled CKA representations, Canonical Correlation Analysis is performed to identify audio and EEG features that are maximally congruent in terms of estimated correlation coefficients [7].

Supervised CKA-based selection of features. Sonification feature sets must be selected to create music following brain patterns but according to distinct emotional conditions. Hence, the alignment is performed separately between each feature set, $\Xi = \{X \in \mathbb{R}^{N_R \times P}, Y \in \mathbb{R}^{N_R \times Q}\}$ being P and Q the number of EEG and Audio features (N_R is the number of trials), to the provided labels, noted as $\Lambda \in \mathbb{Z}$, employing the CKA algorithm that includes an additional transformation to estimate the contribution of every input representation. To be specific, we use the supervised empirical estimate of CKA derived in [5, 7], as follows:

$$\boldsymbol{w}_\Xi{}^* = \underset{\boldsymbol{W}_\Xi}{\arg\max} \frac{\langle \bar{\boldsymbol{K}}_\Xi(\boldsymbol{W}_\Xi), \bar{\boldsymbol{K}}_\Lambda \rangle_F}{||\bar{\boldsymbol{K}}_\Xi(\boldsymbol{W}_\Xi)||_F ||\bar{\boldsymbol{K}}_\Lambda||_F}; \tag{3-2}$$

where notation $||\cdot||_F$ stands for Frobenius norm, $\bar{\boldsymbol{K}} \in \mathbb{R}^{N_R \times N_R}$ is the centered kernel matrix estimated as $\bar{\boldsymbol{K}} = \tilde{\boldsymbol{I}} \boldsymbol{K} \tilde{\boldsymbol{I}}$, $\boldsymbol{K} \in \mathbb{R}^{N_R \times N_R}$ is the kernel matrix, $\tilde{\boldsymbol{I}} = \boldsymbol{I}\text{-}\boldsymbol{1}^\top \boldsymbol{1}/N_R$ is the empirical centering

matrix computed across the trial set that holds N_R, and $\boldsymbol{I}\in\mathbb{R}^{N_R\times N_R}$ is the identity matrix, $\mathbf{1}\in\mathbb{R}^{N_R}$ is the all-ones vector; and$\boldsymbol{K}_{\Xi}\in\mathbb{R}^{N_R\times N_R}$ and $\boldsymbol{K}_{\Lambda}\in\mathbb{R}^{N_R\times N_R}$ are the kernel matrices that match each extracted feature set to the labels, respectively.
The kernel matrix elements, $\xi,\xi'\in\Xi$, are computed on a trial-by-trial basis, respectively, as follows:

$$\kappa_{\Xi}\left(\xi,\xi';\boldsymbol{W}_{\xi}\right)=\exp\left(\text{-}\left((\xi\text{-}\xi')^{\top}\boldsymbol{W}_{\xi}^{\top}\boldsymbol{W}_{\xi}(\xi\text{-}\xi')\right)/2\right), \tag{3-3a}$$

$$\kappa_{\Lambda}\left(\lambda,\lambda'\right)=\delta\left(\lambda,\lambda'\right),\quad \lambda,\lambda'\in\Lambda \tag{3-3b}$$

where $\boldsymbol{W}_{\xi}$ is the matrix linearly transforming the selected $\tilde{\xi}$ and input ξ sets in the form $\tilde{\xi}=\xi\boldsymbol{W}_{\xi}$, with $\tilde{\xi}\in\{\tilde{X}\in\mathbb{R}^{N_R\times P},\tilde{Y}\in\mathbb{R}^{N_R\times Q}\}$, being $\boldsymbol{W}_{\xi}\boldsymbol{W}_{\xi}^{\top}$ the corresponding inverse covariance matrix of the multivariate Gaussian function as in Ecuación (3-3a) [7].
A Gaussian function is used as the first kernel $\kappa_{\Xi}(,)\in\mathbb{R}^{+}$ in Ecuación (3-3a), to assess the pairwise similarity between aligned features due to its universal approximation properties and tractability [156]. The second kernel includes the delta operator $\delta(\cdot,\cdot)$ in Ecuación (3-3b) suitable for dealing with categorical label values.

CCA-based analysis of multimodal features. This unsupervised statistical technique aims to assess the pairwise linear relationship between the multivariate projected feature sets $\widetilde{\Xi}=\{\tilde{X},\tilde{Y}\}$ obtained by supervised CKA-based selection and described in different coordinate systems (EEG and Audio). To this end, both representation sets are mapped into a common latent subspace to become maximally congruent. Namely, the correlation between the EEG and auditory features is maximized across all N_R trials within a quadratic framework constrained to a single-dimensionality latent subspace, as below [166, 7]:

$$\hat{\alpha}_{\tilde{X}},\hat{\alpha}_{\tilde{Y}}=\underset{\alpha_{\tilde{X}},\alpha_{\tilde{Y}}}{\arg\max}\,\alpha_{\tilde{X}}^{\top}\Sigma_{\tilde{X}\tilde{Y}}\alpha_{\tilde{Y}} \tag{3-4a}$$

$$\text{s.t.: }\alpha_{\tilde{X}}^{\top}\Sigma_{\tilde{X}\tilde{X}}\alpha_{\tilde{X}}=1,\quad \alpha_{\tilde{X}}\in\mathbb{R}^{P} \tag{3-4b}$$

$$\alpha_{\tilde{Y}}^{\top}\Sigma_{\tilde{Y}\tilde{Y}}\alpha_{\tilde{Y}}=1,\quad \alpha_{\tilde{Y}}\in\mathbb{R}^{Q} \tag{3-4c}$$

where $\Sigma_{\tilde{X}\tilde{X}}\in\mathbb{R}^{P\times P}$, $\Sigma_{\tilde{Y}\tilde{Y}}\in\mathbb{R}^{Q\times Q}$, and $\Sigma_{\tilde{X}\tilde{Y}}=\tilde{X}^{\top}\tilde{Y}\in\mathbb{R}^{P\times Q}$.

3.1.4. Sonification via Vector Quantized Variational AutoEncoders

The feed-forward encoder and decoder network converts an input time-series $\boldsymbol{\xi}=[\xi_t:\forall t]$, with $\boldsymbol{\xi}\in\Xi$, into a coded form of a discrete finite set (or tokens), $\boldsymbol{z}\in\{z_s:\forall s\in S\}$, having each element of size K. To this end, a latent representation $h_s=\theta_E(\boldsymbol{\xi})$ (with $H\in\{h_s\}$) is encoded to be further element-wise quantized according to the vector-quantized codebook $\{e_k:\forall k\}$ [7]. The VQ-VAE model noted as

$\mu(\boldsymbol{\xi})$ is then trained using the minimizing framework, as below [44, 7]:

$$\mu(\boldsymbol{\xi}) : \min \mathbb{E}\left\{ \|\xi_t - \theta_D(e_{z,t})\|_2^2 : \forall t \right\} + \mathbb{E}\left\{ \|\theta_{SG}(h_s) - e_{z,s}\|_2^2 : \forall k \right\} + \beta \mathbb{E}\left\{ \|h_s - \theta_{SG}(e_{z,s})\|_2^2 : \forall k \right\} \quad (3\text{-}5)$$

where the first term is the reconstruction loss that penalizes for the distance between input $\boldsymbol{\xi}$ and decoded output $\tilde{\boldsymbol{\xi}}=\theta_D(\cdot)$, the second term penalizes for the distance between each encoding value of H and their nearest neighbors e_z in the codebook, and the third term prevents the encoding from strong fluctuations, ruling the weight $\beta \in \mathbb{R}[0,1]$. Besides, notation $\theta_{SG}(\cdot)$ stands for the stop-gradient operation, which passes zero gradients during backpropagation.
Generally speaking, the coding model trained by one auditory signal set $\boldsymbol{\xi} \in \Xi$ can be applied to the generation of acoustic data when feeding to the encoder signals of different nature, $\boldsymbol{\xi}' \in \Xi$, provided their homogeneity is assumed. This model is referred to as $\mu(\boldsymbol{\xi}|\boldsymbol{\xi}')$. In light of this, we suggest that the following conditions be met [7]:

– The VQ-VAE coder includes a parametric spectrum estimation based on regressive generative models fitted on latent representations [75]. Therefore, both sets of signals $(\boldsymbol{\xi}, \boldsymbol{\xi}')$ must have similar spectral content, at the very least, in terms of their spectral bandwidth. That is,

$$\Delta F_{\boldsymbol{\xi}} \simeq \Delta F_{\boldsymbol{\xi}'} \quad (3\text{-}6)$$

– In regression models, both discretized signal representations must be extracted using similar recording intervals and time windows to perform the numerical derivative routines. Furthermore, the VQ-VAE coder demands input representations of fixed dimensions. Hence, the arrangements extracted from $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ must be of similar dimensions.

3.2. Experimental set-up

We introduce a novel approach for translating neural responses to affective music listening into auditory signals, leveraging both auditory and electroencephalographic features that align optimally with the corresponding emotion labels. Our method is evaluated within the stimulus-response paradigm, involving various stages as illustrated in Fig. **3-1**:

- Preprocessing and extracting time-windowed representations: We estimate acoustic features from music data that modulate emotions, alongside Functional Connectivity (FC) measures derived from evoked EEG neural responses. To capture diverse FC characteristics, we explore three strategies: Phase Locking Value, Gaussian Functional Connectivity, and their combination. Additionally, different time windows are evaluated for feature extraction from neural brain responses, specifically focusing on low-level music generation.

- Labeled Correlation Alignment: This stage aims to identify EEG features that align most significantly with the stimulating auditory data for each emotion. To ensure the interpretability of the selected arrangements, we employ a two-step procedure: first, a separate Canonical Correlation Analysis (CCA) is conducted between audio and EEG data using the emotion labels, followed by CCA analysis of the selected feature sets.

  We examine the contribution of electrodes and bandpass-filtered, time-windowed dynamics to the Labeled Correlation Alignment process, while also considering the impact of individual subjects on overall performance.

- Generating Labeled audio conditioning content: We utilize the selected brain neural responses to feed a Vector Quantized Variational AutoEncoder, producing the labeled audio conditioning content.

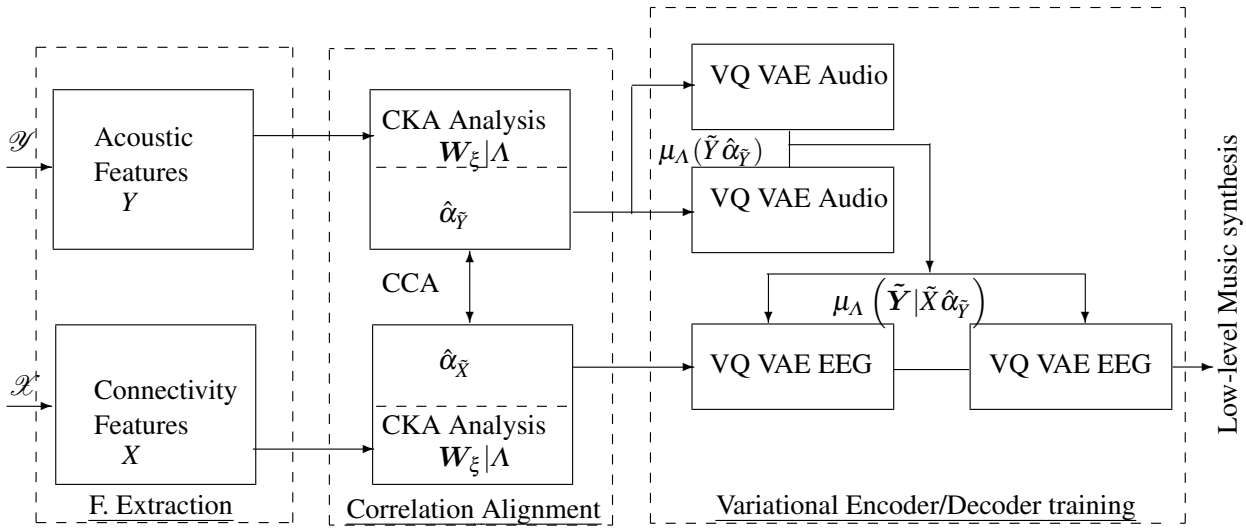


Figura 3-1: Proposed model architecture.

We evaluate the correlation between the captured neural responses and auditory data using Canonical Correlation Analysis (CCA) as a performance measure. The strength of the relationship is quantified by the *r*-squared coefficient, with higher values indicating a stronger association between brain responses and auditory stimuli. To assess the confidence of CCA correlation estimates, we employ leave-one-out cross-validation, specifically, leave-one-subject-out, as demonstrated in a prior study [14].
Furthermore, we assess the discrimination ability of the labeled correlation alignment using a clustering coefficient, denoted as γ and belonging to the positive real numbers ($\gamma \in \mathbb{R}^+$). This coefficient represents the quality of the partition obtained from the CCA correlation values and is computed as follows:

$$\gamma = \left(\frac{\xi_1 - \xi_0}{\max_i \xi_i} + \mathbb{E}(\xi_n - \bar{\xi})^2 \colon \forall n \in N_R \right), \quad \xi_i \in \Xi$$

Here, ξ_0 represents the mean distance between a sample and all other points in the same group, ξ_1 indicates the mean distance between a sample and all other points in the closest group, ξ_n denotes the number of samples within the data set, and $\bar{\xi}$ represents the center of a group. The clustering measure γ strikes a balance between inter-class variability (first term) and intra-class variability (second term). Consequently, larger values of γ indicate more distinct labeled partitions of the extracted features [29].

3.2.1. Preprocessing

EEG Prepocessing

- We performed high-pass filtering on the raw EEG channel set using a 3rd-order zero-phase Butterworth filter with a relatively high cutoff frequency to eliminate linear trends in all N_C electrodes. The raw signal was bandpass filtered between 1 Hz and 45 Hz. Additionally, we extracted the FC feature sets within a bandwidth $f \in N_B$, where $N_B = \lfloor F_s/2 \rfloor$, and F_s represents the EEG sampling frequency. These bandwidths were chosen to encompass physiological rhythms relevant to music appraisal in EEG paradigms, as reported in previous studies [85]. Specifically, the following frequency bands were considered: θ [4 Hz to 8 Hz], α [8 Hz to 12 Hz], and β [12 Hz to 30 Hz].

- To address artifacts, we removed data from occipital electrodes (associated with motor control) that might be highly active due to visual perception of sound stimuli after target presentation [41]. Additionally, poor occipital signals may result from insufficient electrode contact [112]. For three subjects, the impedance had outlier values ($>$ 100 kΩ), and as a result, channels O1 and O2 were excluded from further analysis.

- We re-referenced the EEG data to the common-average electrical activity measured across all scalp channels.

- The EEG responses were resampled, divided into trials, using the onset of each music stimulus as a reference point, and then downsampled at a rate of 80 Hz.

- Finally, the piecewise stationary analysis of EEG and auditory data was performed over a set of time segments with testing values [12, 6, 3, 1,5, 0,75, and 0,375] seconds, using a Hann window with 50 % overlap.

Moreover, the FC features were extracted based on **??**, where the kernel bandwidth parameter of GFC was optimized to minimize the variability of the observed data $p(X|\sigma_\phi)$ as detailed in [4]. This resulted in extracting one real-valued FC matrix of size $N_\phi \times N_\phi$ for each evaluated FC measure and subject on a single-trial basis at instant τ.

$$\widetilde{\sigma}_\phi = \arg\max \operatorname{var}\{p(X|\sigma_\phi)\}$$

The FC matrix was then vectorized to have a vector dimension $N_{\text{FC}} = N_\phi \cdot \frac{N_\phi}{2}$. Consequently, the feature vector derived from individuals (N_S) across all trials (N_R) includes a dimension of $N_{\tilde{X}}^{\lambda} = N\text{FC} \times N_\tau \times N_T \times N_S \times N_L$, extracted for each emotion label λ for the purpose of validating the supervised feature alignment. Note that the extracted EEG feature arrangement doubles in size when both FC measures are concatenated.

Audio Prepocessing

Regarding the auditory stimuli, all recordings were initially sampled at a rate of $44,100$ Hz and then segmented into N_τ sliding windows with $50\,\%$ overlap. In addition, the sampled data underwent a smoothing process by squaring and applying a convolution with a square window. To fulfill the condition specified in 3-6, the stimuli data were further downsampled to 64 Hz using cubic root compression. To ensure compatibility with the dimension of the EEG training set, the acoustic set was adjusted to a similar size, specifically $\dim(\tilde{Y}) \approx \dim(\tilde{X})$. Consequently, within each τ (time segment), we extracted the first Principal Component Analysis (PCA) component from each of the 20 acoustic features mentioned earlier [179]. This array was then augmented with $N_\phi \times 1$ samples of the acoustic envelope. As a result, we obtained a total of $N_\tau \times (20 + N_\phi \times 1)$ acoustic features within each time segment T, which would be utilized in the subsequent alignment procedure.

3.2.2. Results

Electrode Contribution

Initially, we investigate the spatial relevance of each electrode in the scalp EEG montage concerning the relationship between features extracted from neural responses and acoustic stimuli using Labeled Correlation Alignment (LCA). The *r*-squared values obtained through Canonical Correlation Analysis (CCA) after applying CKA matching are shown in Fig. **3-2** for various window intervals N_τ. The correlation estimates are averaged across the label set to provide a comprehensive interpretation. The heatmaps reveal that the correlation range varies and spreads differently across the scalp electrodes depending on the feature extraction method used. The top heatmap shows that Phase Locking Value (PLV) yields the lowest estimates within the range of [0,05-0,59], with only a few electrodes showing significant contribution. In contrast, Gaussian Functional Connectivity (GFC) extends the correlation interval to [0,05-0,73] (middle plot). Combining both measures results in correlation values of [0,10-0,74] (bottom plot), indicating that either strategy of improved FC extraction leads to distinct brain regions being linked to the acoustic stimuli.

Next, we evaluate the influence of each channel by averaging its correlation performance across all tested window intervals. The results are displayed in the matrix row for the entire EEG montage (denoted as $\{17\}$). Several electrodes exhibit negligible contribution irrespective of the extraction method used. We focus on electrodes susceptible to artifacts during music listening paradigms, particularly those associated with brain neural activity in the frontal cortex [101]. Thus, the bottom

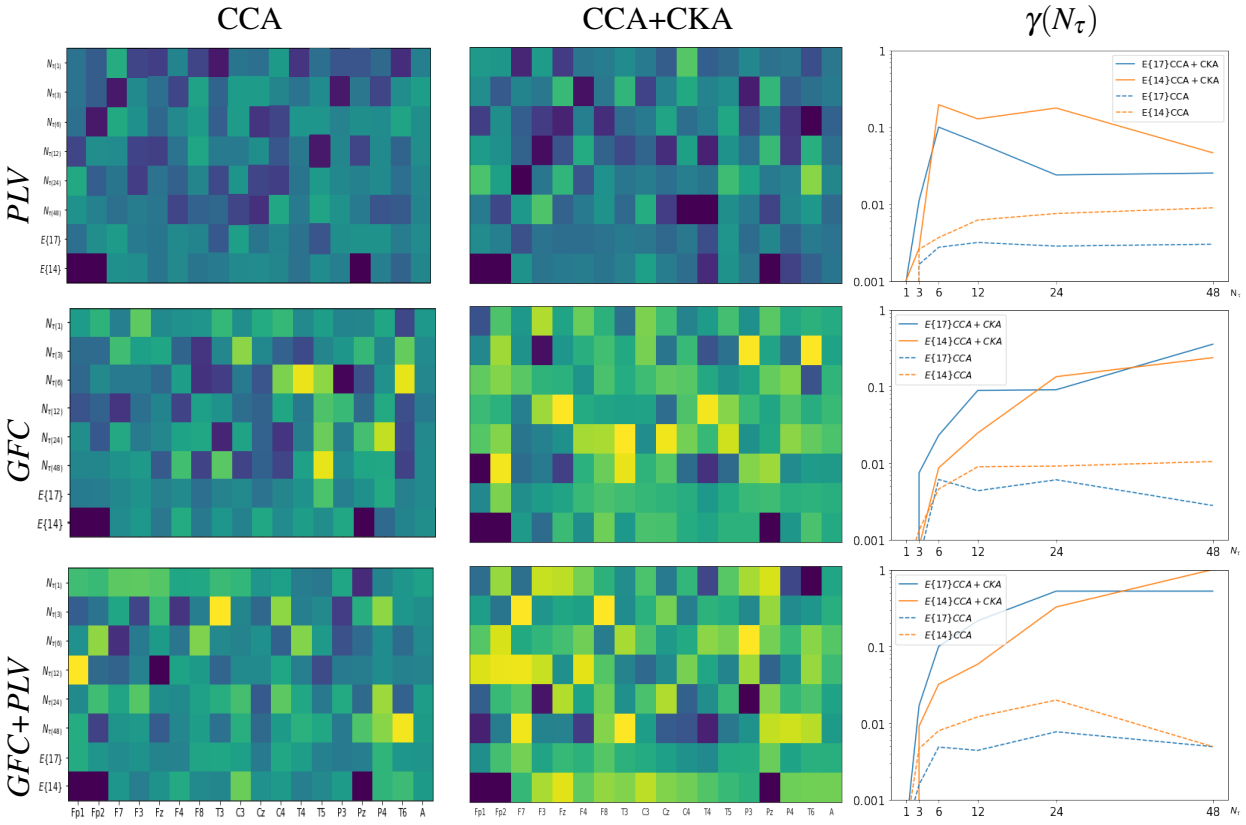


Figura 3-2: The electrode contribution of r-squared values and clustering coefficients γ (right column) obtained by the validated strategies of feature extraction: PLV (Top row), GFC (middle row), and their combination (Bottom row). Notations $\mathbb{E}\{17\}$ stands for all EEG channel signals (i.e., N_C=17) excluding O1,O2 while $\mathbb{E}\{14\}$ denotes without frontoparietal (Fp1, Fp2) and Midline Parietal (Pz) electrodes (N_C=14), respectively. The horizontal axis stands for each electrode according to the standard $10-20$ system. In the right column, the horizontal axis denotes each considered time-windowed set, N_C.

row (denoted as $\mathbb{E}\{14\}$) presents the averaged r-squared values and shows that removing Fp1, Fp2, and Pz electrodes may increase the correlation.

We further assess the discrimination ability of the selected features using the clustering coefficient γ. The right column of Fig. **3-2** illustrates the partition separability of features extracted by PLV (top plot), which is relatively modest due to the low r-squared values. In contrast, GFC leads to more distinct partitions between the extracted EEG features. The combination of GFC and PLV provides the most accurate and separable clustering performance across the tested values of the time window τ. This pattern is observed for both evaluated electrode arrangements: $N_C = 17$ (blue line) and $N_C = 14$ (orange line). Comparing this with a single CCA step (left column) that achieves significantly lower correlation values, regardless of the extraction method used, emphasizes the increased association between neural responses and acoustic stimuli achieved through LCA.

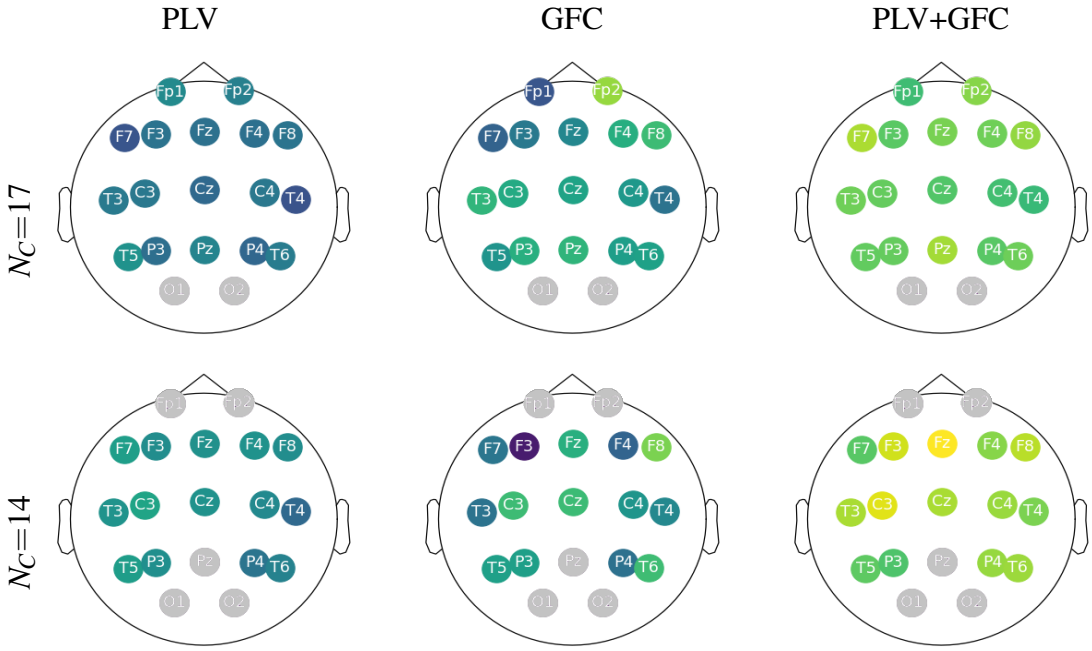


Figura 3-3: Topoplots reconstructed from LCA according to the estimated electrode relationship with the evoking auditory data. The channels affected by artifacts in gray are removed from the coupling analysis.

Lastly, for physiological interpretability, Fig. **3-3** presents the topoplots reconstructed from the FC feature sets according to the correlation with the evoking auditory data performed by LCA. The left column shows weak *r*-squared values evenly distributed over the scalp for PLV. In contrast, GFC increases the contribution from both lobes. This influence is further accentuated by combining GFC with PLV, highlighting electrodes with powerful relevance (right column) and increasing their significance in subsequent sonification stages. Notably, when removing artifact-affected electrodes, the correlation assessments focus more on the frontal and central lobes (painted yellow).

Time-windowed Correlation Estimation

In this study, we explore the impact of time-windowed feature extraction on the performance of Labeled Correlation Alignment (LCA) and examine how distinct the EEG responses remain over time, considering the potential role of changing dynamics in music creation. To explain this characteristic, the upper plot of Fig. **3-4** displays the time-varying clustering coefficient obtained at different window lengths using each feature extraction method examined in the previous section (refer to Fig. **3-2**). The scatter plots demonstrate that the labeled EEG feature partitions become more distinguishable when the window length is narrower than $\tau \leq 3$ seconds. This suggests that capturing affective neural responses with shorter overlapping time segments in feature extraction leads to more separable partitions, regardless of the FC metric. The labeled partitions of extracted

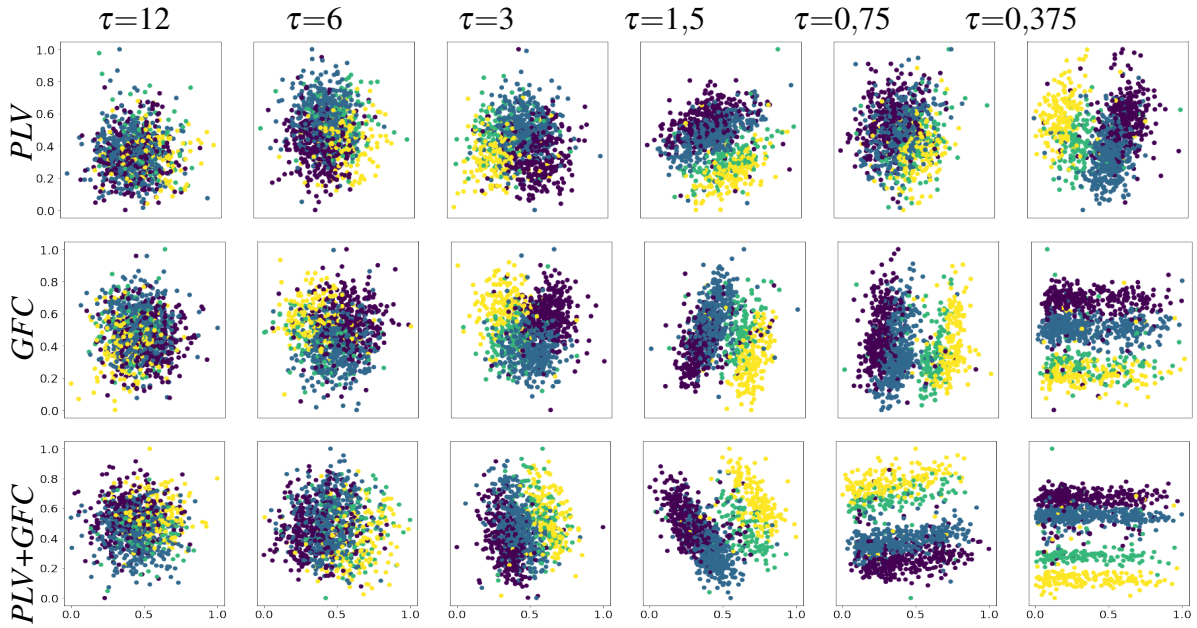
EEG dynamics differ and are more pronounced in GFC than PLV. However, the combination of GFC and PLV yields the best group separation.

Moving on, we analyze the time evolution of LCA specifically for the best strategy of FC representation, i.e., the combination of PLV and GFC. The lower plot in Fig. **3-4** presents the obtained *r*-squared values, indicating that the dynamics extracted at longer lengths ($\tau \geq 3$ seconds) are weak, resulting in intervals with almost zero-valued correlation. In contrast, the extracted features at shorter lengths ($\tau \leq 3$ seconds) exhibit stronger correlations and show fluctuations over time. Implementing the channel removal strategy improves this behavior. Additionally, the plot on the right displays the mean estimation of changes in the time-varying dynamic resolution, calculated as the difference between neighboring correlation values. The results reveal that the separability of affective labels tends to decrease as τ shortens. However, as mentioned previously, this effect can be mitigated with a proper channel selection.

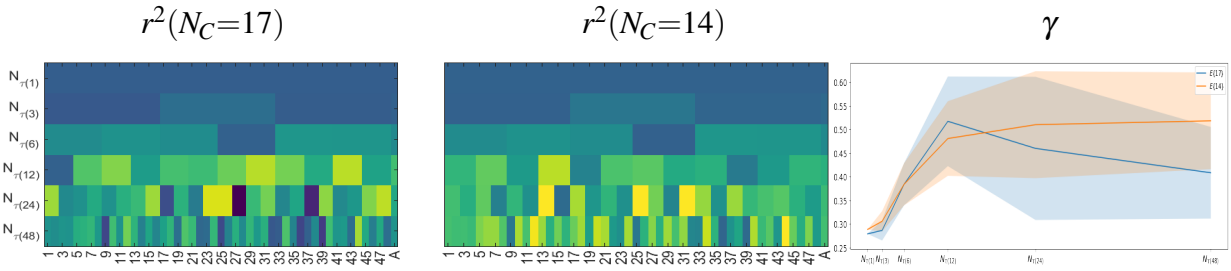
Another essential aspect discussed is the bandpass-filtered feature extraction concerning brain oscillations, which can have musical relevance. Fig. **3-5** shows the *r*-squared and clustering coefficient (γ) values obtained by combining PLV and GFC and extracted at different time windows for three brain oscillations: θ, α, and β. Filtering the lowest band (θ waveform in the blue line) causes smoother changes in the obtained time-varying dynamic resolution compared to the baseline signal encompassing all waveforms (black line). On the other hand, extracting the higher frequency rhythms (α - orange, β - green) leads to faster time-varying changes in the estimated correlation values. However, the rapid changes in *r*-squared imply that the discriminability between affective neural responses fluctuates over time.

To assess the uniformity of the group of test subjects, we analyze the performance of the Labeled Correlation Alignment (LCA) implementation individually across the channel set and at the considered time windows, which were used for feature extraction based on the combination of PLV plus GFC. In the top plot of Fig. **3-6**, we observe an appreciable discrepancy in both mean and variance values among subjects regarding the *r*-squared estimation (green line). Furthermore, a few individuals exhibit a high standard deviation, suggesting that their elicited neural responses deviate significantly from the typical responses in the subject set. To evaluate the discrimination ability that motivates the LCA algorithm, we compute the classification of affective feature sets using a Graph Convolutional Neural Network (GraphCNN) framework, similar to the approach presented in [133]. The blue line represents the calculated classifier accuracy values (mean and standard deviation). To provide a better understanding, we rank all subjects in decreasing order of their achieved mean value, revealing a significant performance gap between the best and lowest performers [7].

To further illustrate this point, we compute the heatmap of electrode contribution from the *r*-squared assessments carried out by both subjects, along with the corresponding reconstructed neural activity topoplots. As shown in the bottom plot of Fig. **3-6**, the best-performing subject (labeled as #1) exhibits a robust relationship between auditory and EEG responses, with distinct brain zones of activation. Moreover, this enhanced performance is observed even within the broadest time window. In contrast, the worst-performing subject (labeled as #27) demonstrates a very



[a] Influence of the time window τ on the performed clustering



[b] Time-resolution encoded by the extracted EEG feature sets.

Figura 3-4: Effect of time-windowed dynamics on the estimated values of r-squared. [a] Quality of clustering between labeled affective neural responses depending on the time window length τ_m measured in s. Outcomes are presented just for the removal channel configuration N_C=14 since it enhances the γ values. [b] Dynamic resolution of neural responses encoded by the extracted feature sets. The influence of both channel removal configurations is evaluated. Of note, only the method combining PLV+GFC is evaluated, and clustering is performed over the reduced set of EEG features using Principal Component Analysis separately for each affective label.

sparse correlation heatmap, suggesting a poor contribution from the central brain zone, which is assumed to be crucial in the Affective Music Listening paradigm [7].

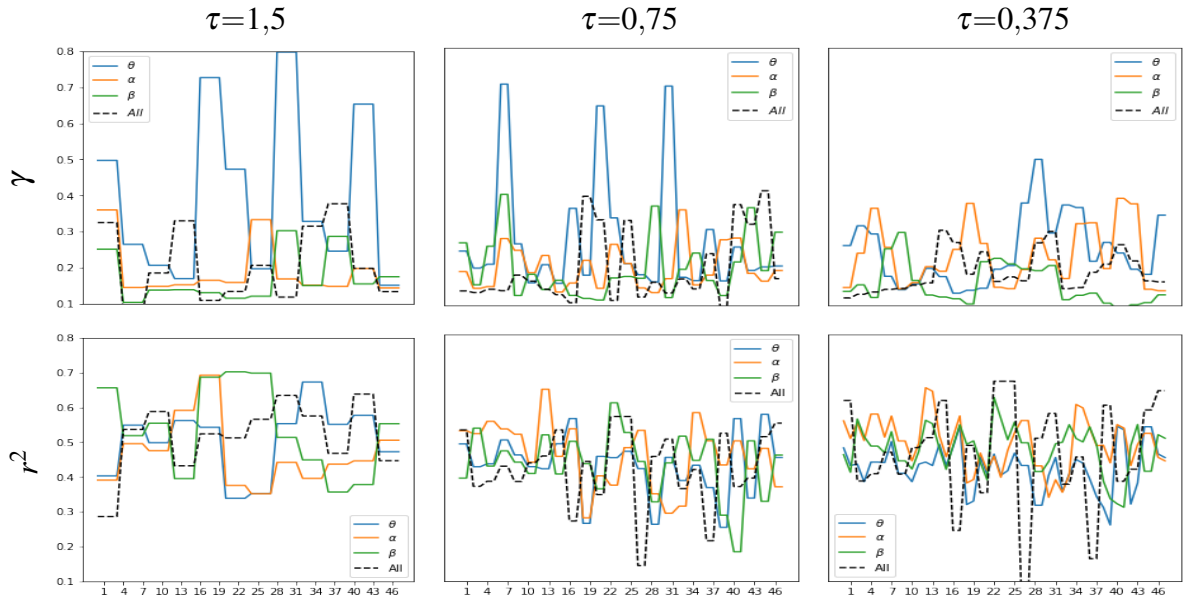
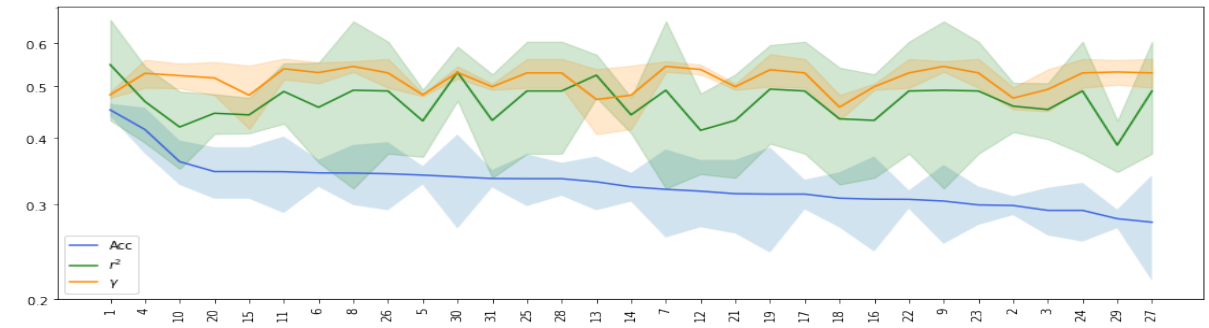


Figura 3-5: Performance variability over time conditioned by the wavebands θ, α and β. Clustering coefficient (top row) and correlation (bottom row) are estimated at short lengths of window τ using the FC extraction combining PLV+GFC. [7]

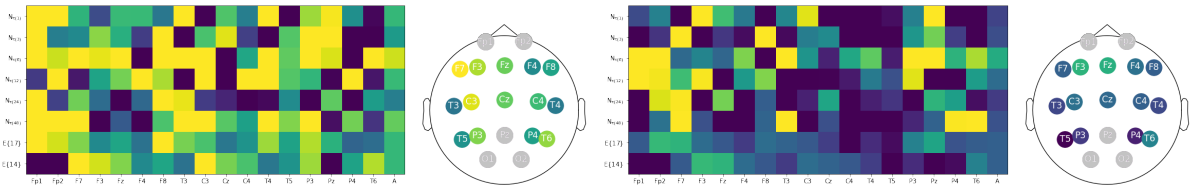
3.2.3. Generation of Affective Acoustic Envelopes

In the last part of the evaluation, we investigate the ability to create music conditioning content using brain neural activity selected by LAC. Specifically, the VQ-VAE framework in Ecuación (3-5) is trained with affective music stimuli, $\tilde{Y}$, and then applied to create auditory data by feeding the autoencoder with the most similar representation of aroused brain neural responses, $\tilde{X}$, i.e., using the model $\mu_\Lambda(\tilde{Y}|\,\tilde{X})$. Due to the highly complex music structure encoded, additional settings are required. Only the acoustic envelope is provided to the encoder as auditory training feature data, without any weighting filter (That is, $\boldsymbol{W}_{\tilde{Y}}=1$), omitting the remaining acoustic features and smoothed to decrease abrupt changes. When providing EEG data to feed the encoder input, the feature sets have an additional dimension to represent neural activity's spatial contribution. We map the EEG feature matrix into a vector representation by adding one convolutional layer to the VQ-VAE input to reduce dimension [7].

In the top row, the left plot of Figura **3-7** illustrates an example of a multichannel EEG response, followed by the extracted FC arrangement (middle plot) and applied to the Labeled Correlation Alignment, estimating the correlation assessments for feeding to the encoder. An example of the generated acoustic envelope in the output is then presented (right plot), reconstructed using VQ-VAE. The right plot illustrates how the envelope resulting from the training model $\mu_\Lambda(\tilde{Y}|\tilde{X}\,\hat{\alpha}_{\tilde{X}})$ is smooth enough (orange line). As a comparison, we show the acoustic output produced when encoding the raw EEG set directly (i.e., $\mu_\Lambda(\tilde{Y}|\tilde{X})$), showing more increased variability and abrupt changes (blue line), which tend to degrade the overall quality of the created music. In the middle



[a] Estimated values of r-squared, γ, and accuracy.



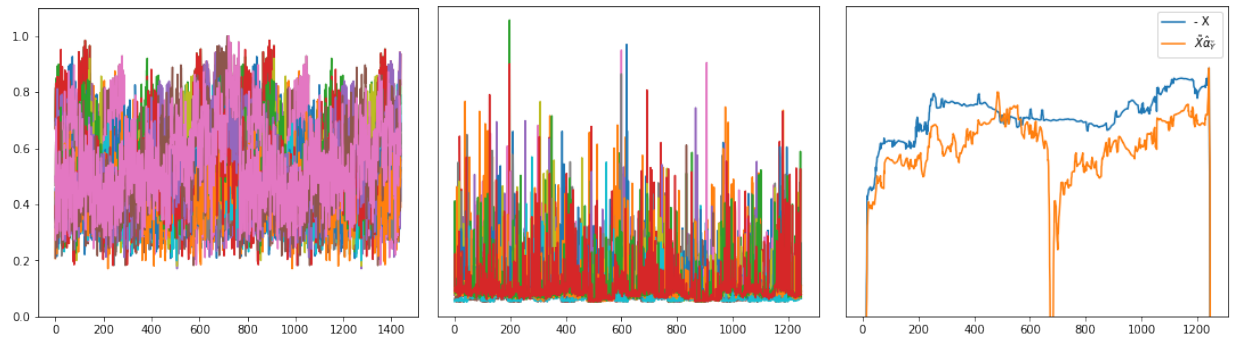
[b] Best-performing subject #1 Worst-performing subject #27

Figura 3-6: Overall subject performance of LCA. [a] Estimated values of r-squared, γ, and classifier accuracy. [b] r-squared heatmaps of electrode contribution and their reconstructed topoplots for subjects #1 and #27. Outcomes are presented for both removal channel configurations $\mathbb{E}\{17\}$ and $\{14\}$ using the FC extraction combining PLV+GFC.

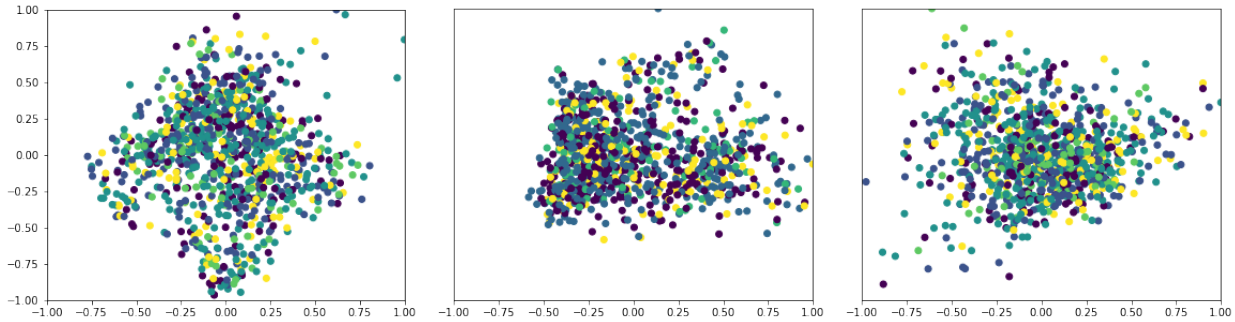
row, we show the clustering results obtained by the sets employed for training: input EEG envelopes (left plot), input FC features (center plot), and generated acoustic envelopes under the model $\mu_\Lambda(\tilde{Y}|\tilde{X}\hat{\alpha}_{\tilde{X}})$ (right plot), which show a low discriminant between affective labeled sets. On the other hand, the Labeled Correlation Alignment makes the compared input training sets distinctive.

3.3. Discussion:

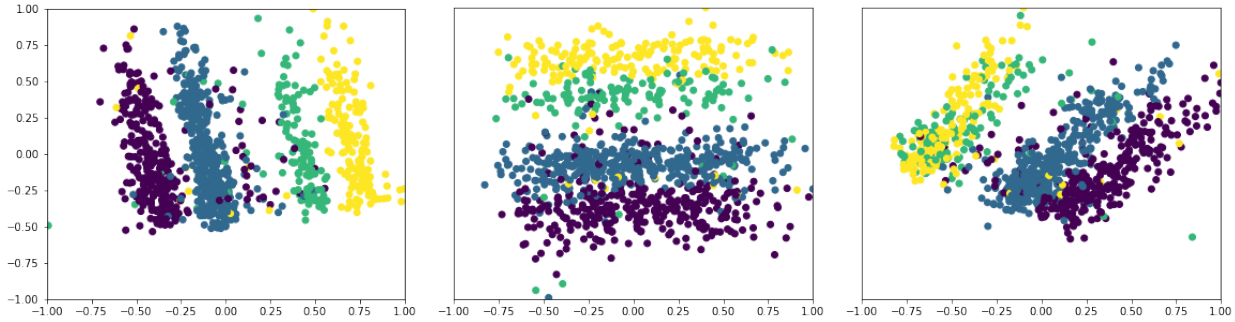
Following the evaluation stage, several noteworthy points emerge: *Feature extraction:* Gaussian Functional Connectivity, when combined with PhaseLockingValue, enhances the relationship assessment of brain activity triggered by acoustic stimuli. However, both FC measures combined provide even better associations between neural responses and acoustic inputs, suggesting that incorporating kernel-based FC can help address inter/intra-subject variability. The validation also underscores the importance of appropriately removing electrodes affected by artifacts to improve EEG feature extraction. Additionally, considering other connectivity measures like Phase-Amplitude Coupling and entropy-based FC representations, commonly used in music appraisal paradigms, may be beneficial. Regarding auditory representations, short-time acoustic envelopes have proven effective in complementing traditional acoustic feature extraction methods. These



[a] Time representation of training sets; All values are normalized for interpretation.



[b] Clustering before LAC implementation



[c] Clustering performed after LAC implementation

Figura 3-7: Sonification via VQ-VAE based on the features extracted by LAC. [a] Time representation of training sets: Input EEG recordings (left plot), extracted FC measures (central plot), and output acoustic envelopes (right plot); [b] Clustering before LAC implementation; [c] Clustering performed after LAC implementation. The illustration is given for the arisen EEG responses (left column), FC measures (central column), and created acoustic envelopes (right column).

envelopes, capturing relationships between neighboring samples, are utilized in the variational encoder network to generate low-level music synthesis. However, for more complex music structures, more sophisticated representations, such as the Musical Instrument Digital Interface format, may be required. *Labeled Correlation Alignment (LCA):* The introduced two-step procedure for aligning multimodal features with the label set is motivated by the limited association observed with

single-step Canonical Correlation Analysis. The absence of label set information in the exploration of relationships results in inadequate discrimination between affective responses. To address this, Centered Kernel Alignment is applied before Canonical Correlation Analysis to select the most relevant representations based on affective labels. The LCA approach allows for greater physiological explanation by incorporating a backward transformation within CKA to estimate the contribution of each extracted feature set. This provides insights such as the frontal and central lobes' increased relevance in the sonification stage due to the focus of correlation estimates on these areas. Moreover, short-time dynamics for narrow windows ($\tau \leq 3$ seconds) still deliver separable affective neural responses. However, bandpass-filtered feature extraction based on brain oscillations may smooth or expedite EEG dynamics, affecting discriminability between affective neural responses. Furthermore, individual differences are apparent, with varying correlation accuracy observed among different subjects. Considering the preceding findings, several factors can be taken into account to improve the alignment of multimodal features. This includes conducting group-level analysis to explore collective contributions across individuals and employing correlation methods that seek optimized projections, such as utilizing deepCCA [86].
Generation of low-level music content. an important observation is that the utilized variational autoencoder effectively generates distinct acoustic envelopes from the EEG representations selected through LCA. However, it is important to note that the current encoder network employs a discrete latent representation in combination with an autoregressive decoder specifically designed for high-quality videos, music, and speech. Consequently, further endeavors are required to address the challenges associated with discrete neural representation, potentially by adopting the predictive VQ-VAE model.

3.4. Summary

This chapter presents a novel approach to sonifying neural responses to affective music listening data using Labeled Correlation Alignment (LCA). The proposed approach addresses inter/intra-subject variability by employing a combination of Phase Locking Value and Gaussian Functional Connectivity. The two-step LCA approach first couples the input features to a set of emotion label sets using Centered Kernel Alignment. This is followed by canonical correlation analysis to select multimodal representations with stronger relationships. LCA also allows for a physiological explanation by including a backward transformation to estimate the matching contribution of each extracted brain neural feature set. Correlation estimates and partition quality are used as performance measures.
The evaluation of the proposed approach utilizes a Vector Quantized Variational AutoEncoder to create an acoustic envelope from the tested Affective Music Listening database. The validation results demonstrate the ability of the developed LCA approach to generate low-level music based on neural activity elicited by emotions while still maintaining the ability to distinguish between acoustic and acoustic outputs.

4 Symbolic music (Piano Roll) and EEG Alignment

4.1. Materials and Methods

4.1.1. Dataset: DEAP

EEG data: The evaluated database is publicly available at [1] and contains EEG recordings and peripheral physiological signals collected from thirty-two participants who viewed 40 one-minute music video excerpts. After watching each excerpt, each subject rated the video according to the four emotion conditions: arousal, valence, like/dislike, dominance, and familiarity. As detailed in [78], the EEG paradigm relies on stimuli selection using retrieval by affective tags, video highlight detection, and an online assessment tool.

From each subject, the EEG signal at 512 Hz cut-off frequency was recorded by an arrangement with 32 channels, namely, Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, O2. For implementing the emotion recognition paradigm, the employed musical stimuli were 40 one-minute music video extracts rated by 14-16 volunteers.

Audio data: This collection holds a set of music video clips selected, taken from YouTube, to evoke one of the tested mental states related to the emotion space [129]. To start, they initially picked 120 stimuli, with half being selected through a semi-automated process and the remaining half chosen manually. Next, a one-minute highlight section was designated for each stimulus. Ultimately, they employed a web-based subjective assessment experiment to narrow it down to 40 final stimuli. The acquisition protocol started with a 2-minute baseline recording of a relaxing state. Each participant listened to a one-minute audio recording during a trial. Two sessions were conducted, each holding 20 trials, and split by a break to check the quality of acquired data and electrode placement. The participant assessed his emotional levels at the end of each trial, labeled as arousal, valence, liking, and dominance.

[1] `https://www.eecs.qmul.ac.uk/mmv/datasets/deap/`

4.1.2. EEGNet:

We will consider a training space consisting of two sets $\mathscr{X}, \Lambda$, where $\mathscr{X} = \mathbf{X}_r \in \mathbb{R}^{C \times T} : r \in R$ represents EEG recordings spanning $T \in \mathbb{N}$ time instants, recorded by a C channel montage. On the other hand, $\Lambda = \boldsymbol{\lambda}_r \in [0,1]^K$ is the label set encoding the data capture of $R \in \mathbb{N}$ single trial signals performed according to the emotion recognition paradigm, with a fixed number $K \in \mathbb{N}$ of emotional states.
In deep learning architectures developed for MI classification, spatial and temporal layers process the time-series EEG input data to predict one-hot class memberships. This can be expressed as:

$$\hat{\boldsymbol{\lambda}} = \mathscr{M}\left(\xi_L \circ \cdots \circ \xi_1\right)\mathbf{X}, \tag{4-1}$$

where $\mathscr{M} : \mathbb{R}^{C \times T} \to [0,1]^K$ is the mapping function or Neural Network model that contains a layer feature map with $P_l \in \mathbb{N}$ elements at the l-th layer, given by:

$$\tilde{\mathbf{X}}l = \xi_l\left(\tilde{\mathbf{X}}l - 1 \otimes \mathbf{W}l + \mathbf{b}l\right), \quad \tilde{\mathbf{X}}l \in \mathbb{R}^{P_l},$$

where $\xi_l : \mathbb{R}^{Pl-1} \to \mathbb{R}^{P_l}$ is a learning function with a non-linear activation, $\mathbf{W}l \in \mathbb{R}^{Pl-1 \times P_l} : l \in L$ represents a set of layer weights, $\mathbf{b}_l \in \mathbb{R}^{P_l}$ is a bias term, and $L \in \mathbb{N}$ is the network depth. The notations $\circ$ and $\otimes$ denote function composition and proper tensor operations.
The estimation framework in Equation 4-1 depends on optimizing the parameter set $\Theta = \mathbf{W}_l, \mathbf{b}l$ across the trial set using the following objective function:

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}\left\{\mathscr{L}(\boldsymbol{\lambda}_r, \hat{\boldsymbol{\lambda}}_r|\Theta) + \gamma\Omega(\Theta) : \forall r \in R\right\}, \tag{4-2}$$

Where $\mathscr{L} : 0,1^K \times [0,1]^K \to \mathbb{R}$ is a loss function, $\Omega(\cdot)$ is a regularization function, and $\gamma \in \mathbb{R}^+$ is a trade-off term governing. The symbol $\mathbb{E}$ denotes the expectation operator. Equation 4-2 can be solved using mini-batch-based gradient descent and back-propagation. For the implementation of supervised feature extraction, we set $\tilde{\mathbf{X}_0} = \mathbf{X}$, $\tilde{\mathbf{X}}_L = \hat{\boldsymbol{\lambda}}$, and a sigmoid function is chosen as the non-linear activation.
Recently, a compact convolutional network called EEG-Net has emerged as a viable solution for addressing the optimization problem mentioned above in the context of EEG classification [82]. This approach offers convolutional kernel connectivity between input and output feature maps, making it applicable across various BCI paradigms. The EEGNet pipeline, shown in Figure **4-1**, begins with a temporal convolution to acquire frequency filters, followed by a depth-wise convolution connected to each feature map to learn frequency-specific spatial filters. Additionally, a separable convolution captures a temporal summary, and a point-wise convolution combines feature maps for class-membership prediction.
In order to have an output that can be directly mapped to the Arousal and Valence plane, the final layer of the EEGNet was changed to have two outputs with a sigmoid activation, with each output corresponding to one of the planes. The MSE (Mean Squared Error) cost function was used to treat the EEGNet as a regressor **4-2**.

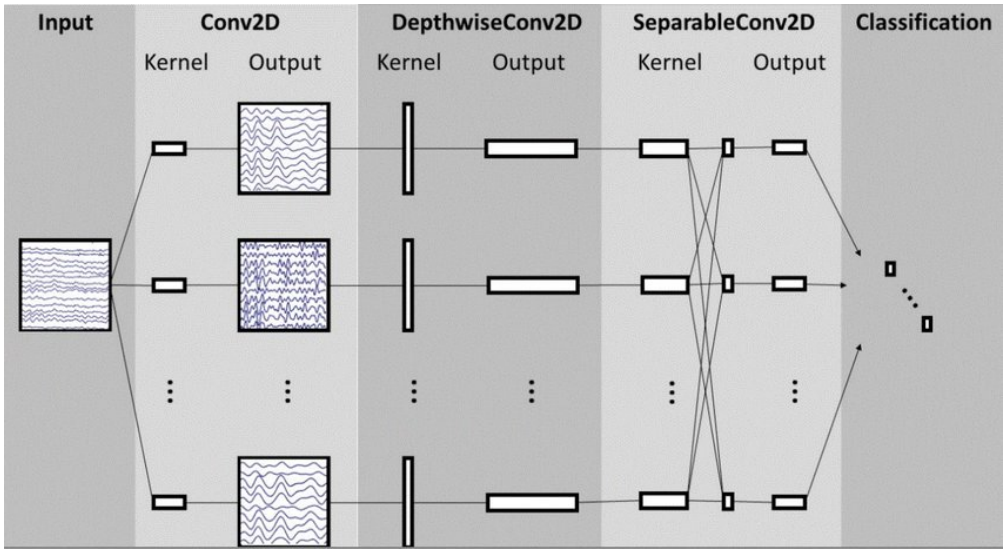


Figura 4-1: EEGNet architecture taken from the original paper [82]

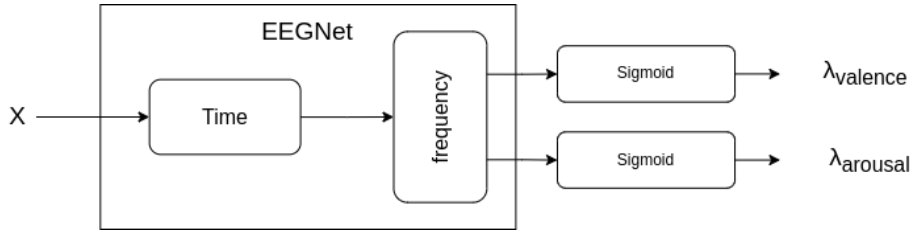


Figura 4-2: EEGNet as a regression network

4.1.3. Autoencoder with CKA Loss:

Piano Roll Autoencoder:

The CKA autoencoder was implemented to reconstruct Piano Roll samples. The problem was approached as a segmentation problem [2], in the sense that the inputs are arrays $\mathscr{Y} = \mathbf{Y}_r \in \mathbb{R}^{Ch \times T} : r \in R$ represents the Piano Roll with $T \in \mathbb{N}$ time instants and Ch the pitch notes. This array can be treated as an image, and the network's target is the mask $\mathbb{Y} \in \{0,1\}^{Ch \times T}$ in which it predicts where there is a note as a 1 and where there is not as a 0.
For the reconstruction of the arrays $\mathbf{Y}$, the Dice coefficient loss is utilized. To induce the labels $\boldsymbol{\lambda}$ in the embedded space, the CKA cost function is applied, as described below.

CKA Loss function [146]:

Let $X \subset \mathscr{X}$ and $Y \subset \mathscr{Y}$ be two pairs of random variables containing samples $x \in X$ and $y \in Y$ respectively. The kernels $\kappa_X : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ and $\kappa_Y : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ can be defined to represent nonlinear relationships among the samples through positive definite functions, resulting in:

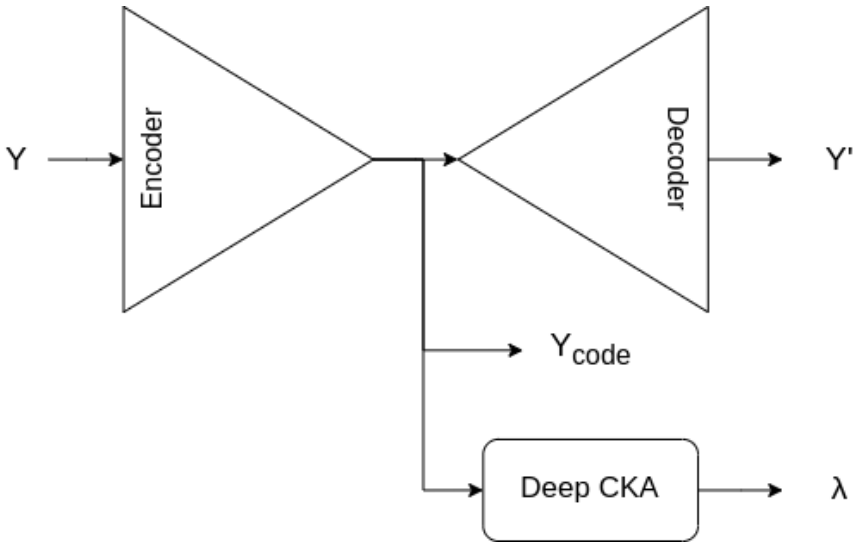


Figura 4-3: Piano Roll Autoencoder

For X:

$$\kappa_X(x,x') = \langle \phi_X(x), \phi_X(x') \rangle_{\mathscr{H}_X}, \tag{4-3}$$

For Y:

$$\kappa_Y(y,y') = \langle \phi_Y(y), \phi_Y(y') \rangle_{\mathscr{H}_Y}, \tag{4-4}$$

In practical applications, when provided with a set of input-output pairs $\{x_n \in \mathbb{R}^P, y_n \in \mathbb{R}^Q\}_{n=1}^N$, we can calculate kernel matrices K_X and K_Y as:

$$K_X[n,n'] = \kappa_X(x_n, x'_n)$$

$$K_Y[n,n'] = \kappa_Y(y_n, y'_n)$$

Next, we estimate the empirical centered kernel alignment, denoted as $\hat{\rho}_{CKA}(K_X, K_Y)$, which lies within the range of 0 to 1, using the following equation:

$$\hat{\rho}_{CKA}(K_X, K_Y) = \frac{\langle K_X^{\sim}, K_Y^{\sim} \rangle_F}{\sqrt{\|K_X^{\sim}\|_F \|K_Y^{\sim}\|_F}}, \tag{4-5}$$

Here, $\|\cdot\|_F$ represents the Frobenius norm, and $\langle \cdot, \cdot \rangle_F$ denotes the inner product. Additionally, the centered kernel matrices in this equation, $K_X^{\sim}$ and $K_Y^{\sim}$, are obtained as:

$K_X^{\sim} = HK_XH$

$K_Y^{\sim} = HK_YH$

Where H is calculated as $I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\top}$, with I_N representing the identity matrix and $\mathbf{1}_N$ as the all-one vector of the appropriate size. This alignment estimation process is a data-driven method to quantify the similarity between the random variables X and Y.

Neighbors Regression:

Using $\boldsymbol{\lambda}_r$ as the original labels for arousal and valence, and $\hat{\boldsymbol{\lambda}}$ as the EEGNet predictions, we compute the Euclidean distance $\boldsymbol{\lambda}_{dist} = \left|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\right|_2$, which then undergoes softmax activation:

$$\text{softmax}(\boldsymbol{\lambda}_{dist}) = \frac{e^{\boldsymbol{\lambda}_{dist}}}{\sum_{i=1}^{n} e^{\boldsymbol{\lambda}_{dist_i}}} \tag{4-6}$$

With $\boldsymbol{Y}_{code}$ being the output of the Piano Roll Encoder, we define $\boldsymbol{X}_{code}$ as:

$$\boldsymbol{X}_{code} = \text{softmax}(\boldsymbol{\lambda}_{dist}) \cdot \boldsymbol{Y}_{code} \tag{4-7}$$

4.2. Experimental set-up

We present an end-to-end deep learning approach to generate symbolic music (piano roll) based on EEG signals associated with induced affective states. The method is assessed within the stimulus-response paradigm and encompasses different stages, as depicted in Figure **4-4**:

- Prepossessing: The preprocessing applied to the EEG signals, the audio-to-MIDI conversion process, and the data segmentation to create the dataset used in the models' training are detailed in this description.
- Supervised Feature Extraction: The EEGNet was employed for emotion classifications, serving as an EEG feature extraction network. Additionally, an autoencoder was utilized for Piano Rolls reconstruction, generating an embedded space that encodes the input signals, representing the autoencoder features.
- Match a generation: The two pre-trained networks were connected through neighbor regression that combines the features of each network. After performing this regression, the EEG-encoded output from the EEGnet network can be used as input for the Decoder of the .Autoencoder Piano Roll"network.

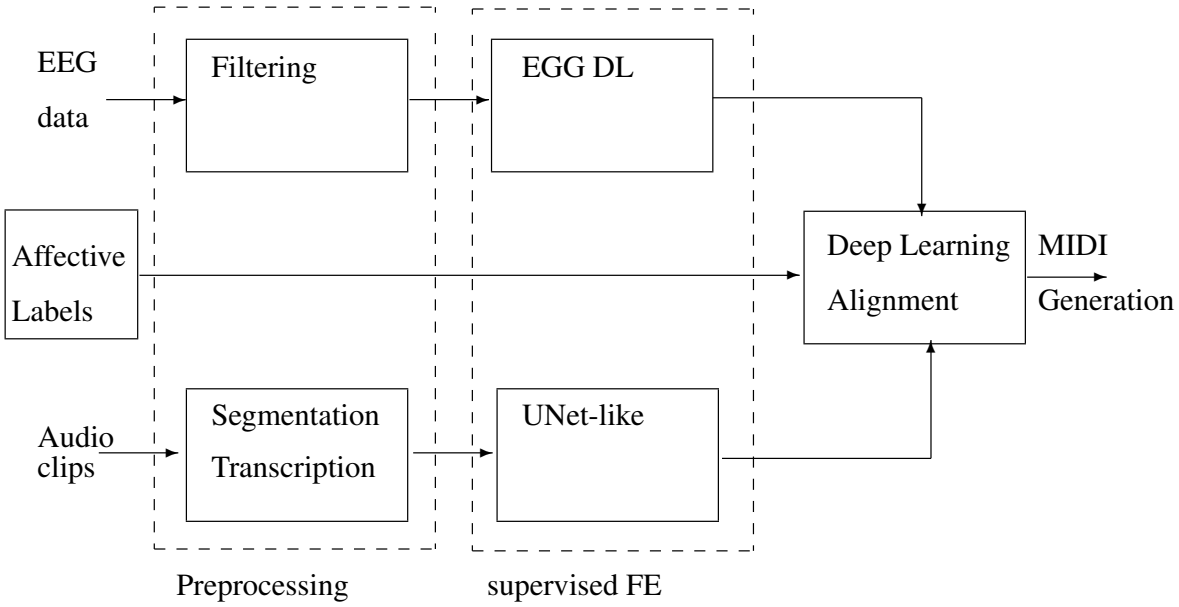


Figura 4-4: scheme of the proposed approach for MIDI generation of EEG data using deep CCA-based models with following steps: *i*) Preprocessing, *ii*) Enhanced Feature Extraction, *iii*) Supervised alignment

4.2.1. Prepossessing:

EEG DATA: As the preprocessing step, the data were downsampled to 128Hz, and the artifacts removed, including EOG, as described in [78]. Then, a band-pass frequency filter from 4-45Hz was applied. Also, the common-reference spatial filtering was performed to reduce the volume effect. Further, the filtered signals were segmented into trials lasting 60 s trials after removing a 3-second pre-trial baseline. Lastly, the EEG channels were reordered to follow the Geneva order for interpretation.

Audio data: All 40 one-minute audio clips from the DEAP dataset are transcribed into MIDI using an automatic music transcription using the Neural Network developed by Spotify's Audio Intelligence Lab [18] with a multi-output structure to improve the resulting frame-level note accuracy. Transcription is performed using the default parameters: Note segmentation is fixed to 0.5, Model Confidence Threshold $--0{,}3$, minimum note length $--11$, and midi tempo $--120$. Further, to get the suitable 2D representations needed by Convolutional DL models, we convert a music score into an image in a tensor (also known as a piano roll or pitch roll) using the package pretty-midi, as suggested in [125]. As a result, we obtain a target array in the form of a binary-valued mask by thresholding the input data.

For evaluation, each one-minute trial was split into ten non-overlapped music and EEG data segments, each lasting six seconds. Note that because we aim at capturing more musical information, we make this interval longer than the four seconds used in [46]. Besides, nine audio partitions with

fade-out endings were removed, yielding 391 suitable testing segments instead of 400, then we ended up working with an array from (400 x 64 x 128).

4.2.2. Supervised Feature Extraction

EEG Features: Because deep learning models with different connective structures may result in distinct sets of extracted EEG features, we evaluate three deep learning models widely used for EEG-Based Brain-Computer Interfaces [82, 107, 132]: EEGnet is a compact convolutional neural network combining depthwise and separable convolutions; TCFussionnet is a fixed hyperparameter-based CNN model that utilizes multiple techniques, such as temporal convolutional networks, separable convolution, depth-wise convolution, and the fusion of layers; and Deep & Shallow ConvNet is a deep convolutional network.
Table **4-1** shows the parameter set-up for the tested EEGNet model, including the used filters F1, F2, and the Depth Multiplier (D). Network architecture is described in detail in [2] and omitted due to its extension.

Tabla 4-1: Detailed EEGnet architecture for MI classification

Layer	*Output Dimension*	*Parameters*
Input Layer	$N_c \times N_t \times 1$	·
Conv2D	$N_c \times N_t \times F_1$	Temporal filter $(F_1) = 4$, $Kernelsize = (1,4)$ $Padding = same$,$Bias = False$
BatchNormalization	·	·
DepthwiseConv2D	$1 \times N_t \times 16$	Depth Multiplier $(D) = 2$,$kernelsize = (N_c, 1)$ $Bias = False$
BatchNormalization	·	·
Activation	·	$Activation = ELU$
AveragePooling2d	$1 \times 32 \times 16$	Pool $size = (1,4)$
Dropout	·	$DropoutRate = 0{,}6$
SeparableConv2D	$1 \times 32 \times 32$	Temporal filter $(F_2) = 32$, $Kernelsize = (1,16)$ $Padding = same$, $Bias = False$
batch Normalization	·	·
activation	·	Activation = ELU
AveragePooling 2d	$1 \times 4 \times 32$	Pool size=(1, 8)
Dropout	·	Dropout Rate = 0.6
flatten	128	·
Dense	*units*	units = $Nclasses$
Activation	*units*	$Activation = Softmax$

[2] publicly available at `https://github.com/hdperezn/Tesis_codes.git`

Music Features: On the other hand, MIDI segmentation is performed as suggested in [47], using a variation of the Unet neural networks termed autoencoder piano roll. This network has architecture parameters shown in Table **4-2**. The programming code for implementing the autoencoder piano roll can be accessed at [3], where the residual layer is not connected sequentially, but to the Add layer.

4.2.3. Match Networks and generation:

Once the EEGNet was trained as a regression model and the piano roll autoencoder using CKA loss was trained, a test data partition was selected for EEGNet prediction. The EEGNet predictions underwent a softmax activation, as shown in Equation 4-6. From this resulting array, the top 5 highest values were extracted. Finally, the product was computed as demonstrated in Equation 4-7. This new array, denoted as $\boldsymbol{X}_{code}$, derived from the training data, is used as input in the decoder of the Piano Roll Autoencoder.

4.2.4. Results

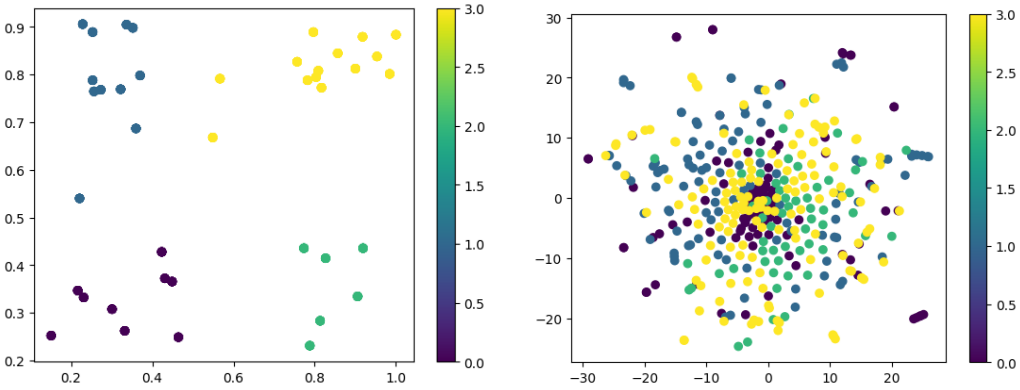
Symbolic music data labels exploration:

Initially, we investigated whether the Piano Roll representation of the musical stimuli by itself was separated following any relationship with the subjects' labels. In Figure **4-5**, part (a) shows how Subject #1 labeled the audio stimuli, with arousal on the X-axis and valence on the Y-axis. Part (b) of the figure plots the two-dimensional reduction ($ncomponents = 2$ and $perplexity = 5$) using t-SNE of the Piano Roll arrays. In both figures, colors represent the class of each audio stimulus, and (c) shows the bottleneck of the Piano Roll Autoencoder with CKA loss, the dots represent the training data, and the xs are the test data.

EEG emotion classification

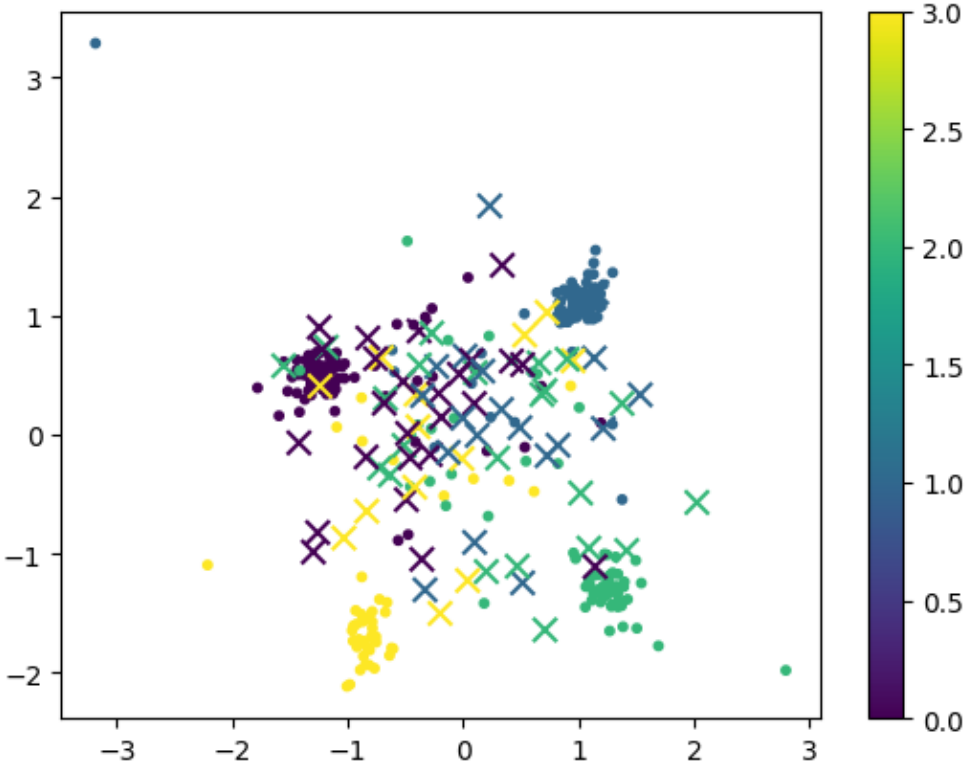
Figure **4-6** illustrates the accuracy results obtained by EEGNET when using EEG signals as input and subject-specific labels as targets. The reported training and testing data are acquired through a 5-fold cross-validation approach, with 80 % of the signals used for training and 20 % for testing. Two class classifications are performed, one for arousal and the other for valence. Additionally, a four-label multiclass paradigm is implemented by combining the labels from valence and arousal. The upper figure was sorted in descending order based on the accuracies in valence and is represented by the orange line. At the same time, the bars indicate to what extent the multiclass accuracy is better or worse for each subject. Similarly, the same procedure is applied to the arousal feature in the lower figure.
The best parameters for the EEGNetThe EEGNet as a classifier training where $kernelLength = 128, F1 = 4, D = 4, F2 = 32, normrate = 0{,}5$ and the drop type was "Dropout"with $DropoutRate =$

[3] publicly available at `https://github.com/hdperezn/EEG_PianoRoll`



(a) Emotion labels set by Sub #1

(b) MIDI TSNE clustering



(c) Embedding space autoencoder + CKA

Figura 4-5: (a) figure is the audio labeled by subject, figure (b) is the piano roll cluster by TSN and (c) shows the embedding space of the Autoencoder Piano Roll with the CKA loss.

Tabla 4-2: Autoencoder Piano Roll - Encoder

Layer	*Output Dimension*	*Parameters*
InputLayer	$N_p \times N_t \times 1$	·
Conv2D	$\frac{N_p}{2} \times \frac{N_t}{2} \times F_1$	$F_1 = 32$, $Kernelsize = 1$ $Padding = same$, $strides = 2$
BatchNormalization	·	·
Activation	·	Activation = relu
Residual(Conv2D)	$\frac{N_p}{4} \times \frac{N_t}{4} \times F_2$	$F_2 = 64$, $Kernelsize = 3$ $Padding = same$, $strides = 2$
Dropout	·	Dropout Rate = 0.5
Activation	·	Activation = relu
SeparableConv2D	$\frac{N_p}{2} \times \frac{N_t}{2} \times F_2$	$F_2 = 64, Kernelsize = 3$ $Padding = same, Bias = False$
BatchNormalization	·	·
Activation	·	Activation = relu
SeparableConv2D	$\frac{N_p}{2} \times \frac{N_t}{2} \times F_2$	$F_2 = 64, Kernelsize = 3$ $Padding = same, Bias = False$
BatchNormalization	·	·
MaxPooling2D	$\frac{N_p}{4} \times \frac{N_t}{4} \times F_2$	$poolsize = 3, strides = 2$
Add[MaxPooling2D, Residual]	·	·
Residual(Conv2D)	$\frac{N_p}{8} \times \frac{N_t}{8} \times F_3$	$F_2 = 128, Kernelsize = 3$ $Padding = same, strides = 2$
Dropout	·	Dropout $Rate = 0{,}5$
Activation	·	Activation = relu
SeparableConv2D	$\frac{N_p}{4} \times \frac{N_t}{4} \times F_3$	$F_2 = 128, Kernelsize = 3$ $Padding = same, Bias = False$
BatchNormalization	·	·
Activation	·	Activation = relu
SeparableConv2D	$\frac{N_p}{4} \times \frac{N_t}{4} \times F_3$	$F_2 = 128, Kernelsize = 3$ $Padding = same, Bias = False$
BatchNormalization	·	·
MaxPooling2D	$\frac{N_p}{8} \times \frac{N_t}{8} \times F_3$	$poolsize = 3, strides = 2$
Add[MaxPooling2D, Residual]	·	·

0,6. In addition to validating our performance with respect to the state of the art, this network was used to refine the time windows in which the classification gave a higher probability. For the rest of this experiment, the MIDI stimulus windows were used where the emotion classification gave a probability greater than 70 %.

Finally, to predict the emotion labels as a range between 0 and 1, the classification layer was

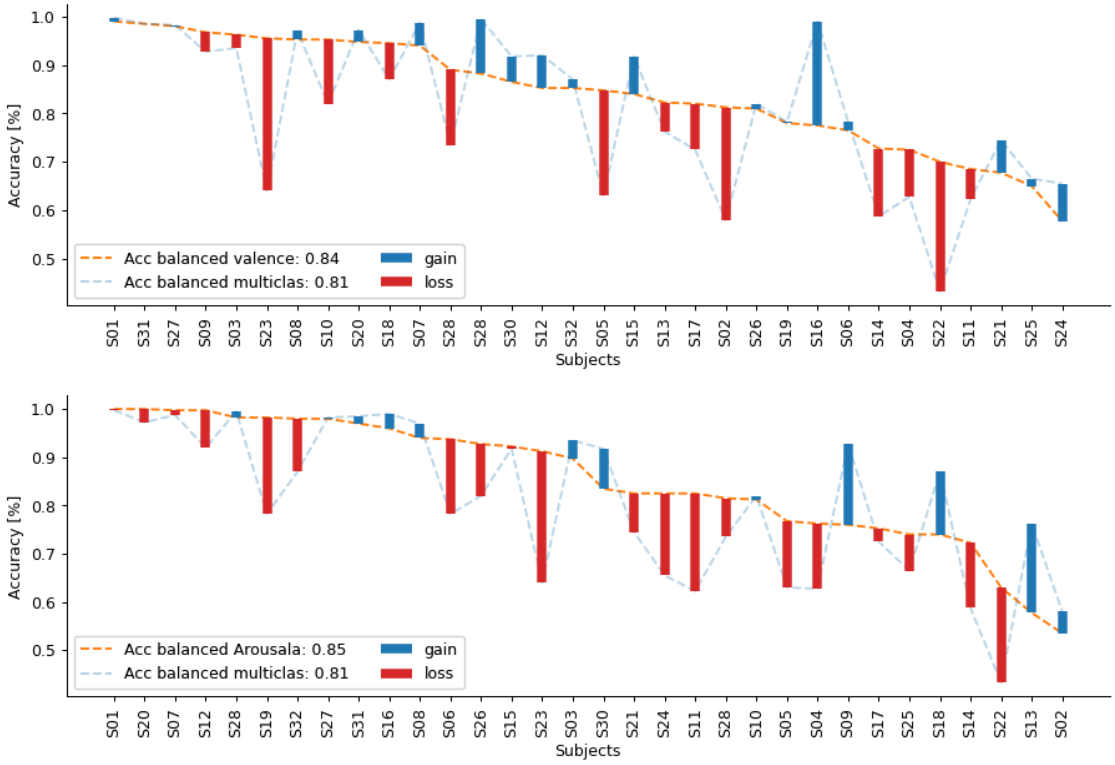
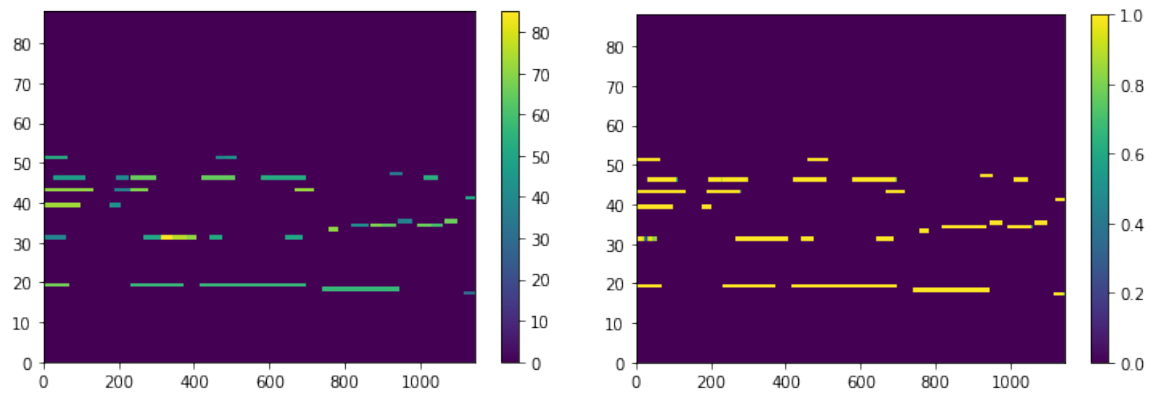


Figura 4-6: Accuracy curves from the emotion recognition paradigm. The upper figure compares the subject-wise accuracy values in multiclass classification versus binary classification for Valence. The lower figure compares the multiclass accuracy with binary classification for Arousal. In both figures, the subjects are sorted in descending order based on their binary classification accuracy.

changed in the network to two regression outputs, one for arousal and the other for valence. As a prediction of network two, this gives us a pair of coordinates that can be mapped in the arousal and valence plane.

Auto-encoder Piano Roll

Figure **4-7** depicts the reconstruction ability of the PianoRoll Autoencoder network for the input data. Panel (a) shows one of the training samples, while panel (b) displays the corresponding prediction generated by the network. The Dice coefficient in the test phase reached 0,902, the recall achieved 0,8911, and the specificity sensitivity attained 0,9988.
Additionally, Figure **4-5** (c) shows the bottleneck of this network. It is important to note how the reconstruction loss plus the CKA loss allows the network to perform well in reconstructing the piano roll arrays. At the same time, The embedded space, the Encoder prediction, maintains an inevitable separation between classes.
Classes separate this embedded space according to the emotion, and the EEGNet prediction in the



(a) Piano Roll sample (b) Piano roll reconstruction

Figura 4-7: (a) Piano Roll sample from the training dataset (b) Is the corresponding reconstruction performed by the Autoencoder PianoRoll

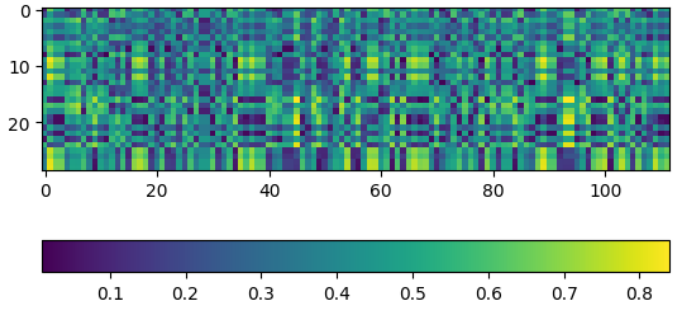
plane of arousal and valence allows us to later make a match between these two spaces.

Match and Generation

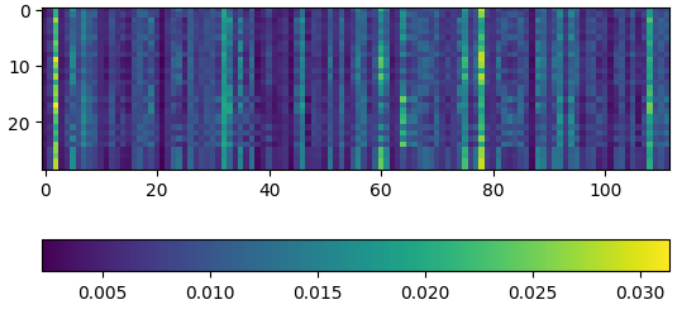
Figure **4-8** shows the distance between $\hat{\boldsymbol{\lambda}}$ predict with test data and $\boldsymbol{\lambda}$ labels of train data for the subject #1. (a) the figure is the single quadratic distance, (b) is the softmax operation applied to the distance, and (c) shows the top 5 highest closest distances for each piano roll sample. The networks was trained using $5 - fold$ cross-validation with an $80/20$ data split
A matrix like the figure in the figure is generated for each of the five folds, so five different predictions were generated for each training. It is important to remember that for each fold, the data generated with the test set never looked at the train set, even though all the data has been reviewed in the cross-validation.

Generation

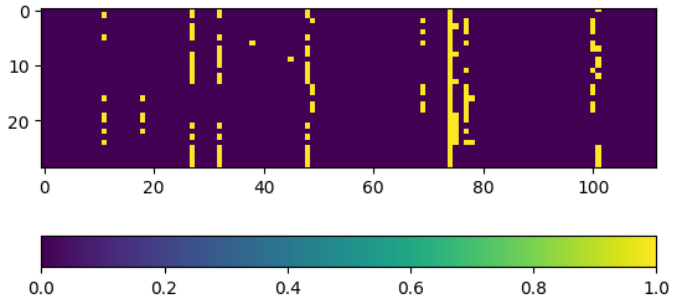
The arrangement that was calculated in figure **4-8** (c) with the test data is the one that, according to equation 4-7, multiplies the prediction of the encoder of the Autoencoder Piano Roll network with the Test data. This resulting matrix will be used as input for the Decoder of the Autoencoder Piano Roll network, and its prediction will result in a Piano Roll array corresponding to each EEG trial used as input.
In figure **4-9**, four characteristics were computed (pitch count, pitch range, average inter-onset-interval, total pitch class transition histogram) for the MIDI signals following the approach in [165]. Their intraset metrics methodology was also adopted to evaluate generative models. The blue line in each figure represents the measurement of a specific characteristic in the training data (it is a constant line because all subjects listened to the same stimuli). At the same time, the orange bars show the performance of the dataset generated for each subject. The subjects were ordered in



(a) Distance matrix



(b) Softmax distance



(c) Top 5 closest distances

Figura 4-8: Distance matrices of EEGNet-regressor predictions

descending order according to the MSE they had in the EEGNet regression.

4.3. Discussion

Subsequently, during the evaluation phase, the following key observations should be highlighted: *Reprocessing and Data*: Concerning the databases, there was a need for a larger dataset since, as observed during training, a set of 400 samples may not be sufficient to utilize large generative models like transformers effectively. The lack of a widely used public multimodal database that incorporates symbolic, audio, and EEG data was also noticed.

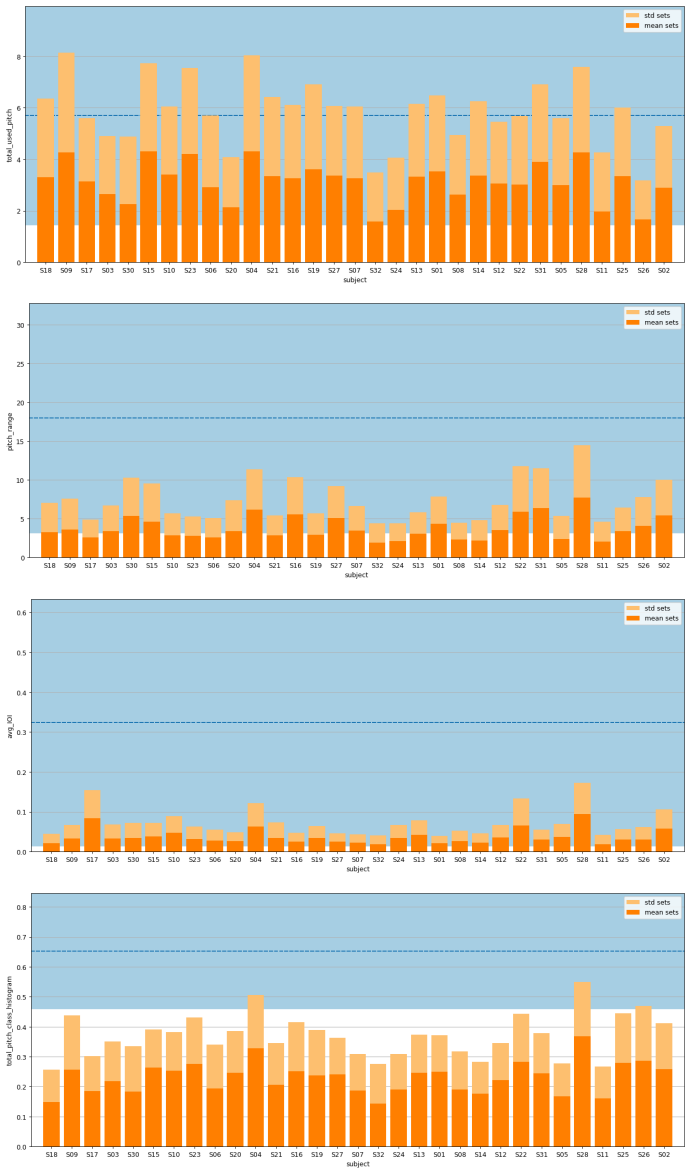


Figura 4-9: Evaluation Features Generated MIDI

Supervised Feature Extraction: In the emotion classification paradigm, we used the EEGNet, as shown in Figure **4-6**, significant accuracy values were achieved, matching or even surpassing some of the state-of-the-art results (see [47]). Although our main goal was not specifically classification, this accuracy demonstrates that the network extracts the necessary features to identify the subject's emotions based on their neural activity. Meanwhile, the Autoencoder PianoRoll achieved outstanding results in its task of reconstructing symbolic musical representations **4-7**. Furthermore, when using the CKA loss in conjunction with the autoencoder **4-3**, the features of the network were separated into different classes, indicating that some label information can indeed be observed in the training data Piano Roll.

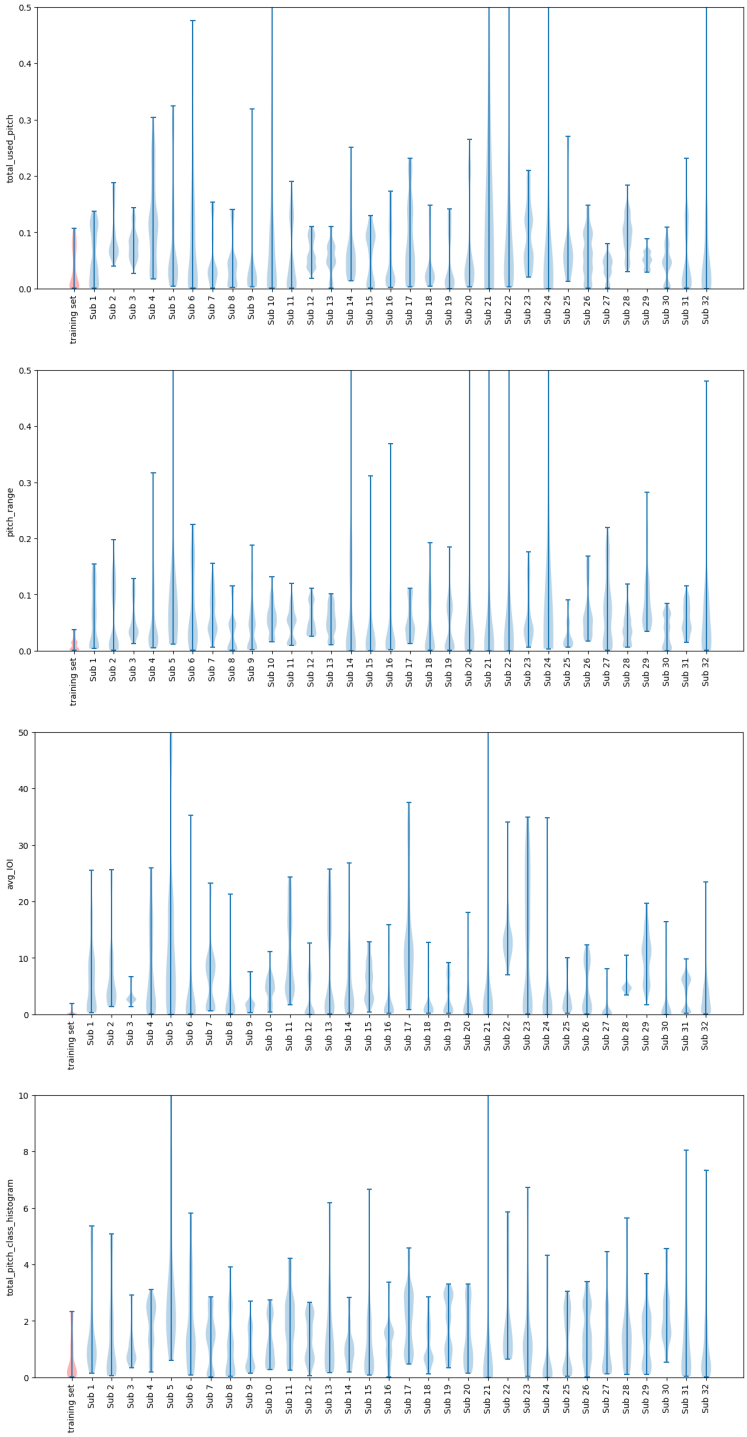


Figura 4-10: Visualization of model characteristics through the PDFs, all subjects fold 1

Match Network: Correlation based on distances to the nearest samples has proven to be an effective technique for matching both spaces. Previously, more complex models were attempted; however,

due to the limited data, finding a tool that did not suffer from overfitting issues became necessary. One of the main disadvantages of this technique is that if two samples share the same neighbors, the resulting generated array will be very similar for these two samples, as each generation is a consequence of the linear combination of embedded codes from the most similar samples. Furthermore, it is crucial to understand that generative models depend entirely on the training data quality. For this reason, MIDI metrics are compared with the same training dataset. This implies that datasets generated by each subject will face the same challenges and deficiencies as the training set. This has a significant impact on databases like this one, where we have limited data, and they were not designed initially as MIDI stimuli; instead, we had to transform raw audio to obtain Piano Roll arrangements.

MIDI Metrics: In the case of the first two characteristics, pitch count and pitch range, the expectation was that the ideal generative model would closely mirror the mean of the training data (represented by the blue line) and maintain a similar variance as indicated by the shaded area. However, accurate generative models will inevitably exhibit discrepancies compared to the training set. We achieved competent results in our case, suggesting that our models do not deviate significantly from the mean regarding the range of generated notes and the number of notes. Nevertheless, the lower standard deviation implies a reduced variety of notes in our compositions.
Concerning rhythm-based features, such as the Average Inter-Onset-Interval, our generated sets did not quite capture the intricacies of the training set's dynamics. This suggests that our generative models faced challenges in rhythm aspects.
The final metric summarizes the models' performance concerning pitch. We have included this figure to illustrate the metrics in which the model competes effectively.

5 Conclusions and Future Work

5.1. Conclusions

- In this study, we propose a cross-subject transfer learning approach to enhance the classification accuracy of elicited neural responses. Our method involves pooling data from labeled EEG measurements and psychological questionnaires using a stepwise multi-space kernel embedding. To validate our approach, we implement the transfer learning within a Deep&Wide framework, pairing the source-target sets based on the BCI inefficiency. The results demonstrate that our method significantly improves the classifier performance for most target individuals when using single or multiple sources.

- In this chapter, a method is presented to convert neural responses from affective music listening data into sound. The approach utilizes Labeled Correlation Alignment (LCA) to identify EEG features that are most congruent with auditory data, based on a given set of emotions. LCA involves two main steps: Supervised Centered Kernel Alignment (CKA)-based feature selection and Canonical Correlation Analysis (CCA)-based analysis. The results obtained from the validation on real-world data demonstrate the effectiveness of the LCA approach in generating low-level music content using neural activity associated with specific emotions, while also being able to distinguish between the produced acoustic envelopes

- We have made significant progress in developing an end-to-end deep learning methodology for generating symbolic music (MIDI) content directly from EEG. Although the model displayed promising potential, it fell short of achieving competitive metrics in some of our rhythm-based evaluations of music quality. To further enhance the music-generation process, we are actively working on refinements and improvements.

5.2. Future work

- In the future, in relation with the multimodality with structured data, we intend to validate the cross-subject transfer learning approach in applications that involve the combination of two or more databases, thereby increasing the number of individuals tested significantly. For example, we aim to include the dataset collected by the Department of Brain and Cognitive Engineering, Korea University in [83], as it contains valuable questionnaire data related to the physiological and psychological condition of subjects. This will enable us to assess

classification performances based on transfer learning at both the intra-subject and inter-dataset levels.

- In future work, we plan to address the necessity of creating a dedicated dataset tailored specifically for EEG-based symbolic music generation. This dataset will encompass a wide range of musical styles, complexities, and emotional nuances, enhancing the training and evaluation of generative models.
- Additionally, we aim to further enhance the End-to-End methodology for symbolic music generation through EEG. To achieve this, we will explore alternative metrics for matching between networks, striving to improve model performance and achieve satisfactory metrics in our evaluations.

5.3. Academic products

5.3.1. Academic papers:

- Alvarez-Meza, A. M., Torres-Cardona, H. F., Orozco-Alzate, M., Perez-Nastar, H. D., & Castellanos-Dominguez, G. (2023). Affective Neural Responses Sonified through Labeled Correlation Alignment. Sensors, 23(12), 5574.
- Collazos-Huertas, D. F., Velasquez-Martinez, L. F., Perez-Nastar, H. D., Alvarez-Meza, A. M., & Castellanos-Dominguez, G. (2021). Deep and wide transfer learning with kernel matching for pooling data from electroencephalography and psychological questionnaires. Sensors, 21(15), 5105.

5.3.2. Others:

- Software registration N: 13-95-219, ÜNET-LIKE FOR PIANO ROLL"
- National invention patent NC2022-0005981, "Sistema y método para generación de piezas musicales basado en respuestas eléctricas cerebrales y técnicas de compocición musical"

Bibliografía

[1] ADOLPHS, Ralph ; ANDERSON, David: The neuroscience of emotion: A new synthesis. Princeton University Press, 2018

[2] AGUIRRE-ARANGO, Juan C. ; ÁLVAREZ-MEZA, Andrés M. ; CASTELLANOS-DOMINGUEZ, German: Feet Segmentation for Regional Analgesia Monitoring Using Convolutional RFF and Layer-Wise Weighted CAM Interpretability. En: Computation 11 (2023), Nr. 6, p. 113

[3] ALARCAO, Soraia M. ; FONSECA, Manuel J.: Emotions recognition using EEG signals: A survey. En: IEEE Transactions on Affective Computing 10 (2017), Nr. 3, p. 374–393

[4] ALVAREZ-MEZA, A ; CARDENAS-PENA, D ; CASTELLANOS-DOMINGUEZ, G: Unsupervised kernel function building using maximization of information potential variability. En: Iberoamerican Congress on Pattern Recognition Springer, 2014, p. 335–342

[5] ALVAREZ-MEZA, A ; OROZCO-GUTIERREZ, A ; CASTELLANOS-DOMINGUEZ, G: Kernel-based relevance analysis with enhanced interpretability for detection of brain activity patterns. En: Frontiers in neuroscience 11 (2017), p. 550

[6] ALVAREZ-MEZA, A. M. ; OROZCO-GUTIERREZ, A. ; CASTELLANOS-DOMINGUEZ, G.: Kernel-Based Relevance Analysis with Enhanced Interpretability for Detection of Brain Activity Patterns. En: Frontiers in Neuroscience 11 (2017), p. 550. – ISSN 1662–453X

[7] ÁLVAREZ-MEZA, Andrés M. ; TORRES-CARDONA, Héctor F. ; OROZCO-ALZATE, Mauricio ; PÉREZ-NASTAR, Hernán D. ; CASTELLANOS-DOMINGUEZ, German: Affective Neural Responses Sonified through Labeled Correlation Alignment. En: Sensors 23 (2023), Nr. 12, p. 5574

[8] ANDREW, G ; ARORA, R ; BILMES, J ; LIVESCU, K: Deep canonical correlation analysis. En: International conference on machine learning PMLR, 2013, p. 1247–1255

[9] ANOWAR, F. ; SADAOUI, S. ; SELIM, B.: Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). En: Computer Science Review 40 (2021), p. 100378. – ISSN 1574–0137

[10] ANTOL, Stanislaw ; AGRAWAL, Aishwarya ; LU, Jiasen ; MITCHELL, Margaret ; BATRA, Dhruv ; ZITNICK, C L. ; PARIKH, Devi: Vqa: Visual question answering. En: Proceedings of the IEEE international conference on computer vision, 2015, p. 2425–2433

[11] APPRIOU, Aurelien ; CICHOCKI, Andrzej ; LOTTE, Fabien: Modern machine-learning algorithms: for classifying cognitive and affective states from electroencephalography signals. En: IEEE Systems, Man, and Cybernetics Magazine 6 (2020), Nr. 3, p. 29–38

[12] ASGHAR, Muhammad A. ; KHAN, Muhammad J. ; FAWAD ; AMIN, Yasar ; RIZWAN, Muhammad ; RAHMAN, MuhibUr ; BADNAVA, Salman ; MIRJAVADI, Seyed S.: EEG-based multi-modal emotion recognition using bag of deep features: An optimal feature selection approach. En: Sensors 19 (2019), Nr. 23, p. 5218

[13] DE AZEVEDO SANTOS, L R. ; SILLA JR, Carlos N. ; COSTA-ABREU, MD: A methodology for procedural piano music composition with mood templates using genetic algorithms. (2021)

[14] BAGHERZADEH, S ; MAGHOOLI, K ; SHALBAF, A ; MAGHSOUDI, A: Recognition of emotional states using frequency effective connectivity maps through transfer learning approach from electroencephalogram signals. En: Biomedical Signal Processing and Control 75 (2022), p. 103544

[15] BAHMANI, M ; BABAK, M ; LAND, W ; HOWARD, J ; DIEKFUSS, J ; ABDOLLAHIPOUR, R: Children's motor imagery modality dominance modulates the role of attentional focus in motor skill learning. En: Human movement science 75 (2020), p. 102742

[16] BASSO, J ; SATYAL, M ; RUGH, R: Dance on the Brain: Enhancing Intra- and Inter-Brain Synchrony. En: Frontiers in Human Neuroscience 14 (2021), p. 586

[17] BHATTACHARJEE, M ; MAHADEVA, P ; GUHA, P: Time-Frequency Audio Features for Speech-Music Classification. En: ArXiv (2018), p. 1–5

[18] BITTNER, Rachel M. ; BOSCH, Juan J. ; RUBINSTEIN, David ; MESEGUER-BROCAL, Gabriel ; EWERT, Sebastian: A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. En: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2022, p. 781–785

[19] BODNER, Mark ; MUFTULER, L T. ; NALCIOGLU, Orhan ; SHAW, Gordon L.: FMRI study relevant to the Mozart effect: brain areas involved in spatial–temporal reasoning. En: Neurological research 23 (2001), Nr. 7, p. 683–690

[20] BRIOT, J ; PACHET, F: Music Generation by Deep Learning - Challenges and Directions. En: ArXiv 1712.04371 (2017), p. 1–17

[21] BRIOT, Jean-Pierre ; HADJERES, Gaëtan ; PACHET, François-David: Deep learning techniques for music generation–a survey. En: arXiv preprint arXiv:1709.01620 (2017)

[22] BRIOT, Jean-Pierre ; HADJERES, Gaëtan ; PACHET, François-David: Deep learning techniques for music generation. Vol. 1. Springer, 2020

[23] CARDONA, L ; VARGAS-CARDONA, H ; NAVARRO, P ; CARDENAS PEÑA, D ; OROZCO GUTIÉRREZ, A: Classification of Categorical Data Based on the Chi-Square Dissimilarity and t-SNE. En: Computation 8 (2020), Nr. 4. – ISSN 2079–3197

[24] CHEN, Yu-An ; WANG, Ju-Chiang ; YANG, Yi-Hsuan ; CHEN, Homer: Linear regression-based adaptation of music emotion recognition models for personalization. En: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2014, p. 2149–2153

[25] DE CHEVEIGNÉ, A ; WONG, D ; DI LIBERTO, G ; HJORTKJAER, J ; SLANEY, M ; LALOR, E: Decoding the auditory brain with canonical component analysis. En: NeuroImage 172 (2018), p. 206–216

[26] CHO, H ; AHN, M ; AHN, S ; KWON, M ; JUN, S: EEG datasets for motor imagery brain-computer interface. En: GigaScience 6 (2017), 05, Nr. 7

[27] CICCARELLI, G ; NOLAN, M ; PERRICONE, J ; CALAMIA, P ; HARO, S ; O'SULLIVAN, J ; MESGARANI, N ; QUATIERI, T ; SMALT, C: Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. En: Scientific reports 9 (2019), Nr. 1, p. 1–10

[28] COLE, Ross: The problem with AI music: song and cyborg creativity in the digital age. En: Popular Music 39 (2020), Nr. 2, p. 332–338

[29] COLLAZOS-HUERTAS, D ; ALVAREZ-MEZA, A ; CASTELLANOS-DOMINGUEZ, G: Image-Based Learning Using Gradient Class Activation Maps for Enhanced Physiological Interpretability of Motor Imagery Skills. En: Applied Sciences 12 (2022), Nr. 3, p. 1695

[30] COLLAZOS-HUERTAS, D. ; CAICEDO-ACOSTA, J. ; CASTAÑO DUQUE, G. A. ; ACOSTA-MEDINA, C. D.: Enhanced Multiple Instance Representation Using Time-Frequency Atoms in Motor Imagery Classification. En: Frontiers in Neuroscience 14 (2020), p. 155

[31] COLLAZOS-HUERTAS, DF ; ÁLVAREZ-MEZA, AM ; ACOSTA-MEDINA, CD ; CASTAÑO-DUQUE, GA ; CASTELLANOS-DOMINGUEZ, G: CNN-based framework using spatial dropping for enhanced interpretation of neural activity in motor imagery classification. En: Brain Informatics 7 (2020), Nr. 1, p. 1–13

[32] COLLAZOS-HUERTAS, D.F. ; ÁLVAREZ-MEZA, A.M. ; CASTELLANOS-DOMINGUEZ, G.: Spatial interpretability of time-frequency relevance optimized in motor imagery discrimination using Deep&Wide networks. En: Biomedical Signal Processing and Control 68 (2021), p. 102626. – ISSN 1746–8094

[33] COLLAZOS-HUERTAS, Diego F. ; VELASQUEZ-MARTINEZ, Luisa F. ; PEREZ-NASTAR, Hernan D. ; ALVAREZ-MEZA, Andres M. ; CASTELLANOS-DOMINGUEZ, German: Deep and wide transfer learning with kernel matching for pooling data from electroencephalography and psychological questionnaires. En: Sensors 21 (2021), Nr. 15, p. 5105

[34] COLLET, C. ; HAJJ, M. E. ; CHAKER, Rawad ; BUI-XUAN, B. ; LEHOT, J. ; HOYEK, N.: Effect of motor imagery and actual practice on learning professional medical skills. En: BMC Medical Education 21 (2021)

[35] CUI, Xu ; WU, Yongrong ; WU, Jipeng ; YOU, Zhiyu ; XIAHOU, Jianbing ; OUYANG, Menglin: A review: Music-emotion recognition and analysis based on EEG signals. En: Frontiers in Neuroinformatics 16 (2022), p. 997282

[36] DADEBAYEV, Didar ; GOH, Wei W. ; TAN, Ee X.: EEG-based emotion recognition: Review of commercial EEG devices and machine learning techniques. En: Journal of King Saud University-Computer and Information Sciences 34 (2022), Nr. 7, p. 4385–4401

[37] DAI, C ; WANG, Z ; WEI, L ; CHEN, G ; CHEN, B ; ZUO, F ; LI, Y: Combining early post-resuscitation EEG and HRV features improves the prognostic performance in cardiac arrest model of rats. En: The American Journal of Emergency Medicine 36 (2018), Nr. 12, p. 2242–2248. – ISSN 0735–6757

[38] DAI, Shuqi ; YU, Huiran ; DANNENBERG, Roger B.: What is missing in deep music generation? a study of repetition and structure in popular music. En: arXiv preprint arXiv:2209.00182 (2022)

[39] DALY, I ; NICOLAOU, N ; WILLIAMS, D ; HWANG, F ; KIRKE, A ; MIRANDA, E ; NASUTO, S: Neural and physiological data from participants listening to affective music. En: Scientific Data 7 (2020), Nr. 1, p. 1–7

[40] DAS, Abhishek ; KOTTUR, Satwik ; GUPTA, Khushi ; SINGH, Avi ; YADAV, Deshraj ; MOURA, José MF ; PARIKH, Devi ; BATRA, Dhruv: Visual dialog. En: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, p. 326–335

[41] DAS, P ; GUPTA, S ; NEOGI, B: Measurement of effect of music on human brain and consequent impact on attentiveness and concentration during reading. En: Procedia Computer Science 172 (2020), p. 1033–1038

[42] DASH, Adyasha ; AGRES, Kat R.: Ai-based affective music generation systems: a review of methods, and challenges. En: arXiv preprint arXiv:2301.06890 (2023)

[43] DAVIS, Steven ; MERMELSTEIN, Paul: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. En: IEEE transactions on acoustics, speech, and signal processing 28 (1980), Nr. 4, p. 357–366

[44] DHARIWAL, P ; JUN, H ; PAYNE, C ; KIM, J ; RADFORD, A ; SUTSKEVER, I: Jukebox: A generative model for music. En: arXiv:2005.00341 (2020), p. 1–20

[45] DI-LIBERTO, G ; MARION, G ; SHAMMA, S: The Music of Silence: Part II: Music Listening Induces Imagery Responses. En: Journal of Neuroscience 41 (2021), Nr. 35, p. 7449–7460

[46] DING, Yi ; ROBINSON, Neethu ; ZHANG, Su ; ZENG, Qiuhao ; GUAN, Cuntai: Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. En: arXiv preprint arXiv:2104.02935 (2021)

[47] DING, Yi ; ROBINSON, Neethu ; ZHANG, Su ; ZENG, Qiuhao ; GUAN, Cuntai: Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. En: IEEE Transactions on Affective Computing (2022)

[48] DONAHUE, C ; MAO, H ; LI, Y ; COTTRELL, G ; MCAULEY, J: LakhNES: Improving Multi-instrumental Music Generation with Cross-domain Pre-training. En: ISMIR, 2019, p. 1–8

[49] DUBUS, G ; BRESIN, R: A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities. En: PLoS ONE 8 (2013)

[50] DÃ¤HNE, S ; BIEÃŸMANN, F ; SAMEK, W ; HAUFE, S ; GOLTZ, D ; GUNDLACH, C ; VILLRINGER, A ; FAZLI, S ; MÃ¼LLER, K: Multivariate Machine Learning Methods for Fusing Multimodal Functional Neuroimaging Data. En: Proceedings of the IEEE 103 (2015), Nr. 9, p. 1507–1530

[51] EBCIOĞLU, Kemal: An expert system for harmonizing four-part chorales. En: Computer Music Journal 12 (1988), Nr. 3, p. 43–51

[52] FERREIRA, Lucas N. ; WHITEHEAD, Jim: Learning to generate music with sentiment. En: arXiv preprint arXiv:2103.06125 (2021)

[53] FIEBRINK, Rebecca ; CARAMIAUX, Baptiste: The machine learning algorithm as creative musical tool. En: arXiv preprint arXiv:1611.00379 (2016)

[54] FLEURY, Mathis ; FIGUEIREDO, Patrícia ; VOURVOPOULOS, Athanasios ; LÉCUYER, Anatole: Two is better? Combining EEG and fMRI for BCI and Neurofeedback: A systematic review. (2023)

[55] FREER, D ; YANG, G: Data augmentation for self-paced motor imagery classification with C-LSTM. En: Journal of neural engineering 17 (2020), Nr. 1, p. 016041

[56] FURUI, Sadaoki: Speaker-independent isolated word recognition based on emphasized spectral dynamics. En: ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 11 IEEE, 1986, p. 1991–1994

[57] GARCIA-MURILLO, D ; ALVAREZ-MEZA, A ; CASTELLANOS-DOMINGUEZ, G: Single-Trial Kernel-Based Functional Connectivity for Enhanced Feature Extraction in Motor-Related Tasks. En: Sensors 21 (2021), Nr. 8

[58] GOMEZ, Patrick ; DANUSER, Brigitta: Relationships between musical structure and psychophysiological measures of emotion. En: Emotion 7 (2007), Nr. 2, p. 377

[59] HE, Qun ; FENG, Lufeng ; JIANG, Guoqian ; XIE, Ping: Multimodal multitask neural network for motor imagery classification with EEG and fNIRS signals. En: IEEE Sensors Journal 22 (2022), Nr. 21, p. 20695–20706

[60] HE, Zhipeng ; LI, Zina ; YANG, Fuzhou ; WANG, Lei ; LI, Jingcong ; ZHOU, Chengju ; PAN, Jiahui: Advances in multimodal emotion recognition based on brain–computer interfaces. En: Brain sciences 10 (2020), Nr. 10, p. 687

[61] HERNANDEZ-OLIVAN, Carlos ; BELTRAN, Jose R.: Music composition with deep learning: A review. En: Advances in speech and music technology: computational aspects and applications (2022), p. 25–50

[62] HERREMANS, D ; CHUAN, C ; CHEW, E: A Functional Taxonomy of Music Generation Systems. En: ACM Computing Surveys (CSUR) 50 (2017), p. 1–30

[63] HILDT, E: Affective Brain-Computer Music Interfaces –Drivers and Implications. En: Frontiers in Human Neuroscience 15 (2021)

[64] HOUSSEIN, Essam H. ; HAMMAD, Asmaa ; ALI, Abdelmgeid A.: Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. En: Neural Computing and Applications 34 (2022), Nr. 15, p. 12527–12557

[65] HUANG, Chih-Fang ; HUANG, Cheng-Yuan: Emotion-based AI music generation system with CVAE-GAN. En: 2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE) IEEE, 2020, p. 220–222

[66] HUI, K ; GANAA, E ; ZHAN, Y ; SHEN, X: Robust deflated canonical correlation analysis via feature factoring for multi-view image classification. En: Multimedia Tools and Applications 80 (2021), Nr. 16, p. 24843–24865

[67] HUNG, Hsiao-Tzu ; CHING, Joann ; DOH, Seungheon ; KIM, Nabin ; NAM, Juhan ; YANG, Yi-Hsuan: Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. En: arXiv preprint arXiv:2108.01374 (2021)

[68] HUTCHINGS, Patrick E. ; MCCORMACK, Jon: Adaptive music composition for games. En: IEEE Transactions on Games 12 (2019), Nr. 3, p. 270–280

[69] JAMES, C ; ZUBER, S ; DUPUIS LOZERON, E ; ABDILI, L ; GERVAISE, D ; KLIEGEL, M: How Musicality, Cognition and Sensorimotor Skills Relate in Musically Untrained Children. En: Swiss Journal of Psychology 79 (2020), Nr. 3-4, p. 101–112

[70] JEON, E ; KO, W ; YOON, J ; SUK, H. Mutual Information-driven Subject-invariant and Class-relevant Deep Representation Learning in BCI. 2020

[71] JI, Shulei ; YANG, Xinyu ; LUO, Jing: A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. En: ACM Computing Surveys (2023)

[72] JUSLIN, P ; VASTFJLL, D: Emotional responses to music: The need to consider underlying mechanisms. En: Behavioral and brain sciences 31 (2008), Nr. 5, p. 559–575

[73] KANT, P ; LASKAR, S ; HAZARIKA, J ; MAHAMUNE, R: CWT Based Transfer Learning for Motor Imagery Classification for Brain computer Interfaces. En: Journal of Neuroscience Methods 345 (2020), p. 108886. – ISSN 0165–0270

[74] KATTHI, J ; GANAPATHY, S: Deep Correlation Analysis for Audio-EEG Decoding. En: IEEE Trans Neural Syst Rehabil Eng 29 (2021), p. 2742–2753

[75] KINGMA, D ; WELLING, M: An Introduction to Variational Autoencoders. En: Foundations and Trends in Machine Learning 12 (2019), Nr. 4, p. 307–392

[76] KO, W ; JEON, E ; JEONG, S ; SUK, H. Multi-Scale Neural network for EEG Representation Learning in BCI. 2020

[77] KOCTÚROVÁ, M ; JUHÁR, J: A Novel approach to EEG Speech activity detection with visual stimuli and mobile BCI. En: Applied Sciences 11 (2021), Nr. 2, p. 674

[78] KOELSTRA, Sander ; MUHL, Christian ; SOLEYMANI, Mohammad ; LEE, Jong-Seok ; YAZDANI, Ashkan ; EBRAHIMI, Touradj ; PUN, Thierry ; NIJHOLT, Anton ; PATRAS, Ioannis: Deap: A database for emotion analysis; using physiological signals. En: IEEE transactions on affective computing 3 (2011), Nr. 1, p. 18–31

[79] KÜHL, N ; GOUTIER, M ; HIRT, R ; SATZGER, G: Machine Learning in Artificial Intelligence: Towards a Common Understanding. En: HICSS, 2019, p. 1–10

[80] KUMAR, S ; SHARMA, A ; TSUNODA, T: Brain wave classification using long short-term memory network based OPTICAL predictor. En: Scientific Reports 9 (2019), 12, p. 1–13

[81] LADDA, A ; LEBON, F ; LOTZE, M: Using motor imagery practice for improving motor performance - A review. En: Brain and Cognition 150 (2021), p. 105705

[82] LAWHERN, Vernon J. ; SOLON, Amelia J. ; WAYTOWICH, Nicholas R. ; GORDON, Stephen M. ; HUNG, Chou P. ; LANCE, Brent J.: EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. En: Journal of neural engineering 15 (2018), Nr. 5, p. 056013

[83] LEE, M ; KWON, O ; KIM, Y ; KIM, H ; LEE, Y ; WILLIAMSON, J ; FAZLI, S ; LEE, S: EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. En: GigaScience 8 (2019), 01, Nr. 5. – ISSN 2047–217X

[84] LEE, M ; YOON, J ; LEE, S: Predicting Motor Imagery Performance From Resting-State EEG Using Dynamic Causal Modeling. En: Frontiers in Human Neuroscience 14 (2020), p. 321. – ISSN 1662–5161

[85] LEIPOLD, S ; GREBER, M ; SELE, S o. ; JÄNCKE, L: Neural patterns reveal single-trial information on absolute pitch and relative pitch perception. En: NeuroImage 200 (2019), p. 132–141

[86] LI, C ; WANG, B ; ZHANG, S ; LIU, Y ; SONG, R ; CHENG, J ; CHEN, X: Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism. En: Comput. Biol. Med 143 (2022), p. 105303

[87] LI, Xiang ; SONG, Dawei ; ZHANG, Peng ; ZHANG, Yazhou ; HOU, Yuexian ; HU, Bin: Exploring EEG features in cross-subject emotion recognition. En: Frontiers in neuroscience 12 (2018), p. 162

[88] LI, Xiang ; ZHANG, Yazhou ; TIWARI, Prayag ; SONG, Dawei ; HU, Bin ; YANG, Meihong ; ZHAO, Zhigang ; KUMAR, Neeraj ; MARTTINEN, Pekka: EEG based emotion recognition: A tutorial and review. En: ACM Computing Surveys 55 (2022), Nr. 4, p. 1–57

[89] LI, Xiaowei ; HU, Bin ; ZHU, Tingshao ; YAN, Jingzhi ; ZHENG, Fang: Towards affective learning with an EEG feedback approach. En: Proceedings of the first ACM international workshop on Multimedia technologies for distance learning, 2009, p. 33–38

[90] LIEBMAN, E ; STONE, P: Artificial Musical Intelligence: A Survey. En: ArXiv 2006.10553 (2020)

[91] LIOI, G ; CURY, C ; PERRONNET, L ; MANO, M ; BANNIER, E ; LÉCUYER, A ; BARILLOT, C: Simultaneous MRI-EEG during a motor imagery neurofeedback task: an open access brain imaging dataset for multi-modal data integration. En: bioRxiv (2019), p. 862375

[92] LONG, Y ; KONG, W ; JIN, X ; SHANG, J ; YANG, C: Visualizing Emotional States: A Method Based on Human Brain Activity. En: ZENG, A (Ed.) ; PAN, D (Ed.) ; HAO, T (Ed.) ; ZHANG, D (Ed.) ; SHI, Y (Ed.) ; SONG, X (Ed.): Human Brain and Artificial Intelligence, 2019, p. 248–258

[93] LOUI, P: Neuroscience of Musical Improvisation. En: Handbook of Artificial Intelligence for Music. Springer, 2021, p. 97–115

[94] MAMMONE, N ; IERACITANO, C ; MORABITO, F: A deep CNN approach to decode motor preparation of upper limbs from timeâ€“frequency maps of EEG signals at source level. En: Neural Networks 124 (2020), p. 357–372. – ISSN 0893–6080

[95] MARTIN, Rod A. ; BERRY, Glen E. ; DOBRANSKI, Tobi ; HORNE, Marilyn ; DODGSON, Philip G.: Emotion perception threshold: Individual differences in emotional sensitivity. En: Journal of Research in Personality 30 (1996), Nr. 2, p. 290–305

[96] MCAVINUE, L ; ROBERTSON, I: Measuring motor imagery ability: A review. En: European Journal of Cognitive Psychology 20 (2008), Nr. 2, p. 232–251

[97] MCFARLAND, D. ; MINER, L. ; VAUGHAN, T. ; WOLPAW, J: Mu and Beta Rhythm Topographies During Motor Imagery and Actual Movements. En: Brain Topography 12 (2004), p. 177–186

[98] MILAZZO, M ; BUEHLER, B: Designing and fabricating materials from fire using sonification and deep learning. En: iScience 24 (2021), Nr. 8, p. 102873

[99] MIRZAEI, S ; GHASEMI, P: EEG motor imagery classification using dynamic connectivity patterns and convolutional autoencoder. En: Biomedical Signal Processing and Control 68 (2021), p. 102584. – ISSN 1746–8094

[100] MISHRA, S ; ASIF, M ; TIWARY, U: Dataset on Emotions using Naturalistic Stimuli (DENS). En: bioRxiv (2021), p. 1–13

[101] MIYAMOTO, K ; TANAKA, H ; NAKAMURA, S: Emotion Estimation from EEG Signals and Expected Subjective Evaluation. En: 2021 9th International Winter Conference on Brain-Computer Interface (BCI) IEEE, 2021, p. 1–6

[102] MIYAMOTO, Kana ; TANAKA, Hiroki ; NAKAMURA, Satoshi: Online EEG-based emotion prediction and music generation for inducing affective states. En: IEICE TRANSACTIONS on Information and Systems 105 (2022), Nr. 5, p. 1050–1063

[103] MOORE, F R.: The dysfunctions of MIDI. En: Computer music journal 12 (1988), Nr. 1, p. 19–28

[104] MORI, K: Decoding peak emotional responses to music from computational acoustic and lyrical features. En: Cognition 222 (2022), p. 105010

[105] MOU, Luntian ; ZHAO, Yiyuan ; HAO, Quan ; TIAN, Yunhan ; LI, Juehui ; LI, Jueying ; SUN, Yiqi ; GAO, Feng ; YIN, Baocai: Memomusic version 2.0: Extending personalized music recommendation with automatic music generation. En: 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) IEEE, 2022, p. 1–6

[106] MUHAMED, A ; LI, L ; SHI, X ; YADDANAPUDI, S ; CHI, W ; JACKSON, D ; SURESH, R ; LIPTON, Z ; SMOLA, A: Symbolic Music Generation with Transformer-GANs. En: Proceedings of the AAAI Conference on Artificial Intelligence Vol. 31, 2021, p. 408–417

[107] MUSALLAM, Yazeed K. ; ALFASSAM, Nasser I. ; MUHAMMAD, Ghulam ; AMIN, Syed U. ; ALSULAIMAN, Mansour ; ABDUL, Wadood ; ALTAHERI, Hamdi ; BENCHERIF, Mohamed A. ; ALGABRI, Mohammed: Electroencephalography-based motor imagery classification using temporal convolutional network fusion. En: Biomedical Signal Processing and Control 69 (2021), p. 102826

[108] NAFEA, Mohamed S. ; ISMAIL, Zool H.: Supervised machine learning and deep learning techniques for epileptic seizure recognition using EEG signalsâ€"A systematic literature review. En: Bioengineering 9 (2022), Nr. 12, p. 781

[109] NATSIOU, A ; O'LEARY, S: Audio representations for deep learning in sound synthesis: A review. En: ArXiv 2201.02490 (2022), p. 1–8

[110] NIRANJAN, D ; BURUNAT, I ; TOIVIAINEN, P ; ALLURI, V: Influence of musical expertise on the processing of musical features in a naturalistic setting. En: Conference on Cognitive Computational Neuroscience, 2019, p. 655–658

[111] NORDSTRÖM, Henrik ; LAUKKA, Petri: The time course of emotion recognition in speech and music. En: The Journal of the Acoustical Society of America 145 (2019), Nr. 5, p. 3058–3074

[112] ORLANDI, S ; HOUSE, S ; KARLSSON, P ; SAAB, R ; CHAU, T: Brain-Computer Interfaces for Children With Complex Communication Needs and Limited Mobility: A Systematic Review. En: Frontiers in Human Neuroscience 15 (2021)

[113] PANDEY, P ; AHMAD, N ; MIYAPURAM, K ; LOMAS, D: Predicting Dominant Beat Frequency from Brain Responses While Listening to Music. En: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) IEEE, 2021, p. 3058–3064

[114] PAPADOPOULOS, Alexandre ; ROY, Pierre ; PACHET, François: Assisted lead sheet composition using flowcomposer. En: Principles and Practice of Constraint Programming: 22nd International Conference, CP 2016, Toulouse, France, September 5-9, 2016, Proceedings 22 Springer, 2016, p. 769–785

[115] PARCALABESCU, Letitia ; TROST, Nils ; FRANK, Anette: What is multimodality? En: arXiv preprint arXiv:2103.06304 (2021)

[116] PATNAIK, Suprava: Speech emotion recognition by using complex MFCC and deep sequential model. En: Multimedia Tools and Applications 82 (2023), Nr. 8, p. 11897–11922

[117] PEKSA, Janis ; MAMCHUR, Dmytro: State-of-the-Art on Brain-Computer Interface Technology. En: Sensors 23 (2023), Nr. 13, p. 6001

[118] PICARD, Rosalind ; SU, David ; LIU, Yan: AMAI: Adaptive music for affect improvement. (2018)

[119] PICARD, Rosalind W.: Building HAL: Computers that sense, recognize, and respond to human emotion. En: Human Vision and Electronic Imaging VI Vol. 4299 SPIE, 2001, p. 518–523

[120] PICARD, Rosalind W.: Affective computing: challenges. En: International Journal of Human-Computer Studies 59 (2003), Nr. 1-2, p. 55–64

[121] PLUMMER, Bryan A. ; WANG, Liwei ; CERVANTES, Chris M. ; CAICEDO, Juan C. ; HOCKENMAIER, Julia ; LAZEBNIK, Svetlana: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. En: Proceedings of the IEEE international conference on computer vision, 2015, p. 2641–2649

[122] PODOBNIK, B ; STANLEY, H: Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. En: Physical review letters 100 (2008), Nr. 8, p. 084102

[123] PURWINS, H ; LI, B ; VIRTANEN, T ; SCHLÜTER, J ; CHANG, S ; SAINATH, T: Deep learning for audio signal processing. En: IEEE Journal of Selected Topics in Signal Processing 13 (2019), Nr. 2, p. 206–219

[124] PURWINS, Hendrik ; BLANKERTZ, Benjamin ; OBERMAYER, Klaus: A new method for tracking modulations in tonal music in audio data format. En: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium Vol. 6 IEEE, 2000, p. 270–275

[125] RAFFEL, Colin ; ELLIS, Daniel P.: Intuitive analysis, creation and manipulation of MIDI data with pretty_midi. En: 15th international society for music information retrieval conference late breaking and demo papers, 2014, p. 84–93

[126] RAHMAN, M ; SARKAR, A ; HOSSAIN, A ; HOSSAIN, S ; ISLAM, R ; HOSSAIN, B ; QUINN, J ; MONI, M: Recognition of human emotions using EEG signals: A review. En: Computers in Biology and Medicine 136 (2021), p. 104696

[127] RAMÍREZ, A ; HORNERO, G ; ROYO, D ; AGUILAR, A ; CASAS, O: Assessment of Emotional States Through Physiological Signals and Its Application in Music Therapy for Disabled People. En: IEEE access 8 (2020), p. 127659–127671

[128] RIMBERT, S ; GAYRAUD, N ; BOUGRAIN, L ; CLERC, M ; FLECK, S: Can a Subjective Questionnaire Be Used as Brain-Computer Interface Performance Predictor? En: Frontiers in Human Neuroscience 12 (2019), p. 529. – ISSN 1662–5161

[129] RUSSELL, James A.: A circumplex model of affect. En: Journal of personality and social psychology 39 (1980), Nr. 6, p. 1161

[130] SANNELLI, C ; VIDAURRE, C ; MULLER, K ; BLANKERTZ, B: A large scale screening study with a SMR-based BCI: Categorization of BCI users and differences in their SMR activity. En: PLOS ONE 14 (2019), 01, Nr. 1, p. 1–37

[131] SANYAL, S ; NAG, S ; BANERJEE, A ; SENGUPTA, R ; GHOSH, D: Music of brain and music on brain: a novel EEG sonification approach. En: Cognitive neurodynamics 13 (2019), Nr. 1, p. 13–31

[132] SCHIRRMEISTER, Robin T. ; SPRINGENBERG, Jost T. ; FIEDERER, Lukas Dominique J. ; GLASSTETTER, Martin ; EGGENSPERGER, Katharina ; TANGERMANN, Michael ; HUTTER, Frank ; BURGARD, Wolfram ; BALL, Tonio: Deep learning with convolutional neural networks for EEG decoding and visualization. En: Human brain mapping 38 (2017), Nr. 11, p. 5391–5420

[133] SHAMSI, F ; HADDAD, A ; NAJAFIZADEH, L: Early classification of motor tasks using dynamic functional connectivity graphs from EEG. En: Journal of neural engineering 18 (2021), Nr. 1, p. 016015

[134] SIMPSON, T ; ELLISON, P ; CARNEGIE, E ; MARCHANT, D: A systematic review of motivational and attentional variables on childrenâ€™s fundamental movement skill development: The OPTIMAL theory. En: International Review of Sport and Exercise Psychology (2020), p. 1–47

[135] SINGH, A ; HUSSAIN, A ; LAL, S ; GUESGEN, H: A Comprehensive Review on Critical Issues and Possible Solutions of Motor Imagery Based Electroencephalography Brain-Computer Interface. En: Sensors 21 (2021), Nr. 6

[136] SINGH, Yeshwant ; BISWAS, Anupam: Robustness of musical features on deep learning models for music genre classification. En: Expert Systems with Applications 199 (2022), p. 116879

[137] SONG, L. ; FUKUMIZU, K. ; GRETTON, A.: Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models. En: IEEE Signal Processing Magazine 30 (2013), Nr. 4, p. 98–111

[138] SONG, XinWang ; YAN, DanDan ; ZHAO, LuLu ; YANG, LiCai: LSDD-EEGNet: An efficient end-to-end framework for EEG-based depression detection. En: Biomedical Signal Processing and Control 75 (2022), p. 103612

[139] SOROUSH, M ; MAGHOOLI, K ; SETAREHDAN, S ; NASRABADI, A: A review on EEG signals based emotion recognition. En: International Clinical Neuroscience Journal 4 (2017), Nr. 4, p. 118

[140] SOUTO, D ; CRUZ, T ; FONTES, P ; BATISTA, R ; HAASE, V: Motor Imagery Development in Children: Changes in Speed and Accuracy With Increasing Age. En: Frontiers in Pediatrics 8 (2020), p. 100

[141] SOYSA, Amani I. ; LOKUGE, Kulari: Interactive machine learning for incorporating user emotions in automatic music harmonization. En: 2010 Fifth International Conference on Information and Automation for Sustainability IEEE, 2010, p. 114–118

[142] STEEDMAN, Mark J.: A generative grammar for jazz chord sequences. En: Music Perception 2 (1984), Nr. 1, p. 52–77

[143] SUBRAMANI, K ; RAO, P: HpRNet : Incorporating Residual Noise Modeling for Violin in a Variational Parametric Synthesizer. En: ArXiv 2008.08405 (2020), p. 1–7

[144] SUGGATE, S ; MARTZOG, P: Screen-time influences children's mental imagery performance. En: Developmental Science 23 (2020), Nr. 6, p. e12978

[145] TAN, C ; SUN, F ; KONG, T ; ZHANG, W ; YANG, C ; LIU, C: A Survey on Deep Transfer Learning. En: KŮRKOVÁ, Věra (Ed.) ; MANOLOPOULOS, Yannis (Ed.) ; HAMMER, Barbara (Ed.) ; ILIADIS, Lazaros (Ed.) ; MAGLOGIANNIS, Ilias (Ed.): Artificial Neural Networks and Machine Learning – ICANN 2018. Cham : Springer International Publishing, 2018, p. 270–279

[146] TOBÓN-HENAO, Mateo ; ÁLVAREZ-MEZA, Andrés M. ; CASTELLANOS-DOMINGUEZ, Cesar G.: Kernel-Based Regularized EEGNet Using Centered Alignment and Gaussian Connectivity for Motor Imagery Discrimination. En: Computers 12 (2023), Nr. 7, p. 145

[147] TORRES-CARDONA, Hector F. ; AGUIRRE-GRISALES, Catalina ; CASTRO-LONDOÑO, Victor H. ; RODRIGUEZ-SOTELO, Jose L.: Interpolation, a model for sound representation based on BCI. En: Augmented Cognition: 13th International Conference, AC 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21 Springer, 2019, p. 471–483

[148] VÄRBU, Kaido ; MUHAMMAD, Naveed ; MUHAMMAD, Yar: Past, present, and future of EEG-based BCI applications. En: Sensors 22 (2022), Nr. 9, p. 3331

[149] VASILYEV, A ; LIBURKINA, S ; YAKOVLEV, L ; PEREPELKINA, O ; KAPLAN, A: Assessing motor imagery in brain-computer interface training: Psychological and neurophysiological correlates. En: Neuropsychologia 97 (2017), p. 56–65. – ISSN 0028–3932

[150] VELASQUEZ, L. ; CAICEDO, J. ; CASTELLANOS-DOMINGUEZ, G.: Entropy-Based Estimation of Event-Related De/Synchronization in Motor Imagery Using Vector-Quantized Patterns. En: Entropy 22 (2020), Nr. 6, p. 703

[151] VEMPATI, Raveendrababu ; SHARMA, Lakhan D.: A Systematic Review on Automated Human Emotion Recognition using Electroencephalogram Signals and Artificial Intelligence. En: Results in Engineering (2023), p. 101027

[152] VISHESH, P ; PAVAN, A ; VASIST, Samarth G. ; RAO, Sindhu ; SRINIVAS, KS: DeepTunes-Music Generation based on Facial Emotions using Deep Learning. En: 2022 IEEE 7th International conference for Convergence in Technology (I2CT) IEEE, 2022, p. 1–6

[153] VUILLEUMIER, Patrik ; TROST, Wiebke: Music and emotions: from enchantment to entrainment. En: Annals of the New York Academy of Sciences 1337 (2015), Nr. 1, p. 212–222

[154] WALLIS, Isaac ; INGALLS, Todd ; CAMPANA, Ellen ; GOODMAN, Janel: A rule-based generative music system controlled by desired valence and arousal. En: Proceedings of 8th international sound and music computing conference (SMC), 2011, p. 156–157

[155] WAN, Z. ; YANG, R. ; HUANG, M. ; ZENG, N. ; LIU, X: A review on transfer learning in EEG signal analysis. En: Neurocomputing 421 (2021), p. 1–14. – ISSN 0925–2312

[156] WANG, J ; XUE, F ; LI, H: Simultaneous channel and feature selection of fused EEG features based on sparse group lasso. En: BioMed research international 2015 (2015)

[157] WANG, N ; XU, H ; XU, F ; CHENG, L: The algorithmic composition for music copyright protection under deep learning and blockchain. En: Applied Soft Computing 112 (2021), p. 107763

[158] WANG, Yan ; SONG, Wei ; TAO, Wei ; LIOTTA, Antonio ; YANG, Dawei ; LI, Xinlei ; GAO, Shuyong ; SUN, Yixuan ; GE, Weifeng ; ZHANG, Wei [u. a.]: A systematic review on affective computing: Emotion models, databases, and recent advances. En: Information Fusion 83 (2022), p. 19–52

[159] WEI, X ; ORTEGA, P ; FAISAL, A. Inter-subject Deep Transfer Learning for Motor Imagery EEG Decoding. 2021

[160] WEINECK, K ; WEN, Olivia X. ; HENRY, Molly J.: Neural entrainment is strongest to the spectral flux of slow music and depends on familiarity and beat salience. En: bioRxiv (2021)

[161] WHISSELL, Cynthia M.: The dictionary of affect in language. En: The measurement of emotions. Elsevier, 1989, p. 113–131

[162] WILSON, J ; STERLING, A ; REWKOWSKI, N ; LIN, M: Glass half full: sound synthesis for fluid–structure coupling using added mass operator. En: The Visual Computer 33 (2017), Nr. 6, p. 1039–1048

[163] WU, D: Hearing the Sound in the Brain: Influences of Different EEG References. En: Frontiers in Neuroscience 12 (2018)

[164] XU, J ; ZHENG, H ; WANG, J ; LI, D ; FANG, X: Recognition of EEG Signal Motor Imagery Intention Based on Deep Multi-View Feature Learning. En: Sensors (Basel, Switzerland) 20 (2020)

[165] YANG, Li-Chia ; LERCH, Alexander: On the evaluation of generative models in music. En: Neural Computing and Applications 32 (2020), Nr. 9, p. 4773–4784

[166] YANG, X ; LIU, W ; LIU, W ; TAO, D: A survey on canonical correlation analysis. En: IEEE Transactions on Knowledge and Data Engineering 33 (2019), Nr. 6, p. 2349–2368

[167] YOON, J ; LEE, M: Effective Correlates of Motor Imagery Performance based on Default Mode Network in Resting-State. En: 2020 8th International Winter Conference on Brain-Computer Interface (BCI), 2020, p. 1–5

[168] YOU, Y ; CHEN, W ; ZHANG, T: Motor imagery EEG classification based on flexible analytic wavelet transform. En: Biomedical Signal Processing and Control 62 (2020), p. 102069. – ISSN 1746–8094

[169] YU, C ; QIN, Z ; MARTIN-MARTINEZ, J ; BUEHLER, M: A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using Artificial Intelligence. En: ACS Nano 13 (2019), Nr. 7, p. 7471–7482

[170] ZELLERS, Rowan ; BISK, Yonatan ; FARHADI, Ali ; CHOI, Yejin: From recognition to cognition: Visual commonsense reasoning. En: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, p. 6720–6731

[171] ZHANG, J ; YIN, Z ; CHEN, P ; NICHELE, S: Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. En: Information Fusion 59 (2020), p. 103–126

[172] ZHANG, K ; XU, G ; CHEN, L ; TIAN, P ; HAN, C ; ZHANG, S ; DUAN, N: Instance transfer subject-dependent strategy for motor imagery signal classification using deep convolutional neural networks. En: Computational and Mathematical Methods in Medicine 2020 (2020)

[173] ZHANG, R ; ZONG, Q ; DOU, L ; ZHAO, X ; TANG, Y ; LI, Z: Hybrid deep neural network using transfer learning for EEG motor imagery decoding. En: Biomedical Signal Processing and Control 63 (2021), p. 102144. – ISSN 1746–8094

[174] ZHANG, Y ; ZHOU, G ; JIN, J ; WANG, X ; CICHOCKI, A: Optimizing spatial patterns with sparse filter bands for motor-imagery based brainâ€“computer interface. En: Journal of Neuroscience Methods 255 (2015), p. 85–91. – ISSN 0165–0270

[175] ZHAO, Kun ; LI, Siqi ; CAI, Juanjuan ; WANG, Hui ; WANG, Jingling: An emotional symbolic music generation system based on lstm networks. En: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) IEEE, 2019, p. 2039–2043

[176] ZHAO, X ; ZHAO, J ; LIU, C ; CAI, W: Deep Neural Network with Joint Distribution Matching for Cross-Subject Motor Imagery Brain-Computer Interfaces. En: BioMed Research International 2020 (2020), p. 1–15

[177] ZHENG, Kaitong ; MENG, Ruijie ; ZHENG, Chengshi ; LI, Xiaodong ; SANG, Jinqiu ; CAI, Juanjuan ; WANG, Jie: EmotionBox: a music-element-driven emotional music generation system using Recurrent Neural Network. En: arXiv preprint arXiv:2112.08561 (2021)

[178] ZHENG, M ; YANG, B ; GAO, S ; MENG, X: Spatio-time-frequency joint sparse optimization with transfer learning in motor imagery-based brain-computer interface system. En: Biomedical Signal Processing and Control 68 (2021), p. 102702. – ISSN 1746–8094

[179] ZHU, J ; WEI, Y ; FENG, Y ; ZHAO, X ; GAO, Y: Physiological Signals-based Emotion Recognition via High-order Correlation Learning. En: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15 (2019), Nr. 3s, p. 1–18

[180] ZHUANG, M ; WU, Q ; WAN, F ; HU, Y: State-of-the-art non-invasive brainâ€“computer interface for neural rehabilitation: A review. En: Journal of Neurorestoratology 8 (2020), Nr. 1, p. 12

[181] ZHUANG, Y ; LIN, L ; TONG, R ; LIU, J ; IWAMOT, Y ; CHEN, Y: G-gcsn: Global graph convolution shrinkage network for emotion perception from gait. En: Proceedings of the Asian Conference on Computer Vision, 2020