NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Emotion Recognition with EEG Data: Leveraging Pretrained Convolutional Neural Networks

David Psiuk

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Emotion Recognition with EEG Data:

Leveraging Pretrained Convolutional Neural Networks

by

David Psiuk

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Professor Fernando José Ferreira Lucas Bação

NOVA Information Management School, Universidade Nova de Lisboa

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

David Psiuk

Lisbon, Portugal

1 July 2025

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my Supervisor, Professor Fernando Bação, for providing guidance and support throughout the ambitious research challenge I proposed. Working on this thesis has been a deeply enriching experience, and I've learned a lot from the entire process.

Special thanks to my colleagues at NOVA, especially Luis, Peter, Noah and Pedro, whose encouragement and intellectual contributions were essential during the challenging phases of this Master's program.

Finally, I am deeply thankful to my family, especially my parents and my sister, whose emotional support was invaluable to the success of this endeavor.

ABSTRACT

Emotion recognition using electroencephalogram (EEG) signals is a key area in brain–computer interfaces with valuable applications in domains such as mental health and clinical diagnostics. To address the challenge of limited data availability in EEG-based emotion recognition, convolutional neural networks (CNNs) pretrained on image data can be leveraged by transforming EEG signals into two-dimensional spectrograms using the Short-Time Fourier Transform (STFT). However, existing literature lacks a systematic comparison of preprocessing methods, feature representations, and CNN adaptation strategies, leaving open questions about best practices in this domain. This study addresses these gaps by systematically evaluating multiple EEG preprocessing configurations and comparing two types of 2D representations: spectrograms and stacked spectrograms, the latter of which has not yet been explored in the context of emotion recognition. The evaluation framework includes four pretrained CNN architectures (MobileNetV2, ResNet50, InceptionV3, and DenseNet121) and compares two approaches to leveraging these models: transfer learning through fine-tuning and feature extraction with an external SVM classifier. All experiments were conducted on the SEED dataset, which contains EEG recordings labeled with positive, neutral, and negative emotional states. The results show that the identified preprocessing configurations substantially improved validation accuracy. While stacked spectrograms demonstrated potential for capturing multichannel EEG dynamics, they were ultimately outperformed by individual spectrograms due to reduced training data availability. Feature extraction surpassed transfer learning in both validation accuracy and computational efficiency. The best overall performance was achieved using InceptionV3 and DenseNet121 in combination with an SVM classifier, reaching a competitive average validation accuracy of 86.67%, and significantly outperforming a baseline SVM model trained on traditional PSD features. The findings offer valuable guidance for future work on EEG-based emotion recognition, highlighting the effectiveness of feature extraction with pretrained CNNs and demonstrating the value of adapting visual deep learning models to neurophysiological data.

KEYWORDS

EEG; Emotion Recognition; Pretrained CNNs; Transfer Learning; Spectrograms

Sustainable Development Goals (SDG):

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BCI	Brain–Computer Interface
CNN	Convolutional Neural Network
EEG	Electroencephalography
SEED	SJTU Emotion EEG Dataset
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
ML	Machine Learning
2D	Two-Dimensional
PSD	Power Spectral Density
ATAR	Automatic Tunable Artifact Removal
Hz	Hertz
LR	Learning Rate

1. INTRODUCTION

Brain–computer interfaces (BCIs) have gained significant attention as a promising field at the intersection of neuroscience and artificial intelligence. A brain–computer interface is a system that enables communication between humans and external devices by interpreting the neural activity generated by the brain (Houssein et al., 2022).

Among the available neural recording techniques, electroencephalography (EEG) is one of the most widely used in BCI research due to its high temporal resolution, portability, non-invasiveness, and relatively low cost (Khosla et al., 2020). EEG captures electrical activity through electrodes placed on the scalp, making it particularly suitable for real-time applications such as assistive technologies and cognitive state monitoring (Martišius & Damaševičius, 2016).

Emotion recognition is a particularly active area within BCI research, bridging fields such as computer science, neuroscience, psychology, biomedical engineering, and medical science (X. Li et al., 2022). While identifying emotional states from brain activity is challenging due to the complexity and variability of EEG signals, it also offers a valuable opportunity to deepen our understanding of how emotions are represented in the brain (X. Wang et al., 2023).

A central resource in this field is the SEED dataset, which was developed specifically for emotion recognition using EEG signals and will also be used in this study. As of September 2024, it has been used in over 6,340 studies across more than 2,320 research institutions worldwide, underscoring the growing interest and potential in this domain (SJTU BCI Lab, 2024).

Beyond its theoretical significance, EEG-based emotion recognition holds practical value across fields such as mental health diagnostics and affective computing (X. Wang et al., 2023). For instance, it can support computer-aided diagnosis of conditions like post-traumatic stress disorder (PTSD) by analyzing neural responses to trauma-related stimuli (Rozgic et al., 2014). Studies have demonstrated that EEG-based emotion recognition can enable real-time detection of emotional states in patients with consciousness disorders, offering promising diagnostic support in clinical settings (Huang et al., 2021). It has shown strong potential to support the treatment of depression, with findings indicating significant cognitive, clinical, and neural improvements in patients (Patil et al., 2023), as well as enhanced emotional awareness and social integration in individuals with autism and other neurodevelopmental conditions (Samal et al., 2024). These examples highlight the potential of EEG-based emotion recognition in both research and applied contexts.

To process EEG data for emotion classification, traditional Machine Learning (ML) models such as Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), and Random Forests have been extensively studied (Ackermann et al., 2016; Mohammadi et al., 2017; Katsigiannis & Ramzan, 2018). While these approaches can yield competitive results, they often require

carefully engineered features and struggle to capture the complex spatial–temporal patterns inherent in EEG signals.

In contrast, convolutional neural networks (CNNs) have demonstrated impressive results in learning rich feature representations from input data, especially images (Szegedy et al., 2016). However, training deep CNNs effectively requires large, labeled datasets and high computational resources, which has led to growing interest in pretrained networks, where models are initially trained on large-scale datasets like ImageNet and can further be reused and adapted for new tasks (Hussain et al., 2019). By transforming EEG signals into two-dimensional (2D) image representations, researchers can leverage pretrained models that have already learned to recognize fundamental patterns and apply this knowledge to tasks like EEG-based emotion recognition, even when labeled training data is limited (Bagherzadeh et al., 2023).

Although substantial progress has been made in EEG-based emotion recognition using pretrained CNNs, key methodological challenges persist. Most notably, prior research has lacked systematic comparisons of preprocessing pipelines, 2D EEG feature representations, and CNN architectures. Moreover, inconsistent modeling strategies, particularly in how pretrained networks are leveraged have led to conflicting results. These limitations complicate efforts to identify robust, generalizable approaches for emotion recognition from EEG data.

To address this gap, the present study systematically evaluates the influence of different preprocessing techniques and 2D EEG representations, including individual spectrograms and stacked spectrograms. Additionally two widely used strategies for leveraging pretrained CNNs are compared: transfer learning and feature extraction. These approaches are evaluated in terms of their classification performance across four pretrained CNN architectures.

By exploring these factors in a unified framework, this research aims to identify optimal practices for EEG-based emotion recognition and offer comparative insights that contribute to the broader understanding of Deep Learning applications in neurophysiological data analysis.

Based on this objective, the study is guided by the following research questions:

RQ1: Which preprocessing and image generation techniques yield the most informative representations for CNN-based EEG emotion classification?

RQ2: How do different 2D feature representations (spectrogram and stacked spectrogram) affect model performance?

RQ3: Which strategy for leveraging pretrained CNNs (transfer learning or feature extraction) yields better performance in EEG-based emotion recognition?

In addition to evaluating classification performance, this study also examines spatial and temporal patterns in the EEG data to gain further insight into model behavior and how these patterns relate to emotional processing.

The remainder of this thesis is structured as follows: Chapter 2 provides a comprehensive review of the existing literature on EEG-based emotion recognition; Chapter 3 outlines the methodology used in this study; Chapter 4 presents and discusses the experimental results of different configurations; Chapter 5 summarizes the key findings and conclusions and Chapter 6 discusses limitations and proposes directions for future research.

2. LITERATURE REVIEW

This chapter provides a structured review of the existing literature on emotion recognition using EEG and the role of pretrained CNNs. It begins with an overview of EEG-based emotion recognition, including methods of data acquisition and the main challenges in the field. Next, it discusses common EEG signal processing techniques, focusing on preprocessing and feature extraction methods. Subsequently, CNNs and strategies for using pretrained models will be presented, followed by an exploration of how EEG features can be represented in a 2D format for CNNs and how these methods are applied to emotion recognition tasks. It concludes by summarizing key findings and highlighting current research gaps.

2.1. EEG FOR EMOTION RECOGNITION

This section overviews EEG data acquisition and its role in emotional computing, highlighting key challenges such as signal complexity, noise, and subject-specific variability.

2.1.1. EEG Data

EEG is a non-invasive technique used to measure the electrical activity of the brain through electrodes placed on the scalp (W. L. Zheng & Lu, 2015). The signals are commonly recorded through an EEG device using the 10-20 international system, as shown in Figure 2.1 (a). In this system, numbers indicate electrode position, while the letters T, F, C, O, and P correspond to different brain regions: Temporal (side of the cortex), Frontal (front of the cortex), Central (midline area), Occipital (back of the cortex), and Parietal (upper rear region), respectively (Ein Shoka et al., 2023).

Since each electrode records the activity of a different brain region, a single recording results in multi-channel EEG data, as shown in Figure 2.1 (b). The data is recorded in Hertz (Hz), with typical sampling rates such as 512 Hz, meaning 512 data points are captured per second (Koelstra et al., 2012).

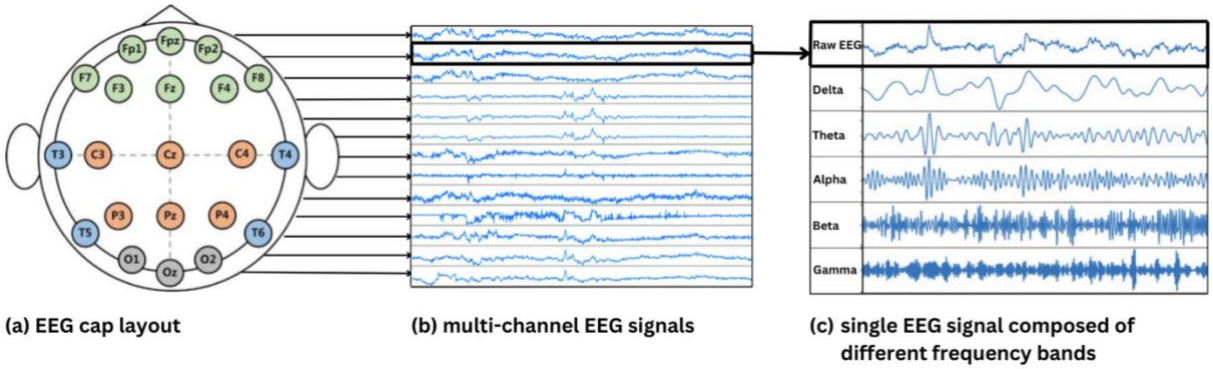

Figure 2.1: Visual Overview of EEG Recording and Signal Properties

When neurons become active, they generate electrical fields around the scalp and since there are billions of neurons in the brain, each electrode captures the cumulative activity of many

neurons at once (Bajaj, 2021). Therefore, the EEG signal obtained from a single electrode results in a complex waveform, which can be viewed as a combination of slow and fast oscillations happening at the same time (Hazarika et al., 1997). Consequently, researchers often segment EEG signals into specific frequency bands, enabling them to isolate and analyze these oscillations more effectively (W. L. Zheng & Lu, 2015). The characteristics of a raw EEG signal and its composition of distinct frequency bands can be observed in Figure 2.1 (c).

These frequency bands are typically categorized as Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (12-30 Hz), and Gamma (>30 Hz) (Zhong et al., 2022). Each of these bands is associated with distinct cognitive states and functions, making them valuable for various applications, including cognitive workload assessment, clinical diagnostics and emotion recognition (Zhong et al., 2022; X. Li et al., 2022). A summary of the frequency bands and their associated cognitive functions is presented in Table 2.1.

Table 2.1: EEG frequency bands and their associated cognitive functions

Frequency Band	**Range**	**Associated Cognitive Functiois**
Delta	1-4 Hz	Deep Sleep
Theta	4-8 Hz	Drowsiness, Creativity, Deep meditation
Alpha	8-12 Hz	Relaxed awareness, attentional processing
Beta	12-30 Hz	Active thinking, problem-solving
Gamma	>30 Hz	High-level Cognition and Information Processing

2.1.2. EEG Emotion Representation and Challenges

Over the past decade, significant efforts have been directed towards emotion recognition using data from diverse sources such as physiological signals, audio and facial expressions (Koelstra et al., 2012). Given that the brain, as part of the central nervous system, regulates the autonomic nervous system involved in emotional processes, directly studying brain activities, particularly through EEG, offers a direct and meaningful approach to understanding emotional cognition and recognizing emotional states (X. Li et al., 2022).

This study utilizes the SEED (SJTU Emotion EEG Dataset), a widely used dataset for EEG-based emotion recognition. Data was collected from participants who watched emotionally evocative film clips, designed to induce positive, neutral, and negative emotions, while their brain activity was recorded using a 62-channel EEG system (W. L. Zheng & Lu, 2015).

When studying emotion recognition with EEG, a critical question is how emotions are represented in the brain. The primary approaches to addressing this involves examining the distribution of different frequency bands and analyzing specific regions of the cortex.

When looking deeper into emotional processing in relation to frequency bands, research consistently indicates that higher frequency EEG bands, particularly Beta and Gamma, are most effective for emotion classification and outperform lower frequency bands (F. Wang et al., 2020; J. Li et al., 2018). Particularly, Gamma band activity has been identified as optimal for distinguishing emotions like happiness and sadness (Yang et al., 2020; M. Li & Lu, 2009). However, W. L. Zheng & Lu (2015) noted that neutral emotions are associated with increased Alpha power, underscoring the relevance of lower frequency bands in emotion recognition. X. Li et al. (2022) support these findings, confirming that Beta and Gamma bands are the most effective while emphasizing that combining all frequency bands achieves superior performance.

Findings also suggest that positive and negative emotions are not confined to a single brain region but are distributed across multiple areas, particularly within the frontal, temporal, and parietal regions (X. Li et al., 2022). Furthermore, hemispheric asymmetries have shown to play a key role in emotion recognition, as electrical relations between asymmetrical EEG channels are linked to emotional intensity and whether the emotion is pleasant or unpleasant (Cimtay & Ekmekcioglu, 2020).

For example, Studies indicate that joyful and happy music increases left frontal activity, while fear and sadness heighten right frontal activity (Schmidt & Trainor, 2001). Further research confirms these asymmetries but also reveals that the left anterior cortex is involved in both positive emotions such as joy and certain negative emotions such as anger and disgust (Aftanas et al., 2006). These findings underscore the complexity of emotional processing and its dependence on distributed neural interactions rather than isolated brain structures.

Several challenges exist in EEG-based emotion recognition, particularly regarding the reliability of emotion labels and inter-subject variability. Y. Wang et al. (2018) noted that while induced emotions should ideally remain stable in EEG emotion datasets, in practice they often fluctuate due to individual sensitivity to stimuli. They observed that the self-reported emotional state was conventionally recorded as an average over the entire trial, creating a mismatch with the true fluctuating intensity (see Figure A.1 in Appendix A). This inconsistency undermines label reliability and negatively impacts model training.

Furthermore, Hamann & Canli (2004) found that differences in personality, biological sex, and genotype can significantly influence the neural mechanisms of emotion processing in different brain regions in emotion processing. Goshvarpour & Goshvarpour (2019) found that women showed higher EEG power across all frequency bands than men, with distinct spatial distribution differences in response to emotional music videos. EEG signals exhibit significant variability not only between genders but, among individuals in general. Numerous studies have demonstrated significant inter-subject differences in EEG patterns, posing a major challenge in the development of a generalizable model for emotion classification (Samek et al., 2013; Sidharth et al., 2023).

2.2. EEG Preprocessing and Feature Extraction

Preprocessing EEG signals and extracting meaningful features are essential steps for effective emotion recognition. Figure 2.2 illustrates the overall workflow, including preprocessing and feature extraction procedures, used to transform raw EEG data into a structured representation suitable for input into a CNN.

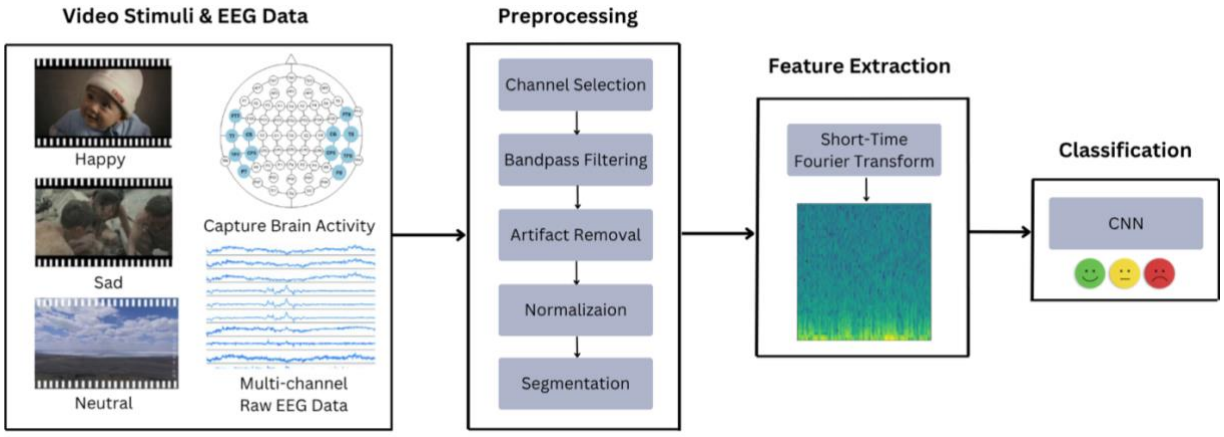

Figure 2.2: General outline of EEG data processing workflow

The following sections describe these key preprocessing and feature extraction methods.

2.2.1. EEG Preprocessing

EEG signal processing is a crucial step in emotion recognition, as raw EEG data is often noisy and prone to artifacts from muscle movements, eye blinks, and external interference (X. Li et al., 2022). To ensure the reliability and effectiveness of EEG-based models, researchers apply various preprocessing techniques, including channel selection, bandpass filtering, artifact removal, normalization and segmentation**.** The following sections will explain these techniques, highlighting their role in improving signal quality and enhancing model performance.

Research has shown that certain brain regions contribute more significantly to emotion classification than others, making channel selection a valuable strategy for improving efficiency without sacrificing accuracy. Li et al. (2018) found that channels in the temporal lobe (sides of the cortex) were the most effective for emotion recognition, while Zhuang et al. (2018) additionally identified significant channels in the prefrontal regions.

However, no single optimal channel selection has been clearly established, as different studies have used varying combinations of electrodes, each achieving competitive results (X. Li et al., 2018; Zhuang et al., 2018, Cimtay & Ekmekcioglu, 2020; W. L. Zheng & Lu, 2015). Notably, W. L. Zheng and Lu (2015) systematically compared different channel configurations and found that a 12-channel profile with electrodes placed mostly in the temporal and parietal regions

(side and upper-back areas of the cortex) achieved the highest performance in their study, achieving even better accuracy than using all available 62 electrodes.

Next, bandpass filtering is a preprocessing technique used to remove noise and restrict EEG signals to relevant frequency ranges. Researchers typically apply high-pass and low-pass filters to eliminate unwanted frequency components and focus on meaningful neural activity (Raghu et al., 2020). However, the specific filtering settings vary across literature, with researchers using different frequency ranges and configurations. For example, Sadiq et al. (2022) applied a bandpass filter ranging from 0.5 Hz to 200 Hz, while J. Li et al. (2018) passed the EEG signals through a filter of 0.3–50 Hz for emotion recognition on the SEED dataset. This variation highlights the lack of a consistent standard for filtering in EEG-based emotion recognition studies.

Additional noises in EEG signals can be depicted by ocular artifacts like eye blinking and cardiac interferences which are dominant below 4 Hz, muscle movements that produce artifacts above 30 Hz, and powerline noise which lie between 50 and 60 Hz (Katsigiannis & Ramzan, 2018; X. Li et al., 2022). Filtering alone cannot remove all these artifacts from the EEG signals so additional artifact removal can be applied to remove unwanted noise in the EEG signals (Katsigiannis & Ramzan, 2018). J. Li et al. (2018) located and removed the artifacts manually. However, this approach is highly time-consuming and impractical for large-scale studies or real-time applications.

Another commonly used method for reducing artifacts such as eye blinks and muscle noise is wavelet-based denoising (X. Li et al., 2022). It operates by decomposing the EEG signal into wavelet coefficients, which represent the signal's frequency components at various time scales. Subsequently, coefficients below a predefined threshold are suppressed before reconstructing the cleaned signal (Donoho & Johnstone, 1994). For instance, Abbas et al. (2023) aimed to preserve relevant signal features while reducing noise in the EEG signal by applying the Discrete Wavelet Transform (DWT).

Additionally, Bajaj et al. (2020) proposed an automatic denoising approach known as Automatic Tunable Artifact Removal (ATAR). This method applies wavelet packet decomposition to the EEG signal and uses an adaptive thresholding mechanism governed by a tunable parameter (β), which controls the aggressiveness of artifact suppression. In contrast to manual or fixed-threshold techniques, ATAR offers a more flexible framework for balancing signal preservation and artifact reduction. It has been adopted in subsequent studies such as (Aslan et al., 2024) who employed ATAR for artifact correction in EEG-based lie detection.

However, some researchers solely apply bandpass filtering for artifact removal (F. Wang et al., 2020; Raghu et al., 2020). As a simpler and more computationally efficient approach, it may be more feasible for real-time applications where minimal processing is required.

Feature normalization is commonly used in EEG-based machine learning tasks to ensure comparability between feature vectors, as their magnitude and range vary depending on

signal characteristics, recording conditions, and individual differences (Katsigiannis & Ramzan, 2018). For example, EEG signals at higher frequencies tend to have smaller magnitudes than those of lower frequencies in EEG data, making normalization essential for balanced feature representation (Katsigiannis & Ramzan, 2018). To address these challenges, researchers commonly center the features by subtracting the mean to reduce bias and stabilize training (Cimtay & Ekmekcioglu, 2020). Normalization helps align the distribution of EEG features, reducing inconsistencies and ultimately improving model stability and generalization across individuals (Kim et al., 2022).

Beyond normalization, some studies further apply baseline correction to ensure consistency in EEG data and prevent extreme values from influencing model training. The baseline correction method utilizes a neutral section of the EEG data, which is often an artifact-free time window before the stimulus, to calculate baseline features (Kim et al., 2022). The extracted EEG features are then divided by the corresponding baseline features to obtain a normalized feature which removes background activity to emphasize changes in the EEG recordings associated with stimuli (Katsigiannis & Ramzan, 2018).

Another important step in EEG classification is segmentaion. Several studies use fixed-length segments for feature extraction rather than entire trials as this approach increases the amount of training data (Yap et al., 2023). Since the length of the EEG signals vary significantly across different stimuli in the SEED dataset, trial segmentation has the additional effect of standardizing the data by ensuring that each emotion has an equal number of samples (F. Wang et al., 2020).

While segment length can significantly influence the resulting feature representation and model performance, the optimal segment duration for capturing emotion-relevant EEG features remains uncertain. For example, F. Wang et al. (2020) segmented each trial into multiple 1-second windows, using these shorter segments as individual feature representations for model training. Similarly, Y. Zhang et al. (2018) retained a 9-second segment from the trials for feature extraction. Pusarla et al. (2022) on the other hand, used the second half of each trial in the SEED dataset, considering approximately 2 minute segments for model training.

2.2.2. EEG Feature Extraction

There are numerous different methods for extracting features from EEG signals. X. Li et al. (2022) differentiated them into Time Domain Features, Frequency Domain Features and Time-Frequency Domain Features, among others. The following section introduces key EEG feature extraction techniques commonly used in research, which are essential for constructing the final feature representation for CNNs.

Time-domain features provide a straightforward approach to extract EEG features by calculating statistical measures such as mean and variance of the EEG signal amplitude. These

metrics capture the temporal characteristics of the EEG signal and provide insights into its time-domain properties (X. Li et al., 2022).

Frequency-domain features offer a complementary view on the EEG signal by analyzing its distribution across various frequency components. This transformation is typically performed using the Fourier Transform, which decomposes a time-domain EEG signal into its constituent frequency components (Cecotti & Graeser, 2008). As described by F. Wang et al. (2020), this method averages frequency content across the entire duration of the signal, enabling the analysis of power within the distinct EEG frequency bands previously discussed in Section 2.1.1.

A commonly used frequency-domain feature in EEG emotion recognition is power spectral density (PSD) (W. L. Zheng & Lu, 2015; Z. Wang et al., 2022). PSD describes the power distribution of the signal across different frequency bands and is derived from the Fourier Transform (Bisina & Azeez, 2017).

Figure 2.3 shows an example of this transformation, where raw EEG data from a 60-second segment and 12 channels was converted from the time domain to the frequency domain. In the resulting plot, the x-axis indicates frequency (Hz), while the y-axis shows the corresponding power associated with each frequency component.

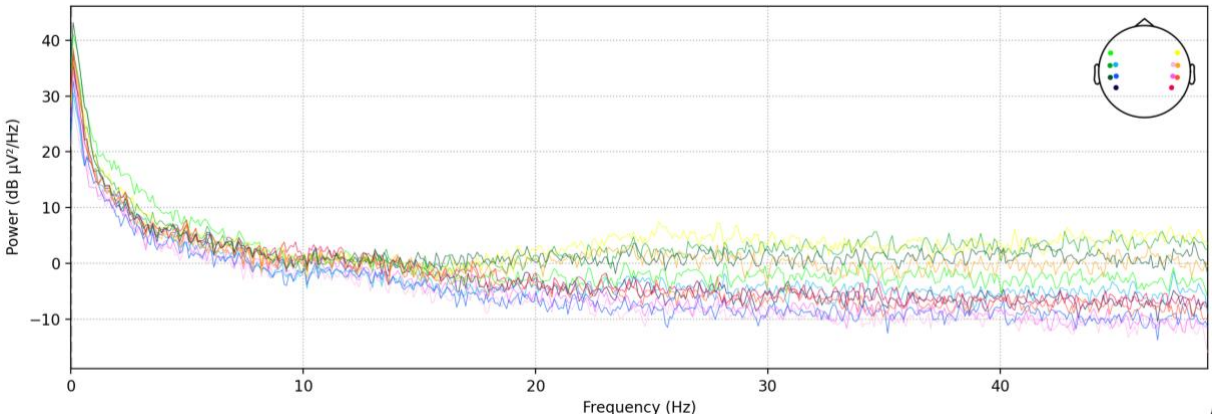

Figure 2.3: Power spectral density of EEG signals

The transition to dynamic time-frequency domain composition features is substantial, since the individual features of the time and frequency domain cannot represent the properties of the EEG signal completely (X. Li et al., 2022). Time-frequency analysis enables to investigate variations in the frequency content of EEG signals over time. Different methods such as the short-time Fourier transform (STFT) are commonly used to instantiate this transition (Mandhouj et al., 2021; Abdulwahhab et al., 2024).

Unlike the standard Fourier Transform, which falsely assumes signal stationarity (F. Wang et al., 2020), the STFT addresses this limitation by dividing the signal into shorter overlapping windows, within which the signal can be considered approximately stationary (J. Li et al.,

2018). Each window is then individually transformed using the Fourier Transform, enabling the analysis of how the signal's frequency components evolve over time.

The STFT is commonly used to produce a spectrogram, which visually represents the signal's strength in the time-frequency domain (Mandhouj et al., 2021). As illustrated in Figure 2.4, a spectrogram is essentially a heatmap where the x-axis represents the time, the y-axis represents the frequency bins, and the color indicates the intensity or power of the signal at each frequency across time points (Pusarla et al., 2022).

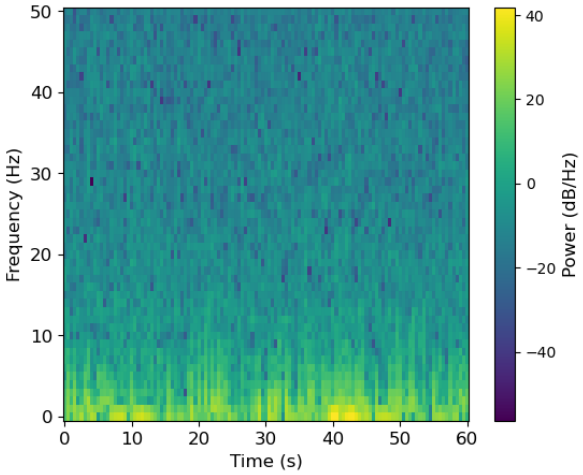

Figure 2.4: Time-frequency representation of EEG data: spectrogram

In addition to time-domain, frequency-domain, and time-frequency features, entropy-based methods such as Differential Entropy (DE) have also been explored in the literature to quantify the complexity of EEG signals, particularly in emotion recognition tasks (Duan et al., 2013; W. L. Zheng & Lu, 2015). This reflects the diversity of feature representations developed for EEG-based analysis, although they are primarily designed for traditional ML models. Consequently, such features were not included in this study due to the added complexity of adapting them to a format compatible with convolutional neural networks.

2.3. Convolutional Neural Networks for EEG Data

Convolutional Neural Networks (CNNs) represent a significant advancement in the field of Deep Learning, particularly in the domain of computer vision (Krizhevsky et al., 2017). They are designed to automatically and adaptively learn spatial hierarchies of features from input images, making them exceptionally effective for tasks such as image classification, object detection, and segmentation (Z. Li et al., 2022). Additionally, CNNs have been adapted and extensively utilized in the field of EEG classification, where they demonstrate strong capabilities in capturing spatial and temporal patterns within EEG signals (Cimtay & Ekmekcioglu, 2020; Abdulwahhab et al., 2024).

2.3.1. Introduction to CNNs

A typical CNN consists of several layers, each serving a distinct purpose in the feature extraction process. The fundamental layers include convolutional layers, pooling layers, and fully connected layers (Raghu et al., 2020). Convolutional layers apply a set of learnable filters (kernels) to the input image, producing feature maps that capture specific patterns or features across all locations of the input data (LeCun et al., 2010). To control how the convolution kernels interact with the input, padding can be applied to preserve border information by adding zero values, while stride determines the step size of the filter movement, affecting the output resolution (Z. Li et al., 2022).

Following the convolutional layers, pooling layers are applied to reduce the spatial dimensions of the feature maps, thereby decreasing the computational load while retaining useful information (Z. Li et al., 2022). Finally, fully connected layers are commonly used in conjunction with softmax for multi-class classification, mapping the extracted features to class probabilities (Krizhevsky et al., 2017; Z. Li et al., 2022). A general architecture of a CNN can be seen in Figure 2.5.

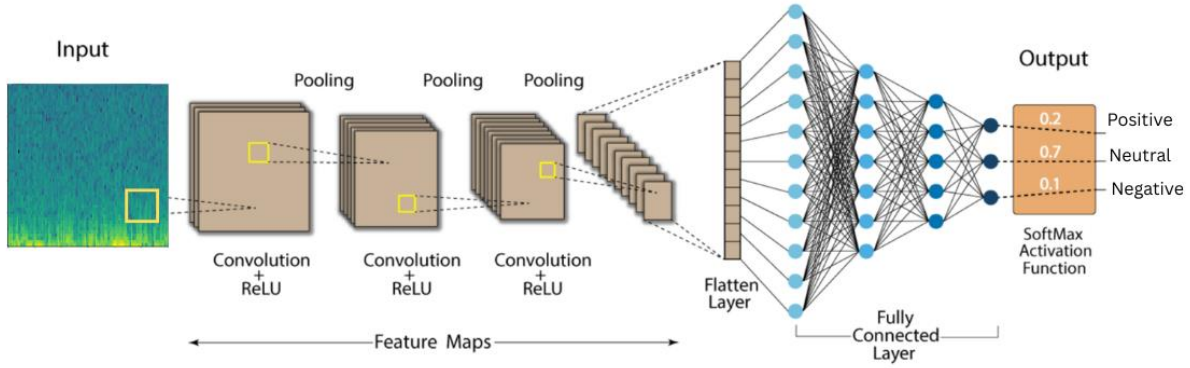

Figure 2.5: Architecture of a Convolutional Neural Network (adapted from Rguibi et al., 2022)

One of the most remarkable aspects of CNNs is their ability to learn features hierarchically. Early layers typically learn low-level features, such as edges and textures, while deeper layers capture more abstract representations, such as shapes and objects (Zeiler & Fergus, 2013). This hierarchical feature learning enables CNNs to generalize well across various tasks and datasets, significantly reducing the need for manual feature engineering (Razavian et al., 2014).

2.3.2. Leveraging Pretrained CNNs

While the capabilities of CNNs offer substantial advantages, the successful training of such models often demands extensive labeled datasets and considerable computational resources. To overcome these challenges, researchers have increasingly turned to pretrained CNNs as a practical alternative (Gao et al., 2025). These models are initially trained on large-scale image

datasets such as ImageNet, which contains millions of labeled images across a wide range of categories (Arshed et al., 2023). These models retain rich, hierarchical feature representations that capture low and high-level visual patterns (Alam et al., 2024). Such representations are highly generalizable and can be repurposed for tasks in different domains (Krishnapriya & Karuna, 2023).

In the context of EEG-based tasks, transforming signals into 2D formats like spectrograms enables the use of pretrained CNNs originally developed for natural image classification (Sadiq et al., 2022). This approach has the advantage of reusing feature detectors trained on diverse and large-scale datasets, making model training particularly useful in scenarios with limited labaled data such as EEG emoiton recognition (W. Li et al., 2022).

Pretrained CNNs can be adapted to new tasks through different strategies. In the literature, two commonly employed methods are transfer learning and feature extraction.

Transfer learning (also referred to as fine-tuning) typically involves initializing a CNN with weights from a pretrained model and adapting it to a new task by replacing the final classification layer with a new one tailored to the target output space (Tajbakhsh et al., 2016). One or more layers of the network are then retrained to adjust the learned features to the target data distribution (X. Li et al., 2022). Empirical results have demonstrated that fine-tuned models can outperform or match CNNs trained from scratch, even in domains with markedly different characteristics, such as medical imaging (Tajbakhsh et al., 2016).

Feature extraction refers to leveraging the pretrained CNN as a fixed feature encoder by transforming input images into numerical feature vectors that capture spatial and structural information (Razavian et al., 2014). Subsequently, these feature vectors can serve as input to external ML classifiers for downstream tasks such as classification (Raghu et al., 2020). This approach is computationally efficient, as it requires only a single forward pass through the CNN without retraining (Sadiq et al., 2022).

While both strategies are technically forms of transfer learning since they involve transferring knowledge from a source domain to improve performance in a target domain (Pan & Yang, 2010). In empirical research, they are referred to separately (Sadiq et al., 2022; Raghu et al., 2020). In this study, the convention will be followed: the term transfer learning refers to the adaptation of the CNN's final classification layer followed by fine-tuning, whereas feature extraction denotes using a pretrained network to generate fixed feature representations for an external machine learning classifier. The implementation details of both strategies are described in 3.3.3.

2.3.3. EEG Feature Representations for CNNs

Applying CNNs on EEG-based classification tasks requires transforming the data into a representation that matches the model's input format (X. Li et al., 2022). CNNs were originally designed for image processing, and pretrained models typically expect 2D image inputs (Raghu et al., 2020). Consequently, a crucial step in this study is to identify a feature representation of the EEG signal that aligns with these requirements while preserving the integrity of the underlying data. Therefore, the emotion recognition task with EEG data will be transformed to a computer vision task (J. Li et al., 2018). In the following, the most effective 2D EEG feature representations identified in the literature will be explored, along with their limitations and practical implications.

As priorly mentioned in section EEG Feature Extraction, spectrograms are time-frequency domain features derived using the STFT. These features are conveniently represented in a 2D format and have been used for EEG classification tasks in research.

For example, Shen et al. (2024) employed spectrograms as a representation of EEG signals to enable real-time detection of epileptic seizures by directly feeding the spectrograms into a CNN. Similarly, Sadiq et al. (2022) made use of scalograms as a time-frequency feature representation of motor and mental imagery EEG data. Based on 18 selected channels from the motor cortex region, they generated a total of 18 images corresponding to a single trial of a subject. Pusarla et al. (2022) generated a total of 15x15x62 = 13,950 spectrograms across the entire SEED dataset, derived from 15 subjects, each completing 15 trials, with data recorded from 62 distinct EEG channels for emotion recognition. Consequently, each spectrogram represents the data from a single channel within a single trial of an individual subject.

While spectrograms yield promising results in different areas, a key limitation in these studies is the independent processing of each EEG channel. The papers suggest that representations are generated separately per channel, disregarding the multichannel dynamics inherent in EEG data. This approach risks losing critical inter-regional brain activity and channel dependencies, which are valuable information for emotion recognition, potentially limiting the model’s ability to capture the full complexity of EEG patterns.

To incorporate multi-channel information, researchers have modified spectrograms to capture inter-channel relationships in a 2D format, potentially enabling models to learn more complex EEG dynamics.

X. Li et al. (2017) constructed a channel-by-scale grid by segmenting EEG signals into fixed-length windows, summing scalogram values along the time axis to capture energy distributions across scales, and stacking these vectors from multiple channels. Similarly, F. Wang et al. (2020) proposed electrode-frequency distribution maps (EFDMs) for emotion recognition using the SEED dataset, as illustrated in Figure 2.6 (a). They applied the STFT and stacked the values from multiple channels to form a compact 32×64 image, where 32

represents channels and 64 frequency bins. The proposed 2D representations were then used as input to a custom-designed CNN.

While these approaches achieved promising results, the small image size necessitates a customized CNN, making them impractical for use with pretrained models, which typically require larger input dimensions, such as 224 × 224 (Sadiq et al., 2022).

To address this limitation, Raghu et al. (2020) transformed EEG signals into 2D spectrogram stacks that incorporate multi-channel data while enabling the use of pretrained CNNs for classifying seven variants of epileptic seizures. They generated spectrograms of the whole recording for 19 selected EEG channels and vertically concatenated them into a single tall image, integrating multi-channel information.

As seen in Figure 2.6 (b), the x-axis corresponds to time, while the y-axis combines frequency and channel-specific data into a unified vertical dimension. The increased image size allows to effectively resize the inputs to match the expected dimensions of pretrained networks. However, no prior research has explored this specific representation for EEG-based emotion recognition.

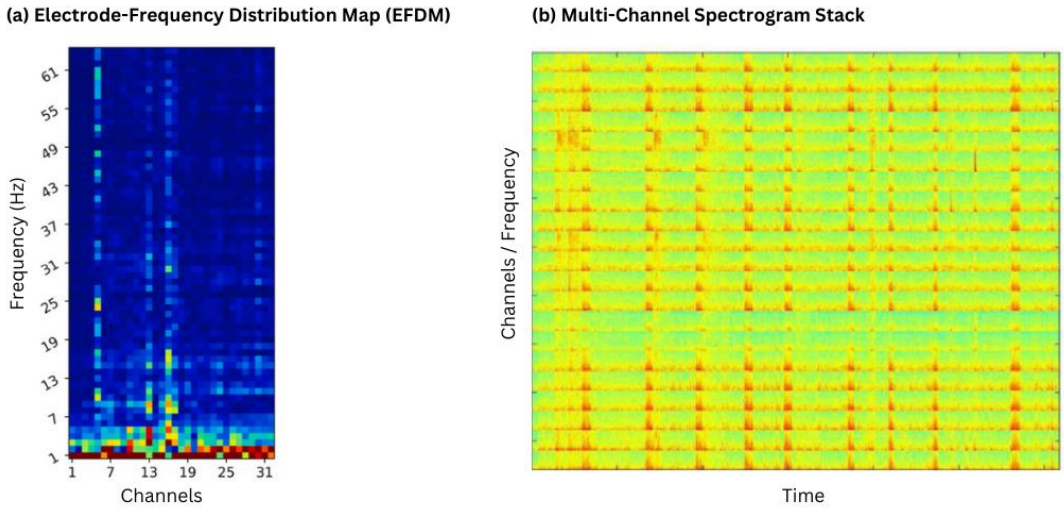

Figure 2.6: 2D Multi-Channel Time-Frequency Representations. Adapted from (F. Wang et al., 2020) and (Raghu et al., 2020).

Additionally, topographical mapping approaches have been explored to incorporate the spatial arrangement of EEG electrodes into EEG feature representations. For instance, J. Li et al. (2018) generated electrode-based maps for different frequency bands where each pixel corresponds to a certain electrode and Differential Entropy values were inserted. Similarly, Thodoroff et al. (2016) calculated the spectral power values across channels for three frequency bands and mapped them topographically into a single 32×32×3 image for seizure detection. While these approaches effectively capture spatial information, they produce small feature maps that require custom models and are impractical for use with pretrained networks designed for larger inputs.

2.3.4. Pretrained CNNs in EEG Emotion Recognition

Numerous studies have explored the integration of 2D EEG representations with pretrained CNNs to improve emotion recognition performance. By transforming EEG signals into image-like formats such as spectrograms, researchers have been able to apply well-established CNN architectures originally developed for natural image processing.

For example, Asghar et al. (2019) leveraged the pretrained AlexNet model by extracting features from spectrograms, which were then refined using a Bag of Deep Features (BoDF) approach and classified using an external SVM, achieving an accuracy of 93.80% on the SEED dataset. Similarly, Pusarla et al. (2022) developed a customized architecture based on the pretrained DenseNet121, feeding spectrograms into the pretrained network and classifying emotions with an additional SVM.

In contrast, Cimtay & Ekmekcioglu (2020) applied windowing and reshaping of raw EEG signals before feeding it into a pretrained InceptionResNetV2 model. Their method did not use a 2D image representation and achieved a lower accuracy of 78.34% on the SEED dataset, suggesting that representations such as spectrograms may provide more informative inputs for CNN-based emotion classification compared to raw signals. Furthermore, F. Wang et al. (2020) pretrained their proposed CNN on the SEED dataset before applying it to an additional emotion recognition dataset. After fine-tuning the network on the new data, they achieved an accuracy of 82.84%.

Collectively, these studies demonstrate the diverse strategies employed to leverage pretrained CNN architectures for EEG-based emotion recognition. However, most research customizes a single pretrained model with different configurations, making it difficult to draw direct comparisons or identify the most effective adaptation strategies. This highlights the absence of a standardized evaluation framework in the field, a gap that has been addressed more systematically in other EEG-related domains.

For example, Raghu et al. (2020) conducted a systematic evaluation of ten pretrained CNN architectures, including AlexNet, ResNet50, DenseNet201, InceptionV3, and others, using stacked spectrograms as input for seizure classification. They directly compared the two previously discussed approaches: feature extraction, where pretrained features were classified using an SVM, and transfer learning, where the final layers of the CNNs were replaced and retrained using a softmax classifier. Their results showed that feature extraction generally outperformed transfer learning across most models.

A similar study by Sadiq et al. (2022) investigated ten pretrained CNNs for mental and motor imagery EEG classification. They compared the same two strategies for leveraging pretrained models as Raghu et al. (2020). Interestingly, they derived at contrasting results: transfer learning consistently outperformed feature extraction across all architectures by more than 10% validation accuracy.

The contradictory outcomes of these methods may be attributed to differences in the EEG domain, namely epileptic seizure detection and motor imagery data. This raises an important question: which approach is more suitable for EEG-based emotion recognition? The divergence in findings underscores the need for a systematic and controlled comparison of transfer learning and feature extraction strategies within the specific context of emotion recognition.

2.4. Summary and Research Gap

This literature review examined key techniques for EEG-based emotion recognition using pretrained CNNs. It introduced the fundamental characteristics of EEG data, including its complex waveforms and frequency bands, and discussed their relevance to emotional state classification.

Various EEG preprocessing techniques such as filtering, artifact removal, channel selection, and segmentation were introduced. Feature extraction methods spanning time-domain, frequency-domain and time-frequency domain were reviewed such as 2D EEG feature representations like spectrograms and stacked spectrograms.

The application of CNNs in EEG classification was analyzed, emphasizing their ability to learn spatial and temporal features. Finally, the role of leveraging pretrained CNNs in emotion recognition was examined, emphasizing its significance in addressing challenges like sparse data and computational ressources through methods including transfer learning and feature extraction.

The use of pretrained CNNs for EEG-based emotion classification has gained increasing traction in research, reflecting broader advances in Deep Learning for neurophysiological signal analysis. However, several research gaps remain underexplored:

1. Comparative Analysis of Preprocessing and Image Generation Techniques

   Most existing studies adopt fixed preprocessing and image generation pipelines, limiting the ability to compare their relative effectiveness. This study systematically evaluates a range of preprocessing and image generation configurations to identify the most effective combinations for EEG-based emotion recognition.

2. Evaluation of EEG Feature Representations

   While 2D representations like spectrograms are commonly used, few studies compare different spatial encodings within the same experimental framework. Stacked spectrograms, in particular, have not yet been evaluated for emotion recognition, despite their use in related EEG classification tasks. This work addresses this gap by comparing per-channel and stacked spectrograms to determine their effectiveness in emotion classification.

3. Evaluation of Pretrained CNN Architectures

Different CNN architectures have demonstrated success in various EEG emotion recognition tasks, yet pretrained models are often heavily modified, leaving their overall effectiveness in this domain uncertain. This study addresses this gap by systematically evaluating the performance of different pretrained CNNs such as MobileNetV2, ResNet50, InceptionV3 and DenseNet121.

4. Strategies for Leveraging Pretrained CNNs: Transfer Learning vs. Feature Extraction

In EEG classification, studies in domains such as seizure detection and motor imagery have reported conflicting results regarding the relative effectiveness of feature extraction and transfer learning when using pretrained networks. This study directly investigates these model adaptation techniques in the context of EEG-based emotion recognition, aiming to provide comparative insights into their performance under a unified evaluation framework.

By systematically evaluating different EEG preprocessing techniques, feature representations, pretrained CNN architectures, and modeling strategies, this study aims to provide valuable insights into the most effective approaches for EEG-based emotion recognition.

3. METHODOLOGY

This chapter outlines the methodological framework used to investigate EEG-based emotion recognition through spectrogram representations and pretrained CNNs. It details the dataset, such as preprocessing and image generation configurations. To ensure robustness, various of these configurations will later be empirically tested and evaluated. Finally, the evaluation framework for the emotion recognition task is introduced, forming the basis for the experiments presented in Chapter 4.

3.1. DATASET

This study utilizes the SEED (SJTU Emotion EEG Dataset), a dataset specifically designed for research in EEG-based emotion recognition (W. L. Zheng & Lu, 2015). The dataset comprises recordings from 15 participants, each of whom completed three experimental sessions spaced approximately one week apart. During each session, participants watched 15 carefully selected film clips intended to elicit positive, negative, or neutral emotions.

EEG signals were recorded using the ESI NeuroScan System with 62 scalp electrodes arranged according to the international 10–20 system. The EEG signals were originally sampled at 1000 Hz and subsequently down sampled to 200 Hz to reduce computational load and facilitate analysis. A 2D representation of the electrode configuration is shown in Figure 3.1 (a), along with a 3D representation in Figure 3.1 (b).

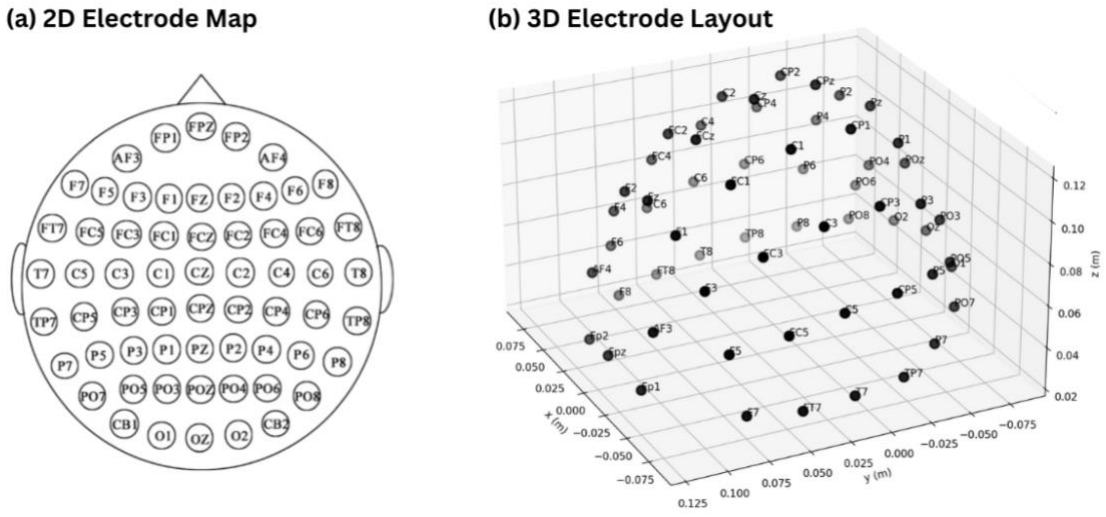

Figure 3.1: Electrode layout representations following the international 10–20 system

The duration of the film clips ranged from 3:04 to 4:26 minutes, with an average length of 3:38 minutes. Each clip represented a single trial, resulting in a total of 45 trials per subject across the three sessions. After viewing each clip, participants completed a self-assessment to report their experienced emotional state, providing subjective validation for the intended emotional labels.

The use of film clips as emotion-inducing stimuli is further supported by findings from Zhuang et al. (2018), who showed that movie-induced emotions produce EEG patterns closely resembling those of self-induced emotions, reinforcing the validity of this experimental approach.

This dataset has become a widely adopted benchmark in the field of EEG-based emotion recognition, making it highly suitable for assessing how effectively deep learning models can capture and interpret emotional processing.

3.2. PREPROCESSING & IMAGE GENERATION

The transformation of raw EEG recordings into spectrogram representations suitable for CNNs involves a series of preprocessing and image generation steps. Each of these steps can significantly affect the resulting data quality and model performance. The following subsections describe these procedures and outline the rationale behind each design choice.

Given the variability inherent in EEG signals, multiple configurations were considered and empirically explored. A systematic performance-based evaluation of the configurations is presented in the Pre-Evaluation in Chapter 4.1, where the most effective pipeline is identified.

3.2.1. Preprocessing

Preprocessing aims to enhance the signal quality, reduce noise, and standardize the data prior to the image generation. Core steps in this process include bandpass filtering, additional artifact removal, normalization, and spectrogram transformation using the STFT. All procedures in this study were applied to a selected subset of 12 EEG channels: FT7, FT8, T7, T8, C5, C6, TP7, TP8, CP5, CP6, P7, and P8. This configuration was adopted based on findings from the original SEED study, which demonstrated that this specific subset outperformed the full 62-channel configuration in classification tasks (W. L. Zheng & Lu, 2015).

The first step in the preprocessing pipeline is bandpass filtering, which removes low-frequency drifts and high-frequency noise outside the frequency band relevant for EEG-based emotion recognition. In this work, filtering was applied using a finite impulse response (FIR) filter, consistent with common practice in EEG preprocessing (Bagherzadeh et al., 2024).

In this study, several bandpass configurations were explored: 0.3-50 Hz, 1–50 Hz, 1–70 Hz, and 1–90 Hz. The final setting was selected based on performance comparisons detailed in Section 4.1.1. Figure 3.2 illustrates the effect of bandpass filtering (1–90 Hz) on raw EEG data of two channels. After filtering, the signals display fewer extreme oscillations and appear more stable and visually coherent. Although amplitude variations remain across channels, the filtered signals are better aligned in scale and more suitable for downstream analysis.

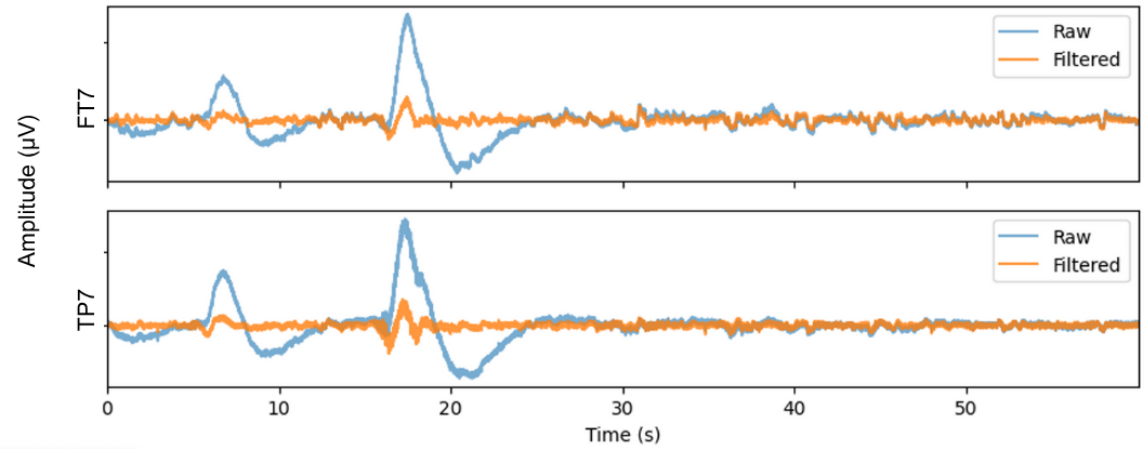

Figure 3.2: Comparison of raw and filtered EEG signals

In addition to bandpass filtering, further artifact removal can improve EEG signal quality by addressing noise sources such as ocular, cardiac, and muscular activity. In this study, two automated artifact correction techniques were explored: Wavelet-based denoising and Automatic Tunable Artifact Removal (ATAR). Both methods were implemented using the SPKIT library and applied to the raw EEG signal.

For wavelet denoising, the signal was decomposed using the Daubechies 3 (db3) wavelet. An optimal threshold based on the method by (Donoho & Johnstone, 1994) was applied to remove low-magnitude coefficients, and the signal was then reconstructed from the remaining components. For ATAR, the parameter β, which controls the aggressiveness of artifact suppression, was set to 0.2, as higher values did not show noticeable improvement in preliminary tests. A comparative evaluation of these artifact removal methods is presented in Chapter 4.1.1.

To ensure consistency in the input representations, channel-wise normalization was applied independently for each trial. Specifically, the signal values of each EEG channel were standardized by subtracting the mean and dividing by the standard deviation computed over that trial. This transformation preserves the temporal and spectral patterns within each channel while placing the data on a comparable numerical scale.

Although this normalization does not affect the relative dynamics of the signal, it helps reduce scale differences between channels, which may be caused by differences in how well the electrodes are connected to the scalp and signal strength. Since the resulting spectrograms are computed per channel and processed independently by the model, using a common statistical scale helps reduce input variability and may support more consistent learning across different channels.

Baseline correction was not applied in this study, as the SEED dataset does not include dedicated baseline recordings for each subject. While an alternative approach could involve using the average signal from neutral-labeled trials as a surrogate baseline, this method was

considered unreliable due to potential variability in emotional responses, which could introduce inconsistencies and reduce the effectiveness of the correction.

To convert EEG signals into time–frequency representations, the STFT was applied. As priorly explained, STFT decomposes the signal into overlapping windows and computes the Fourier transform within each segment. It is defined in Equation 3.1 (Mandhouj et al., 2021; Pusarla et al., 2022):

$$STFT\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - m] \cdot e^{-jwn} \quad (3.1)$$

where $x[n]$ is the input signal at time n and $w[n - m]$ the window function. In this study, a Hamming window was used, which is commonly employed in EEG analysis (Pusarla et al., 2022; Abdulwahhab et al., 2024)). The window function is defined in Equation 3.2:

$$w(n) = 0.5 \left[1 - cos\left(\frac{2\pi n}{M-1}\right)\right], \quad 0 \leq n \leq M-1 \quad (3.2)$$

where n is the window length and M the sampling number (F. Wang et al., 2020).

A window length of 200 samples (corresponding to 1 second, given the 200 Hz sampling rate of the SEED dataset) and 50% overlap was selected as this configuration is commonly used in EEG-related studies (Lu et al., 2012). This configuration achieves a frequency resolution of 1 Hz and 120-time bins for a 60-second segment, offering a practical balance between time and frequency resolution critical for capturing EEG patterns related to emotional processing.

The STFT values are further used to produce a spectrogram, which visually represents how the signal's strength is distributed across time and frequency. Specifically, each value in the spectrogram represents the power at a specific time-frequency bin, calculated as the squared magnitude of the STFT coefficients (Mandhouj et al., 2021).

3.2.2. Image Generation

Following the preprocessing pipeline and the calculation of STFT-based spectrogram values, the next step involves defining the settings used for image generation. This process includes applying a logarithmic scale, segmenting the data, such as selecting appropriate color mapping and interpolation methods.

The power values of the spectrogram vary widely in scale, strong low-frequency components can overshadow weaker patterns at higher frequencies. To address this, a logarithmic transformation is applied, which compresses the dynamic range and enhances the visibility of subtle spectral features across the full frequency axis. Table 3.1 illustrates the effect of this transformation across different magnitudes:

Table 3.1: Effect of logarithmic scaling on power values

Power (V^2/Hz)	0,01	0,1	1	10	100
Log Power (10 + $\log_{10}$(Power))	-20	-10	0	10	20

Figure 3.3 compares spectrograms without and with log scaling. The log-scaled version reveals finer patterns, particularly in higher frequencies, likely making the input more informative for CNNs trained on visual representations.

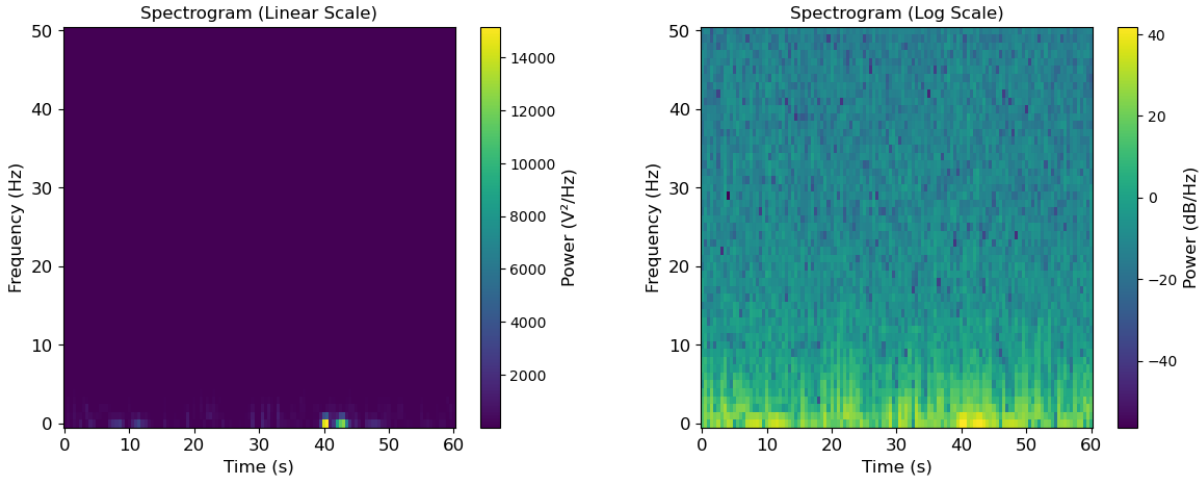

Figure 3.3: Spectrogram before and after logarithmic transformation

After generating the spectrograms, a critical decision involves selecting the segment length used for training the CNN. As priorly mentioned, studies employ varying segment lengths ranging from a few seconds to several minutes. Longer segments preserve temporal context and provide richer information but result in fewer training samples, which may limit model generalization. Shorter segments increase the number of training examples, possibly improving the model's capacity to generalize, while also reducing the temporal resolution and potentially less informative input.

This study compares different segment lengths, particularly 3 minutes, 1 minute, 30 seconds and 10 seconds. Figure A.2 in Appendix A illustrates how the varying segment lengths influence the resulting spectrograms.

To address the limited availability of training data, overlapping segments can be used to expand the dataset while preserving temporal resolution and potentially improving generalizability. However, excessive overlap may introduce redundancy, increasing the risk of overfitting. Therefore, this study proposes a 25% overlap for one-minute spectrogram

segments as a practical compromise. Figure 3.4 illustrates this segmentation process for a single trial.

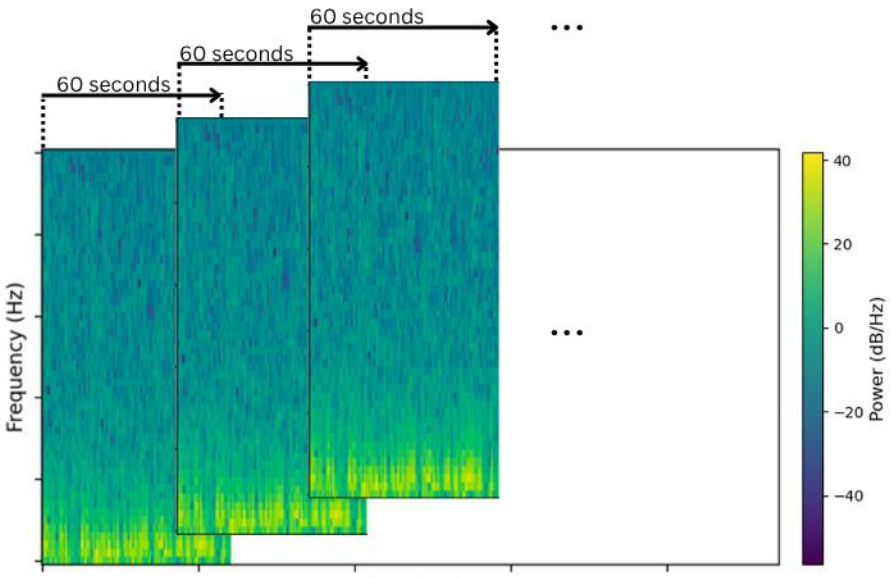

Figure 3.4: Segmentation using 25% overlap for 60-second spectrogram segments

The varying segment lengths were evaluated empirically in the pre-evaluation phase (see Section 4.1.2), and the best-performing configuration was selected for further use. By comparing both long and short segments, the study captures both sustained emotional responses and finer temporal variations.

Another important factor in image-based EEG classification is how spectrogram values are visually represented through color mapping. To ensure consistent input for the CNN, the color scale should remain uniform across all trials. Therefore, the same color should correspond to the same signal intensity level.

This study compares different approaches for defining the color range in spectrogram visualization on a per-subject basis. In the first approach, the global minimum and maximum values across all trials were used to set the lower and upper bounds of the colormap, preserving the full dynamic range. In the second, percentile-based thresholds were applied to clip extreme values, reducing the influence of outliers and enhancing the contrast of the image. Both methods provide consistent visual representation across trials but highlight different aspects of the data.

Figure 3.5 shows how the selected value range impacts the appearance of the spectrogram. Using the full range of values (a) can result in low-contrast representations that obscure important details. Alternatively, restricting the scale to percentiles (e.g., 0.1%–99% or 1%–99%) produces higher contrast images (b and c), where frequency changes are more pronounced.

Additionally, interpolation can be applied to improve image smoothness. In this work, bilinear interpolation was used (see Figure 3.5 (d)), where each output pixel is computed as a weighted average of the four nearest neighbors, resulting in smoother transitions. While interpolation does not alter the underlying data, it may influence how pretrained image models interpret the spectrograms.

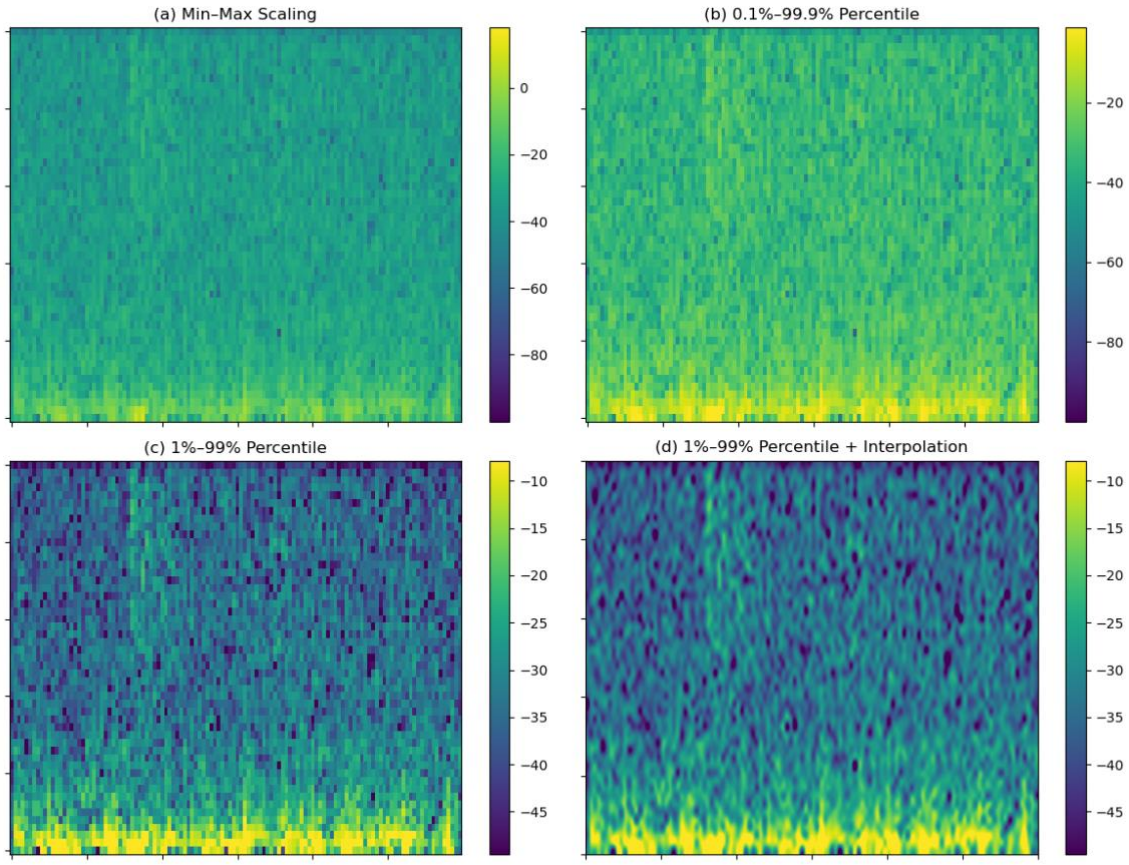

Figure 3.5: Spectrograms under varying scaling and rendering techniques

Since the effect of these visual formatting choices on model performance is not yet explored, multiple configurations—including color scale bounds and interpolation—were empirically tested (see Section 4.1.2) to determine the most effective settings for emotion recognition from EEG spectrograms. These results may also provide beneficial insights for other EEG-based tasks that rely on time-frequency representations.

As previously discussed, conventional spectrograms represent the signal from a single EEG channel. However, this channel-wise approach may not fully capture the spatial dynamics of brain activity. To address this, stacked spectrograms were additionally constructed by horizontally concatenating the spectrograms of the 12 selected channels into a single image. This approach is inspired by related work (Raghu et al., 2020) used for EEG based seizure prediction and aims to preserve both temporal and spatial structure across channels. The

resulting images were resized to match the input dimensions of the pretrained CNNs. An example of such a stacked spectrogram is shown in Figure A.3 in Appendix A.

Although data augmentation techniques such as random flipping or rotation are widely used in general image classification tasks (Z. Wang et al., 2024), they were not applied in this study. Given that spectrograms represent structured time–frequency information along fixed axes, such transformations could possibly distort the semantic layout of the data and hinder model learning.

3.3. EVALUATION FRAMEWORK

As outlined in the preceding chapter, both preprocessing and image generation configurations significantly influence the outcome of the 2D feature representation. In parallel, different strategies for leveraging pretrained models have been discussed to efficiently train a classification model that can be adapted for emotion recognition. To systematically address these components, this study adopts a two-stage machine learning framework:

1. Pre-Evaluation Phase, aimed at identifying optimal preprocessing and image generation parameters.

2. Final Evaluation Phase, which focuses on comparing the performance of different CNN-based architectures under two model adaptation strategies.

Together, these phases form a structured evaluation framework outlining the most effective methods during data preparation and model training. Before detailing the two evaluation phases, the evaluation protocol used to assess performance is introduced below.

3.3.1. Evaluation Protocol

This study follows the evaluation protocol proposed by (W. L. Zheng & Lu, 2015) in their original paper in which they introduced the SEED database. This evaluation method is further applied across several additional studies in the field (e.g., F. Wang et al., 2020; Song et al., 2020, W. Liu et al., 2016). In this protocol, the first 9 trials of each session are used as the training set, while the last 6 trials are reserved for validation. Evaluation is conducted on a per-subject basis, and the final performance metric is computed as the average classification accuracy across subjects.

This original dataset split strategy is chosen to ensure direct comparability with prior work using the SEED dataset. While cross-validation is commonly used in other contexts, the original protocol enables consistent benchmarking and reduces computational cost.

To assess model performance, two common evaluation metrics are employed: accuracy and F1-score. Accuracy measures the proportion of correctly classified instances, while the F1-score provides a more balanced assessment especially useful in uneven class distributions.

The mathematical definitions of these metrics are given in Equations 3.3 and 3.4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (3.3)$$

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN} * 100 \quad (3.4)$$

where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively (Sadiq et al., 2022).

Additionally, emotion classification tasks are typically assessed with confusion matrices to visualize per-class prediction performance (Pusarla et al., 2022). These matrices provide a detailed breakdown of correct and incorrect classifications across all emotion categories, enabling a more granular interpretation of the model's strengths and weaknesses.

To generate trial-level predictions from the model's spectrogram-level outputs, an aggregation mechanism is required. In the case of per-channel spectrograms, a majority voting scheme is applied across all channel-level predictions, as illustrated in Figure 3.6 Each trial consists of multiple segments, and each segment contains EEG signals from 12 channels producing 12 independent spectrograms. For each segment, the final label is determined by selecting the most frequent label among the 12 channel-level predictions.

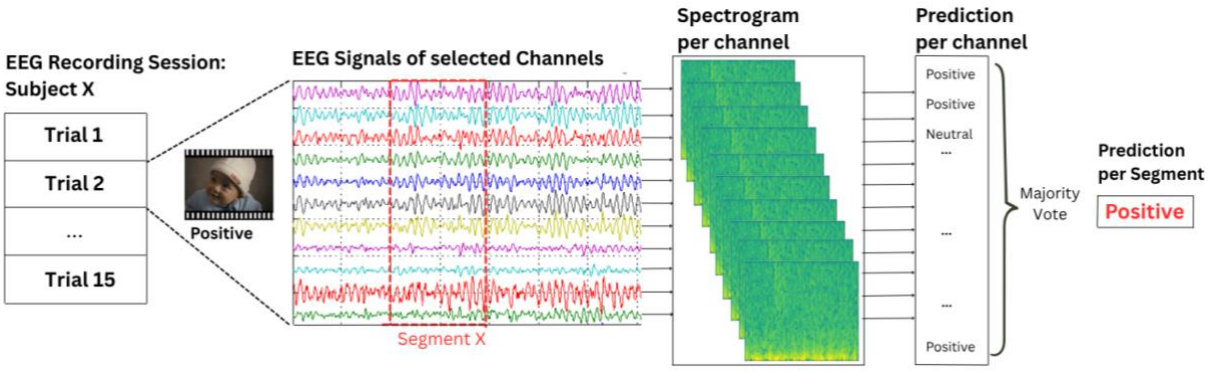

Figure 3.6: Majority voting across EEG channels to derive at segment-level predictions

In contrast, stacked spectrograms integrate information from all channels into a single input, yielding a unique prediction per segment by design. A second majority voting step is then applied across all segments within a trial to determine the final emotion label per trial. This voting strategy ensures that the final prediction reflects the consensus of multiple inputs while remaining computationally simple.

3.3.2. Pre-Evaluation

The Pre-Evaluation Phase establishes the foundations for the final evaluation by systematically comparing various preprocessing and image generation configurations previously described. Its objective is to identify a robust and computationally efficient pipeline that maximizes model performance in EEG-based emotion recognition.

Due to the high computational cost associated with evaluating all possible combinations of parameters, this phase is conducted solely on Subject 1 of the SEED dataset. The most effective techniques derived from this analysis are then applied in the Final Evaluation Phase, which is then held across different subjects to enable a more generalized performance. The parameter space explored in this phase covers two major stages of the pipeline: signal preprocessing and image generation.

In the preprocessing stage, different bandpass filter configurations are tested alongside two additional artifact removal strategies, namely ATAR and wavelet-based denoising. For image generation, color mapping strategies based on min-max value limits and percentile-based clipping are compared, as well as the effect of different interpolation settings. Additionally, varying EEG segment lengths are evaluated to establish the most informative input configuration.

Finally, the three learning rates (LR) 0.001, 0.0001, and 1e-5 are tested to identify the most effective configuration for use in the subsequent final evaluation phase. The two pretrained CNNs ResNet50 and MobileNetV2 are modified by replacing the final layers to match the target task and are utilized for evaluation through transfer learning. Training is performed using the ADAM optimizer. Each configuration is trained for 30 epochs with a batch size of 32.

3.3.3. Final Evaluation

Following the identification of optimal preprocessing and 2D feature representation techniques in the pre-evaluation phase, the final evaluation phase assesses the overall EEG-based emotion classification performance across different modelling techniques. Specifically, this study compares different spectrogram representations (per-channel vs. stacked) and CNN architectures pretrained on ImageNet. The pretrained models are evaluated using two distinct strategies for leveraging pretrained CNNs:

A) Transfer Learning using pretrained network
B) Feature Extraction using pretrained network

Method A refers to replacing the final layers of the pretrained CNN to suit the emotion classification task. More specifically, a global average pooling layer is used to flatten the output of the final convolutional block, followed by a dropout layer for regularization, and a fully connected layer with three output units corresponding to the emotion classes. A softmax activation completes the architecture, enabling probabilistic multi-class classification. The

modified network is then fine-tuned on EEG-based emotion recognition data, allowing it to adapt to the new target task.

Method B by contrast, treats the CNN as a fixed feature extractor, resulting in a flattened one-dimensional feature vector that represents the image. These feature representations can be stored externally and subsequently be passed to an external classification model. A Support Vector Machine (SVM) is commonly employed for this purpose (Asghar et al., 2019) and is likewise adopted in this work.

This approach is adapted from prior work by Sadiq et al. (2022) and Raghu et al. (2020), who applied these two methods to classify EEG signals in the contexts of motor imagery and seizure prediction, respectively. The complete sequence of processing steps is visualized in Figure 3.7.

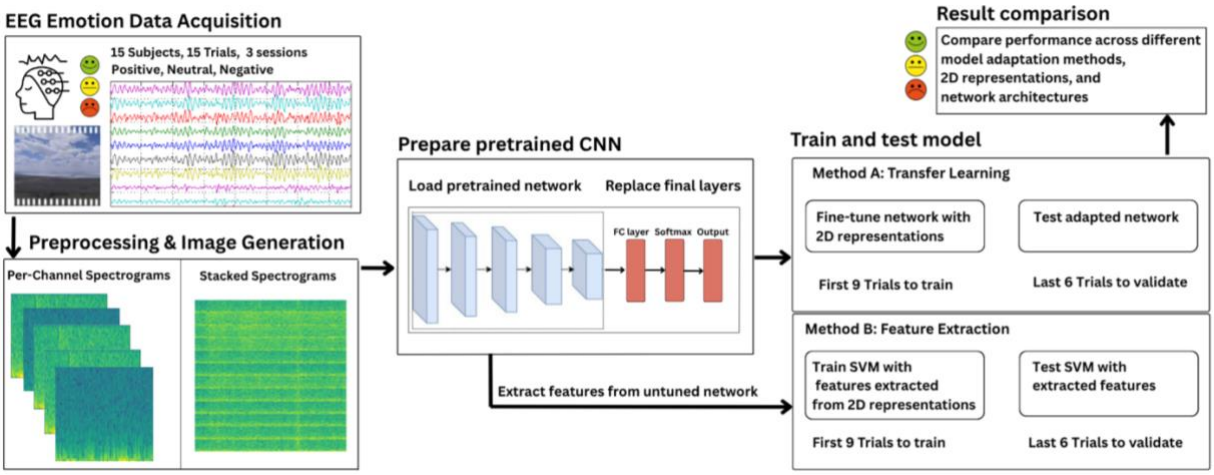

Figure 3.7: EEG emotion recognition workflow using 2D representations and two methods for leveraging pretrained CNNs

Table 3.2 summarizes the pretrained CNN architectures used in this study: MobileNetV2**,** ResNet50**,** InceptionV3**,** and DenseNet121 along with key architectural characteristics.

Table 3.2: Architectural characteristics of pretrained CNNs used in this study

Layers	**MobileNetV2**	**ResNet50**	**InceptionV3**	**DenseNet121**
Input Layer	(224, 224, 3)	(224, 224, 3)	(229, 229, 3)	(224, 224, 3)
Pretrained Model Output	(7, 7, 1280)	(7, 7, 2048)	(5, 5, 2048)	(7, 7, 1024)
Global Average Pooling	(1, 1280)	(1, 2048)	(1, 2048)	(1, 1024)
Total number of parameters	2.261.827	23,593,859	21,808,931	7,040,579

It includes the input dimensions expected by each network, the shape of the feature maps produced by the final convolutional block, and the total number of parameters.

These four architectures were selected to represent a diverse range of CNN designs, balancing computational efficiency, depth, and representation power. MobileNetV2 provides a lightweight, mobile-friendly architecture with relatively few parameters, while ResNet50, InceptionV3, and DenseNet121 are deeper models capable of capturing more complex hierarchical features. All models are widely used in image-based transfer learning tasks and come with well-established pretrained weights on the ImageNet dataset, making them suitable for adaptation to spectrogram-based EEG classification.

It is important to note that this thesis does not aim to analyze the internal design of these architectures. Instead, the focus lies in evaluating how these pretrained models perform under different adaptation strategies and spectrogram representations. As such, detailed architectural comparisons are beyond the scope of this work.

To contextualize the results, a baseline classifier is included using a SVM trained on PSD features from the original SEED publication, representing a traditional EEG classification pipeline without transfer learning.

4. RESULTS AND DISCUSSION

This chapter presents and interprets the experimental results obtained in both the pre-evaluation phase and the final evaluation, followed by a discussion of their implications in the context of EEG-based emotion recognition. Further performance patterns are briefly analyzed across trials and channels to provide deeper insight into the spatial and temporal dynamics relevant to emotion recognition.

4.1. PRE-EVALUATION RESULTS

This section summarizes the impact of various preprocessing and image generation parameters on emotion recognition performance. In addition, the effect of a different LR is evaluated to identify optimal training settings. The objective is to determine a robust and computationally efficient configuration for the final evaluation phase. All experiments are performed on data from the first subject, using transfer learning with ResNet50 and MobileNetV2 architectures pretrained on ImageNet.

4.1.1. Preprocessing

The first parameter evaluated in the pre-evaluation phase is the bandpass filter configuration. In this study, four configurations are compared: 0.3–50 Hz, 1–50 Hz, 1–70 Hz, and 1–90 Hz. The frequency axis of the spectrograms is dynamically adjusted to reflect the selected range.

As shown in Figure 4.1 (a), both ResNet50 and MobileNetV2 achieve higher validation accuracy when higher frequency bands are included. These findings suggest that higher-frequency components, particularly within the gamma band, play a significant role in enhancing emotion recognition performance. This observation is consistent with research indicating that emotional processing is particularly reflected in higher frequency bands (X. Li et al., 2022).

Furthermore, setting the lower frequency cutoff at 1 Hz instead of 0.3 Hz yields better results, possibly due to reduced low-frequency noise. The 1–90 Hz bandpass filter configuration proves to be the best-performing setup overall, capturing a broad and informative range of frequencies relevant to emotion recognition. Across all filter configurations, ResNet50 consistently outperforms MobileNetV2, a trend observed throughout subsequent experiments.

In addition, Figure 4.1 (b) demonstrates the strong impact of majority voting across channels. Instead of relying on per-channel predictions independently, aggregating the outputs across all channels improves validation accuracy from 0.56 to 0.83 for the 1–90 Hz frequency range. This result highlights the variability in signal quality across channels and confirms the stabilizing effect of majority voting for trial-level classification.

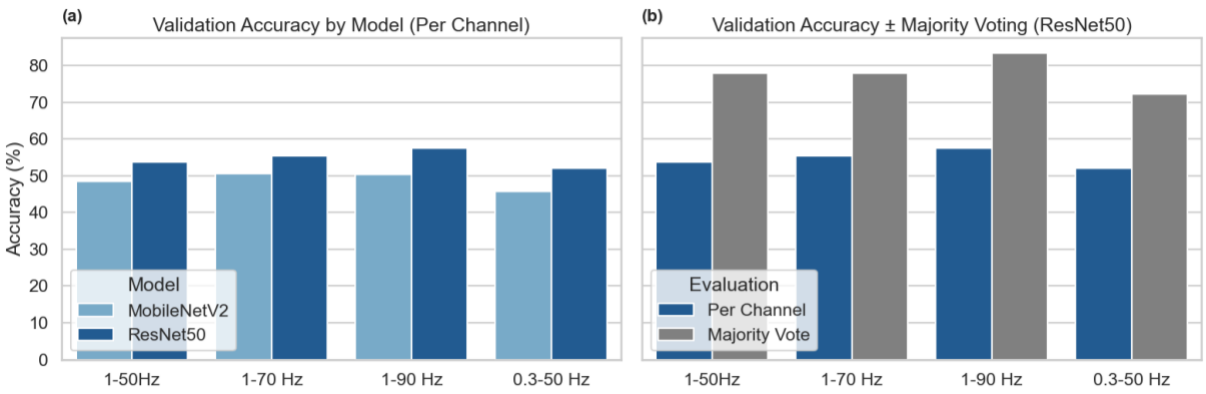

Figure 4.1: Impact of bandpass filtering and majority voting on validation accuracy

Next, additional artifact removal is examined as part of the preprocessing pipeline. Two techniques are considered: Wavelet Denoising and the ATAR algorithm.

As depicted in Figure 4.2, both artifact removal methods show no substantial impact for per channel predictions. However, when combined with majority voting, the best overall performance is achieved without any additional artifact removal. This can be attributed to the fact that majority voting already mitigates the effect of noisy channels by aggregating predictions, reducing the benefit of denoising. Furthermore, additional artifact removal might not only suppress noise but also informative signal components.

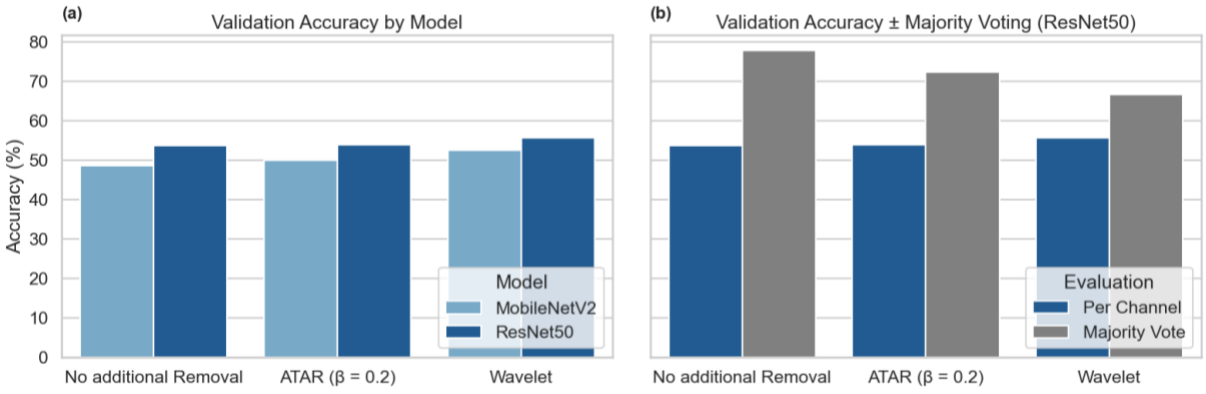

Figure 4.2: Impact of additional artifact removal and majority voting on validation accuracy

4.1.2. Image Generation

The color mapping range and interpolation configurations can significantly affect the resulting input representations. As discussed in Section 3.2.2, these parameters influence the dynamic range and smoothness of the spectrograms, which can in turn affect classification performance.

In this study, three color range strategies were compared: percentile-based thresholds at 99%/1% and 99.9%/0.1%, as well as a min-max value range computed across all trials for a single subject. Additionally, the impact of bilinear interpolation was evaluated for the 99%/1% condition.

As shown in Figure 4.3, percentile-based color mapping, particularly at 99.9% / 0.1%, consistently yields the best classification performance across models. In contrast, using the full min-max value range results in significantly lower accuracy. This may be because using a wide dynamic range spreads the intensity values too sparsely, which reduces visual contrast in the spectrograms and makes it more difficult for the CNN to detect relevant features. Meanwhile, the use of bilinear interpolation led to a decrease in validation accuracy. This suggests that a more abrupt representation of the spectrogram may better preserve local structures and facilitate model training.

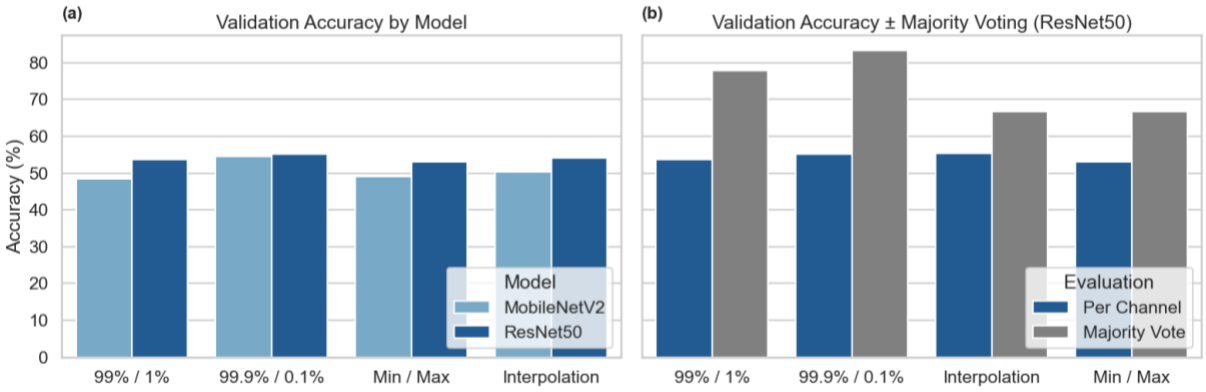

Figure 4.3: Impact of color mapping, interpolation and majority voting on validation accuracy

Furthermore, various segment lengths were tested to examine the trade-off between data granularity, model performance, and computational cost. Specifically, the following configurations were evaluated: 10 seconds, 30 seconds, 1 minute, 1 minute with 25% overlap, and 3-minute segments. Notably, shorter segments significantly increase the number of training samples from 2,169 images for 1-minute segments to over 11,880 images for 10-second segments per subject.

Figure 4.4 (a) shows that training with 10-second segments extends computation time to over 90 minutes for a single subject for ResNet50, compared to approximately 30 minutes for 1-minute segments. As expected, ResNet50 requires more computational resources than the lighter MobileNetV2, due to its deeper architecture.

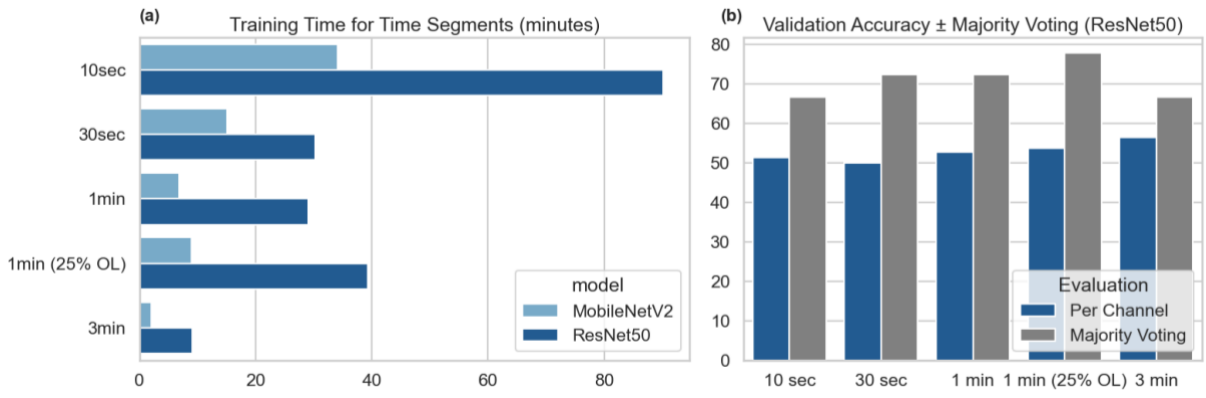

Figure 4.4: Impact of varying time segments on training time and validation accuracy

As shown in Figure 4.4 (b), pretrained CNNs achieve higher accuracy with longer segments, possibly due to better temporal context and reduced noise in longer EEG windows. Among the tested options, the proposed 1-minute segments with 25% demonstrated the best overall performance and is thus adopted for the final evaluation.

Based on the pre-evaluation results, the settings summarized in Table 4.1 were selected for the subsequent experiments.

Table 4.1: Final preprocessing and image generation parameters

Bandpass filter	**Artifact removal**	**Color mapping**	**Interpolation**	**Segment length**
1-90Hz	None	99.9% percentile	None	1 minute, 25% Overlap

As summarized in Figure A.4 in Appendix A, these optimized parameters improved validation accuracy for the first subject from 66.7% to 83.3%, demonstrating the importance of careful parameter tuning prior to full-scale evaluation. The optimal settings will further be applied to the stacked spectrograms, which integrate all channels into a single representation.

4.1.3. Learning Rate

Next, the optimal learning rate is examined by evaluating model performance for the final spectrogram representation. Figure 4.5 shows the training and validation curves before majority voting for MobileNetV2 and ResNet50 using three LRs applied to the first subject.

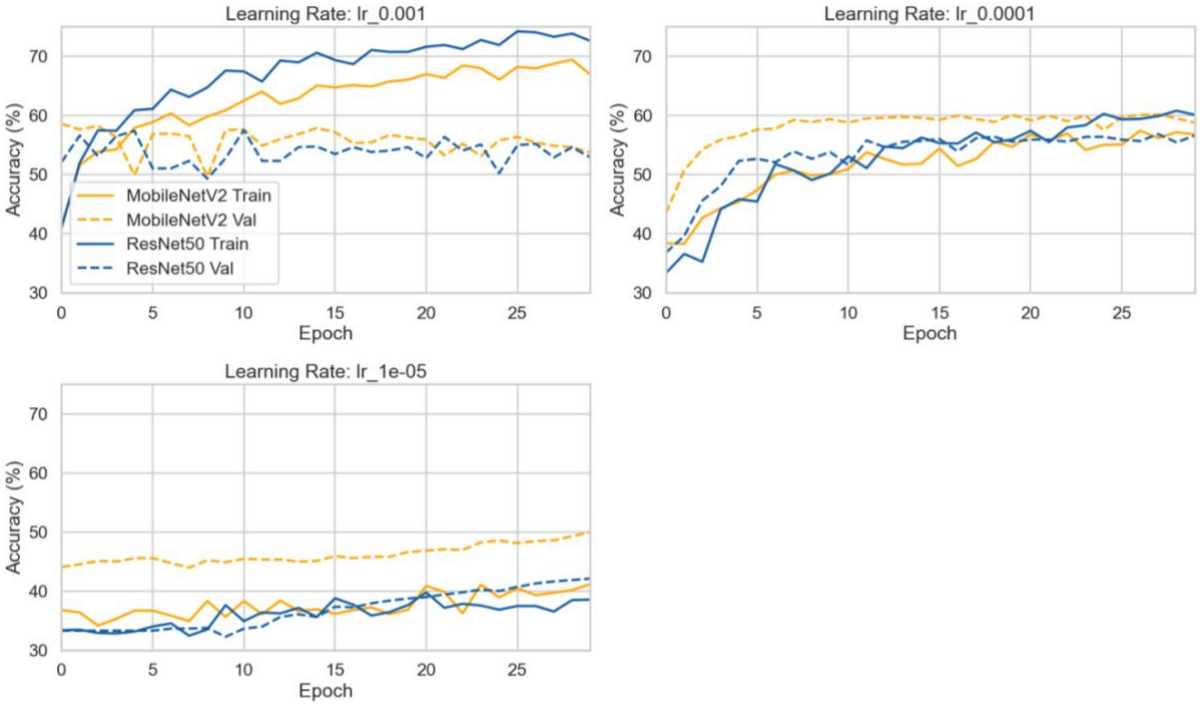

Figure 4.5: Model convergence across learning rates for final spectrogram configuration

A LR of 0.001 leads to no significant improvement in validation accuracy and substantial overfitting as training performance consistently exceeded validation performance. This suggests that the step sizes are too large, causing resulting in unstable learning and poor generalization.

In contrast, a LR of 1e-5 results in the lowest overall performance. The model appears to underfit, as the training accuracy remains significantly lower than the validation accuracy. This discrepancy may be due to the presence of a Dropout layer, which is active only during training. As a result, predictions made during evaluation may be more stable.

The most favorable outcome is observed with a LR of 0.0001, which yields the highest validation accuracy for both model architectures and the most stable convergence during training. This aligns with prior findings, where 0.0001 was also identified as an optimal LR for fine-tuning pretrained CNNs on EEG datasets (Sadiq et al., 2022). Accordingly, a LR of 0.0001 is selected for the final evaluation across multiple subjects. The Dropout rate is set to 0.2, as higher values led to severe underfitting in earlier experiments.

Although prior work suggests benefits in fine-tuning deeper CNN layers to learn task-specific features, preliminary experiments in this study revealed severe overfitting when additional layers were updated. As shown in

Figure A.5, updating the last 10 or 20 layers increasingly reduced generalization and performance. This is likely caused by the limited size of the training data, which is insufficient to support updating a high number of parameters in deeper layers. Consequently, this study limits fine-tuning to the final classification layer to ensure stability.

When using stacked spectrograms, a higher LR of 0.001 results in improved performance compared to a lower value of 1e-5, as shown in Figure 4.6. This is likely due to the smaller training set size in the stacked setup, where each segment corresponds to a single input image rather than one per channel, necessitating larger updates to achieve effective learning. Consequently, a LR of 0.001 is used for training with stacked spectrograms.

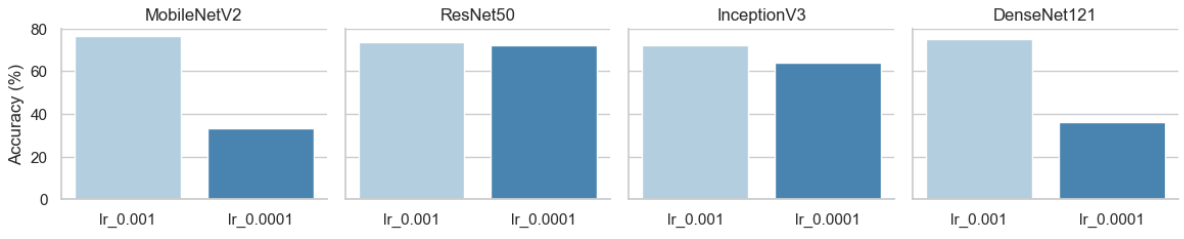

Figure 4.6: Validation accuracy across models for varying learning rates for spectrogram stacks

4.2. FINAL EVALUATION RESULTS

With the optimal preprocessing and spectrogram generation settings established in the pre-evaluation phase, the groundwork is set for the final evaluation phase, as outlined in Section 3.3.3. The evaluation is conducted on the last five subjects of the SEED dataset and includes four pretrained CNN architectures: ResNet50, MobileNetV2, InceptionV3, and DenseNet121. Per-channel and stacked spectrograms are employed as input representations and evaluated using both strategies for leveraging pretrained networks: transfer learning and feature extraction.

4.2.1. Transfer Learning using Pretrained CNNs

First, the transfer learning strategy will be evaluated, in which the final layers of the pretrained CNNs are replaced and fine-tuned using a softmax classifier.

Table 4.2 presents the classification accuracy and F1-scores achieved by each model using transfer learning applied to per-channel spectrograms. Performance varies notably across architectures, with ResNet50 achieving the highest average validation accuracy of 83,33% and F1-score of 83,07%, followed by DenseNet121 with an accuracy of 81.11% and F1-score of 80.29%. MobileNetV2 performs consistently worse across all subjects, with an average accuracy of 76,67%.

Table 4.2: Performance across models and subjects for per-channel spectrograms using transfer learning

	MobileNetV2	**ResNet50**	**InceptionV3**	**DenseNet121**
Subject 11	83,33	83,33	88,89	83,33
Subject 12	83,33	94,44	83,33	83,33
Subject 13	61,11	77,78	66,67	83,33
Subject 14	67,67	72,22	61,11	61,11
Subject 15	88,89	88,89	88,89	94,44
Accuracy Mean	76,67	**83,33**	77,78	81,11
F1 Mean	76,25	**83,07**	76,76	80,29

These results indicate that deeper architectures such as ResNet50 and DenseNet121 are better suited to capturing the complex temporal and spatial patterns in per-channel EEG spectrograms. The relatively close values between accuracy and F1-score suggest balanced

classification performance across emotion classes, although class-specific differences are further discussed in the following.

Confusion matrix analysis in Figure 4.7 reveals that all models perform best when classifying positive emotions. InceptionV3, for instance, achieves class-level accuracy as high as 97% for positive emotions, despite ResNet50 demonstrating the highest overall mean performance. This indicates that certain architectures may be particularly well-suited to detecting specific emotional states.

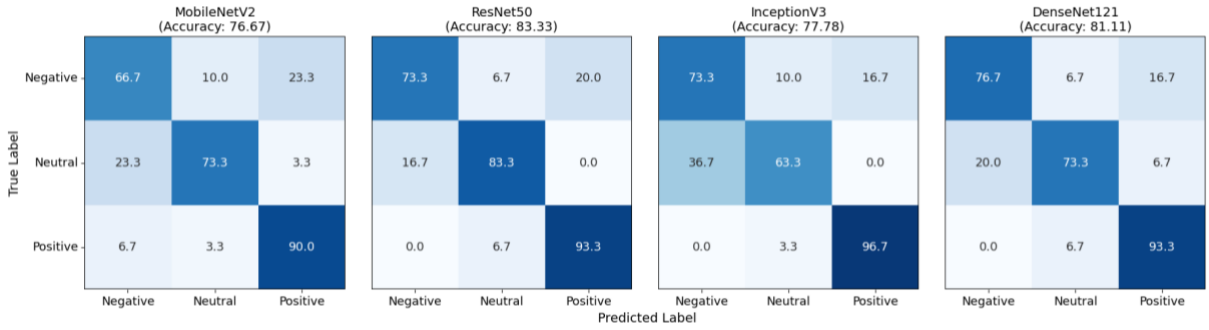

Figure 4.7: Confusion matrices across pretrained models for per-channel spectrograms using transfer learning

In contrast, neutral and negative emotions are more challenging to classify. Classification accuracy for neutral emotions ranges from 63% with InceptionV3 to 83% with ResNet50, while accuracy for negative emotions varies between 67% with MobileNetV2 and 77% with DenseNet121. Notably, no single model consistently outperforms across all emotional categories, as each architecture exhibits relative strengths for specific labels.

Misclassifications often occur between neutral and negative emotions, which are known to share overlapping EEG characteristics (J. Li et al., 2018). For example, InceptionV3 tends to confuse these emotions, misclassifying 36,7% of neutral instances as negative. These confusions may reflect the overlapping nature of brain responses to certain emotional stimuli.

As illustrated in Figure 4.8 there is considerable variation in performance across individual subjects.

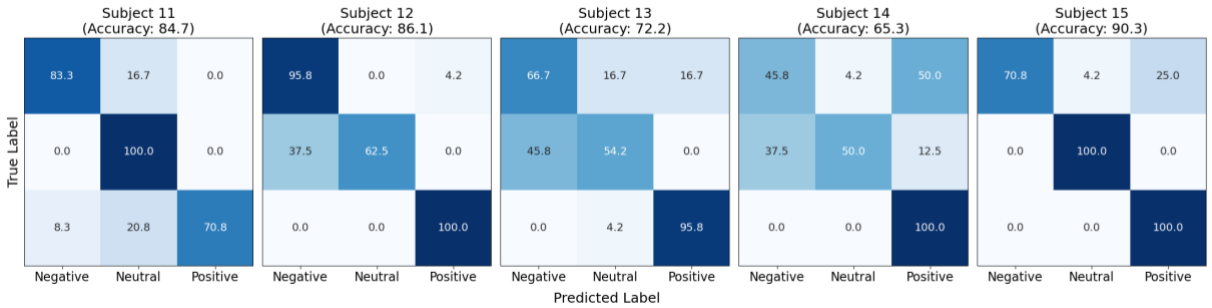

Figure 4.8: Confusion matrices across subjects for per-channel spectrograms using transfer learning

This observation is consistent with previous findings that highlight substantial inter-subject variability in EEG-based emotion recognition, particularly due to differences in regional brain activity and emotional perception (W. Li et al., 2024). These variations manifest not only in overall accuracy but also in the model's ability to distinguish specific emotional categories across subjects.

Average accuracy across all models ranges from 65% for Subject 14 to 90% for Subject 15, with Subjects 11, 12, and 13 achieving 85%, 86%, and 72%, respectively. Correct classification of positive emotions exceeds 95% for all subjects except Subject 11, where it drops to 70%. However, the distribution of false positives and false negatives varies considerably, reflecting individual differences in neural responses to emotional stimuli.

For instance, while Subject 11 exhibits clear signals for neutral emotions, Subjects 13 and 14 show frequent confusion between neutral and negative classes. These discrepancies may result from subjective differences in emotional perception. A video labeled as neutral might be experienced as humorous or irritating, thereby inducing atypical EEG responses. It is also possible that emotional responses in these subjects were weaker or more ambiguous, reducing the clarity of associated brain activity.

Next, results for stacked spectrograms are presented, in which data from all EEG channels are combined into a single input representation. As shown in Figure 4.9, the overall average validation accuracies across subjects are 80% for MobileNetV2, 82,22% for ResNet50, 77,78% for InceptionV3, and 75,56% for DenseNet121.

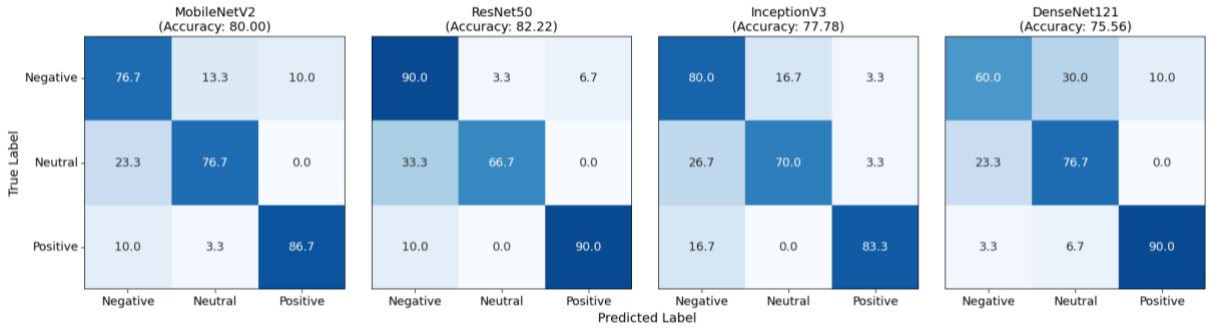

Figure 4.9: Confusion matrices across pretrained models for spectrogram stacks using transfer learning

Compared to per-channel spectrograms, the confusion matrices indicate a more balanced classification performance across all three emotional categories. While per-channel spectrograms exhibited a clear advantage in recognizing positive emotions, the stacked representation improves performance for negative and neutral emotions with a slight cost to positive emotion accuracy. This trade-off results in similar overall accuracy between the two representation methods.

Among the models, ResNet50 again achieves the highest overall accuracy. However, it exhibits a tendency to misclassify neutral emotions as negative, with 33% of neutral trials predicted

incorrectly. Conversely, DenseNet121 struggles with negative emotion recognition, misclassifying 30% of negative trials as neutral. These patterns are consistent with earlier observations.

Analyzing model performance across individual subjects offers additional insight into how subject-specific EEG patterns influence classification outcomes. As shown in Figure 4.10, average accuracies for Subjects 11 through 15 are 86%, 85%, 65%, 62%, and 96%, respectively. Consistent with earlier results using per-channel spectrograms, Subject 15 again achieves the highest accuracy, now improved by 6 percentage points, suggesting a strong alignment between this subject's EEG responses and the emotional content of the video stimuli.

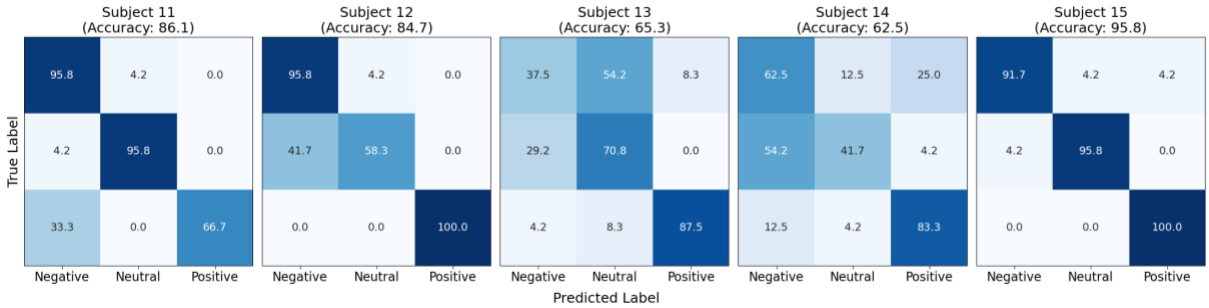

Figure 4.10: Confusion matrices across subjects for spectrogram stacks using transfer learning

Positive emotions remain the most reliably classified category across all subjects, with Subject 11 again exhibiting comparatively lower performance on this class. The distribution of correct predictions and misclassifications aligns with the overall trends observed for per-channel spectrograms, supporting the consistency of the findings across representation strategies. However, negative emotions are classified more accurately with stacked spectrograms, particularly for Subjects 11, 12, and 15, indicating that integrating spatial information across channels enhances the model's ability to distinguish negative affect.

Subjects 13 and 14 continue to show the lowest overall accuracy. These variations are likely influenced by multiple factors, including individual differences in emotional perception, regional brain activity, and signal quality. For example, Subjects 13 and 14 show a tendency to confuse neutral with negative or positive emotions, which may reflect subjectively ambiguous reactions to the video stimuli or weaker emotional responses overall. Such variability further underscores the importance of incorporating subject-specific considerations in model design and evaluation.

As shown in Figure 4.11 (a), the overall classification performance remains comparable between per-channel and stacked spectrogram representations, with only minor variations across architectures.

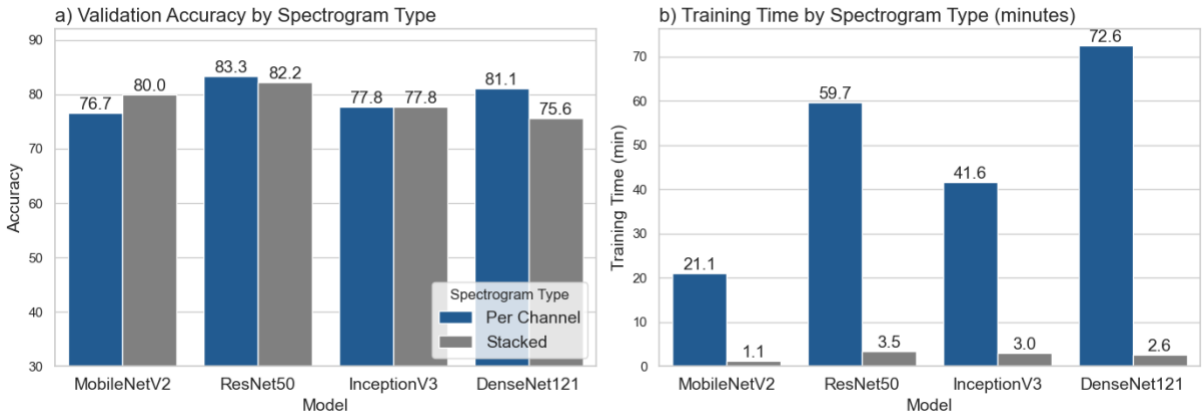

Figure 4.11: Performance comparison across spectrogram types and models using transfer learning

For MobileNetV2, stacked spectrograms yield slightly higher validation accuracy compared to per-channel inputs, suggesting that simpler models may benefit from the compactness of the stacked format. ResNet50 consistently achieves the highest performance across both configurations, with accuracies of 83,3% (per-channel) and 82,2% (stacked), indicating robustness to input format.

In contrast, DenseNet121 which performs well with per-channel spectrograms, achieves the lowest accuracy when trained on stacked inputs, possibly due to the limited dataset size, which may be insufficient for deeper architectures. Similarly, InceptionV3 performs comparably across formats but is slightly outperformed by MobileNetV2 in the stacked configuration, further supporting the idea that simpler architectures are more effective under data-constrained conditions.

Combining all channels into a single stacked spectrogram reduces the training set size by a factor of 12. Consequently, training times are significantly reduced, as shown in Figure 4.11 (b). With stacked spectrograms, the average model training times range from 1 to 3.5 minutes per subject, with MobileNetV2 being the fastest and ResNet50 the slowest. In contrast, training with per-channel spectrograms requires between 21 and 73 minutes per subject, with DenseNet121 incurring the highest computational cost.

From a performance perspective, per-channel spectrograms offer a slight advantage in overall accuracy, but the stacked representation provides substantial gains in training efficiency. It is plausible that with a larger dataset, stacked spectrograms could outperform per-channel inputs in both performance and efficiency, as they already show improved recognition of negative and neutral emotions, as previously discussed.

4.2.2. Feature Extraction using Pretrained CNNs

This section evaluates the use of pretrained CNNs as fixed feature extractors, with classification performed by an external machine learning model. Specifically, feature representations are extracted from the final convolutional layer and used to train an SVM, enabling faster experimentation and potentially enhance performance. The evaluation is conducted using per-channel and stacked spectrogram representations.

First, each per-channel spectrogram is passed through a pretrained CNN only once. The resulting feature vector which numerically represents the spectrogram image is saved externally for later use, which significantly reduces training time compared to end-to-end CNN training. The dimensionality of the extracted features depends on the architecture of the pretrained model, resulting in feature vectors of size 1280, 2048, 2048, and 1024 for MobileNetV2, ResNet50, InceptionV3, and DenseNet121, respectively.

Before training the SVM, the feature vectors are standardized by removing the mean and scaling to unit variance based on the training set statistics. This transformation ensures that the features are on a comparable scale to facilitate effective learning. The same transformation is then applied to the validation set to maintain consistency and prevent data leakage.

To determine suitable hyperparameters for the SVM classifier, a grid search was conducted on Subject 1 using the features extracted from each pretrained CNN. The goal was to identify the optimal regularization strength and kernel type. The best-performing configuration across all models employed a radial basis function kernel and a regularization parameter of C = 10, whereas C controls the trade-off between model complexity and training error (Pusarla et al., 2022). This parameter combination was subsequently applied to all subsequent subjects and models in this evaluation.

Table 4.3 summarizes the overall average validation accuracy and F1-score across subjects for each model architecture. The best performance was achieved using features extracted from InceptionV3 and DenseNet121, both attaining an average validation accuracy of 86,67% and an F1-score of 86,47% and 86,26%, respectively. MobileNetV2 and ResNet50 also performed competitively, achieving average accuracies of 82,22% and 81,11%.

Table 4.3: Performance across models for per-channel spectrograms using feature extraction

	MobileNetV2	**ResNet50**	**InceptionV3**	**DenseNet121**
Accuracy Mean	82,22	81,11	**86,67**	**86,67**
F1 Mean	81,65	80,72	**86,47**	86,26

The results suggest that deeper architectures such as InceptionV3 and DenseNet121 produce more discriminative features for emotion recognition when used in conjunction with an external SVM.

Analyzing the confusion matrices provides deeper insight into the classification behavior of each model, revealing both strengths and recurring error patterns. As shown in Figure 4.12, and consistent with previous findings, positive emotions remain the most reliably predicted category across all CNN architectures, with class-wise accuracies ranging from 90% for MobileNetV2 to 97% for DenseNet121. For neutral emotions, average classification accuracy ranges from 77% to 83%, while negative emotion predictions vary between 73% and 83%.

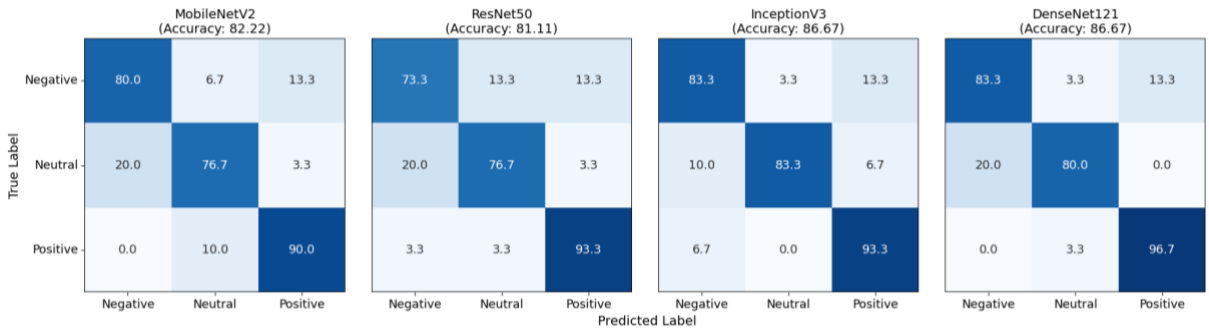

Figure 4.12: Confusion matrices across pretrained models for per-channel spectrograms using feature extraction

Although the extracted features used with the SVM lead to slightly more stable predictions, the overall structure of the confusion matrices remains similar. The most frequent misclassification involves neutral emotions being predicted as negative, a trend observed across all models. Additionally, 13% of negative emotions are misclassified as positive, highlighting a secondary but notable error pattern.

Subject-level confusion matrices in Figure 4.13 reveal that this particular misclassification is primarily driven by Subject 14, for whom 50% of negative trials are incorrectly labeled as positive.

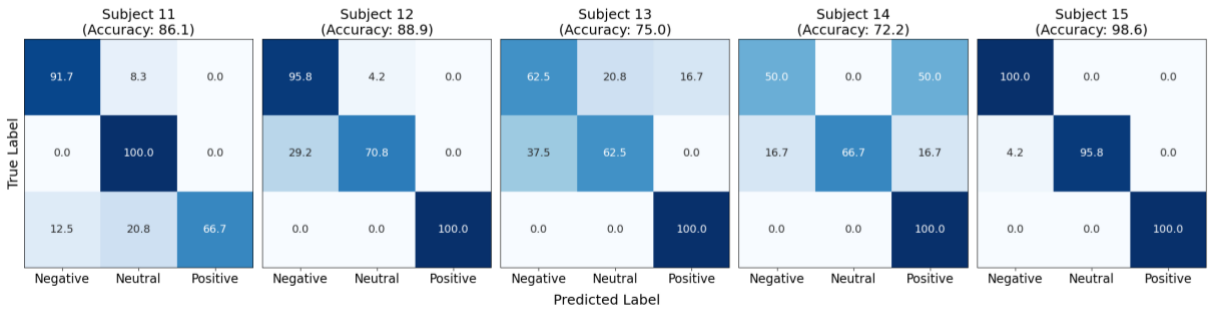

Figure 4.13: Confusion matrices across subjects for per-channel spectrograms using feature extraction

This error does not appear in the same form across other subjects. For instance, Subject 13, in line with broader trends, shows the most frequent confusion between negative and neutral emotions. These results underscore the individuality of EEG-based emotional responses and the importance of accounting for inter-subject variability in model evaluation.

Lastly, stacked spectrogram representations are used for feature extraction. To account for the informative difference in representations a separate hyperparameter search was conducted for each model using the same configurations as before. The optimal parameters varied across architectures and are summarized in Table A.1 in Appendix A.

Table 4.4 shows the average validation accuracy and F1-score for each model. The best performance was achieved using features extracted from ResNet50, with an accuracy of 81,11% and F1-score of 79,83%. InceptionV3 and DenseNet121 followed closely, while MobileNetV2 exhibited the lowest performance, with an average accuracy of 65,56% and F1-score of 61,92%.

Table 4.4: Performance across models for stacked spectrograms using feature extraction

	MobileNetV2	**ResNet50**	**InceptionV3**	**DenseNet121**
Accuracy Mean	65,56	**81,11**	77,78	75,56
F1 Mean	61,92	**79,83**	77,66	75,34

The performance gap between models widens when using stacked spectrograms. MobileNetV2 performs particularly poorly. Conversely, ResNet50 maintains relatively high performance, indicating greater robustness to input representation changes. These trends reinforce the notion that network depth and architectural complexity play a crucial role when training classifiers on compressed representations like stacked spectrograms.

As shown in Figure 4.14, predictions made under this experimental setting remain relatively unstable, particularly for neutral and negative emotions.

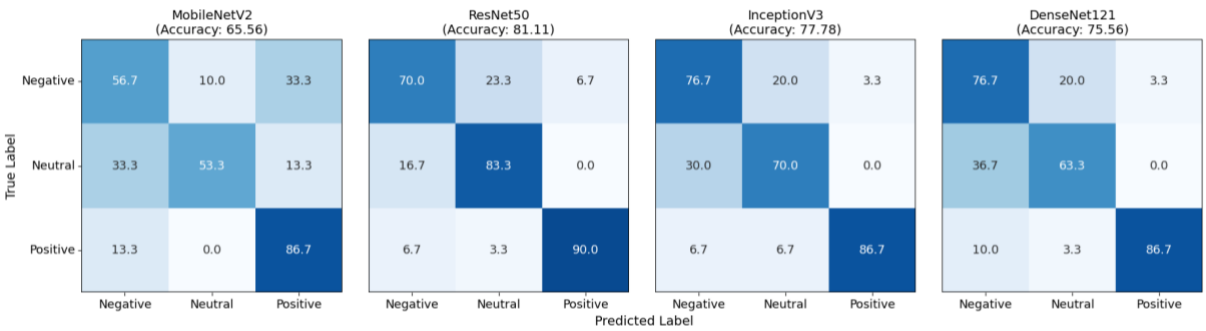

Figure 4.14: Confusion matrices across pretrained models for spectrogram stacks using feature extraction

As shown in Figure 4.15 (a), using stacked spectrograms for feature extraction and SVM classification results in consistently lower performance compared to per-channel spectrograms. Validation accuracies for per-channel representations range from 81,1% to 86,7%, while stacked spectrograms only range from 66,6% to 81,1%, depending on the underlying CNN. This performance decline is primarily attributed to the reduced dataset size when combining all EEG channels into a single stacked representation, as previously discussed.

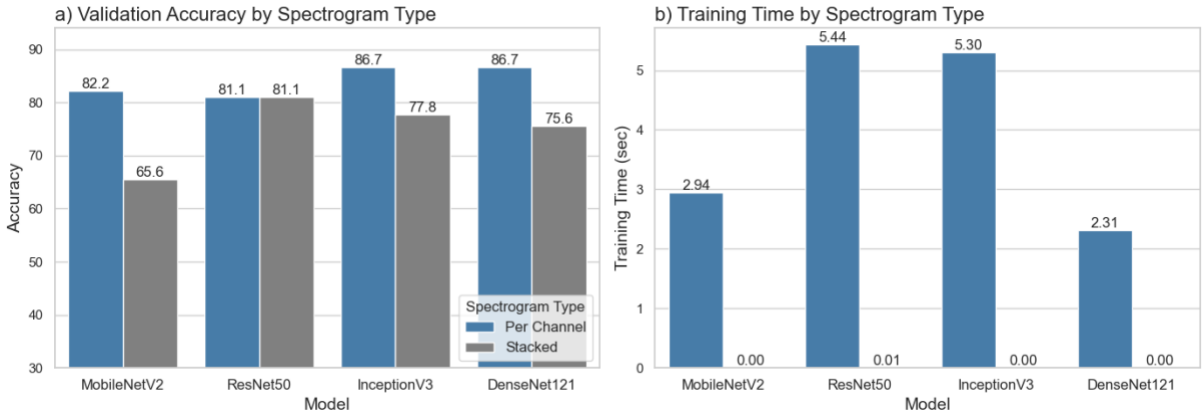

Figure 4.15: Performance comparison across spectrogram types and models using feature extraction

This limitation appears to affect the feature extraction strategy more significantly than transfer learning, suggesting its greater sensitivity to data scarcity. Features extracted from MobileNetV2 inhabit the most significant drop in performance reducing from 82,2% to 66,6% when trained with stacked spectrograms.

Future work may benefit from the integration of data augmentation techniques to address this limitation or using shorter segments in spectrogram stacks, as it increases the dataset size while still preserving sufficient information through the combined representation.

Although training efficiency is noticeably improved with stacked spectrograms, as shown in Figure 4.15 (b), with average training time per subject dropping to under 1 millisecond for all models compared to 2 to 5 seconds with per-channel inputs, this computational gain comes at a considerable cost in classification performance. Overall, the results underscore that in contexts where model accuracy is critical, per-channel representations remain the more reliable choice, even with slightly longer training times.

4.2.3. Transfer Learning vs. Feature Extraction

The following section presents a direct comparison between the two strategies for leveraging pretrained networks: transfer learning and feature extraction. As established in previous sections, single-channel spectrograms consistently outperformed stacked representations. Therefore, this comparison is restricted to results obtained with per-channel spectrograms.

Prior studies have reported mixed findings on the relative effectiveness of these methods, with Raghu et al. (2020) showing that feature extraction with an SVM outperformed transfer learning for seizure prediction, and Sadiq et al. (2022) finding the opposite for motor and mental imagery tasks.

As shown in Figure 4.16, the results of this study indicate that feature extraction using an additional SVM classifier generally outperforms transfer learning through fine-tuning. Validation accuracies for the feature extraction approach range from 81,11% to 86,67%, compared to 76,67% to 83,33% for transfer learning. The best overall performance was achieved using feature extraction with InceptionV3 and DenseNet121, both reaching a validation accuracy of 86,67%.

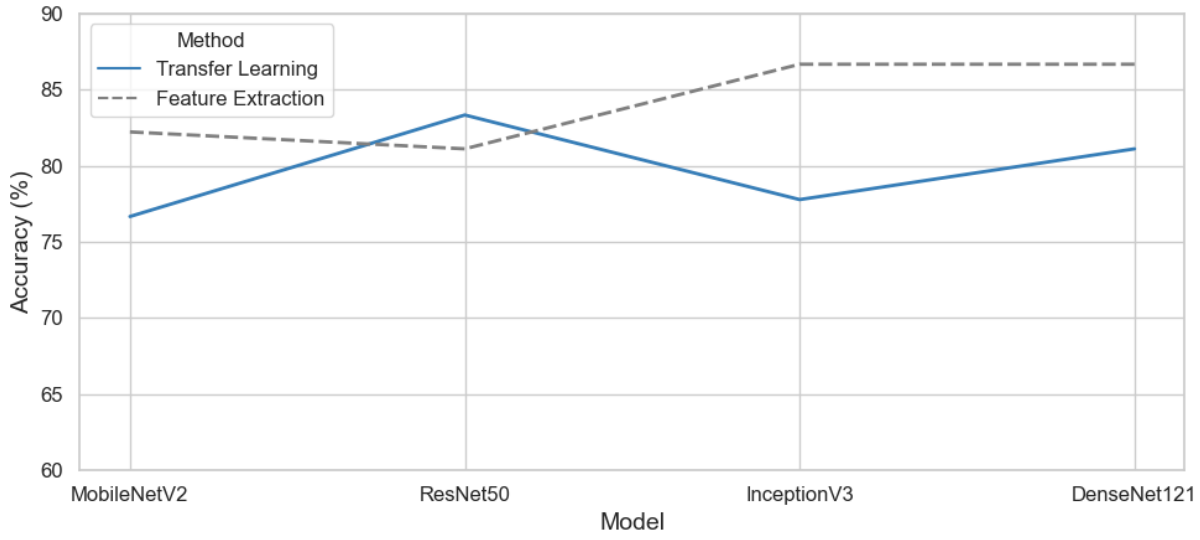

Figure 4.16: Comparison of transfer learning and feature extraction across models for validation accuracy

These findings support prior work by Raghu et al. (2020), who found InceptionV3 to perform best among ten pretrained models when used for feature extraction tasks, reinforcing its suitability for leveraging pretrained knowledge in EEG-based classification. Three of the four models (MobileNetV2, InceptionV3, and DenseNet121) demonstrate a notable improvement in performance for this method, with gains of 5%, 9%, and 6%, respectively.

In contrast, ResNet50 performs slightly better when fine-tuning the network. This may be due to architectural differences, or the fact that preprocessing and image generation parameters were initially optimized with this strategy using ResNet50, potentially biasing results in its favor.

In addition to inreased performance, feature extraction offers substantial advantages in training efficiency. As shown in Figure 4.17, training times for the transfer learning strategy range from 21 minutes for MobileNetV2 to 73 minutes for DenseNet121 per subject, primarily

due to the repeated forward and backward passes through the network during each epoch and input.

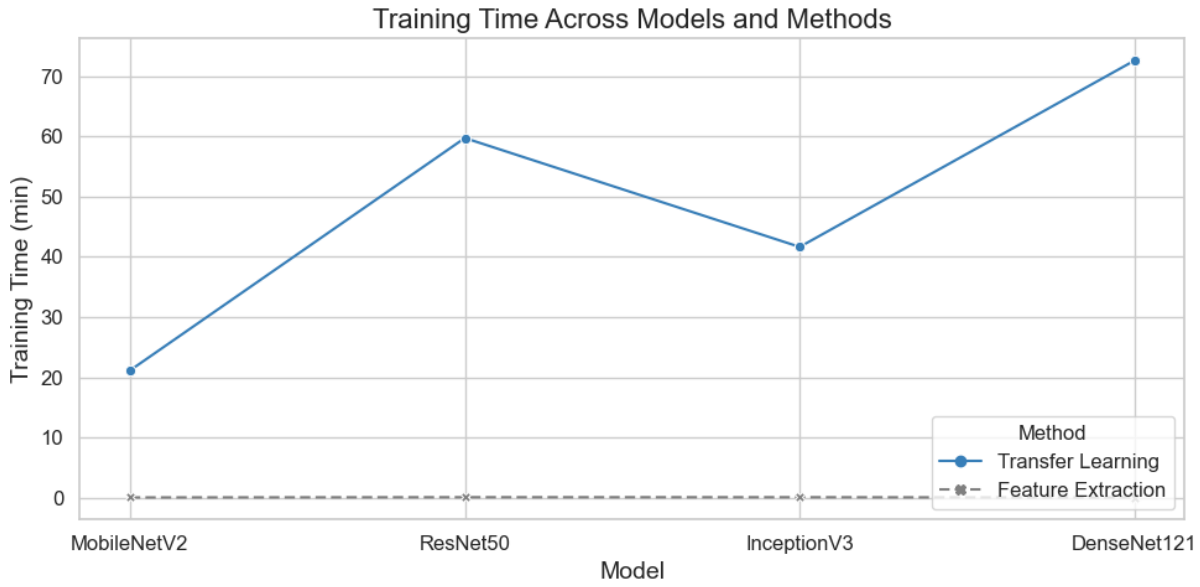

Figure 4.17: Comparison of transfer learning and feature extraction across models for training time

In contrast, feature extraction requires only a single forward pass per image to obtain features, which are then stored externally and reused during SVM training. Subsequently, the training process only takes 2 to 5 seconds per subject, depending on the dimensionality of the feature vectors.

To summarize, leveraging pretrained models as fixed feature extractors with an additional SVM yields the highest overall performance on the SEED dataset, while also requiring significantly less training time. This highlights the substantial advantage of this model adaptation strategy, particularly in the context of EEG-based emotion recognition.

4.2.4. Baseline and Literature Comparison

To contextualize the results, the proposed approach is also compared to a traditional EEG classification pipeline that does not involve transfer learning techniques. In this setup, a SVM was trained using precomputed PSD features provided within the SEED dataset. These features were extracted across five frequency bands (Delta, Theta, Alpha, Beta, and Gamma) and concatenated into a single feature vector per segment.

A grid search with the same hyperparameter settings used in earlier SVM evaluations was conducted to ensure a fair comparison. Similarly, majority voting was applied across segments to generate trial-level predictions, consistent with the methodology used for spectrogram-based models. The average validation accuracy and F1-score for the PSD-based classifier were 67,7% and 65,5%, respectively. These results are significantly lower than those achieved with

spectrogram-based CNN feature extraction, highlighting the advantage of deep feature representations in capturing complex EEG patterns for emotion recognition.

Table 4.5 provides a comparative summary of prior studies employing the SEED dataset with varying modeling approaches, feature representations and evaluation methods. Subject-dependent evaluation refers to the evaluation method applied in this study and is consistent with the approach used in the other referenced works.

Table 4.5: Comparison of EEG emotion recognition studies on the SEED dataset

Study	**Model**	**Features**	**Eval. Method**	**Accuracy (%)**
(W. L. Zheng & Lu, 2015)	DNN	DE	subject-dependent	86,05
(W. Liu et al., 2016)	Deep AutoEncoder	PSD, DE	subject-dependent	82,11
(J. Li et al., 2018)	HCNN	Topographical Map from DE Features	subject-dependent	88,2 (Gamma) 86,2 (Beta)
(Asghar et al., 2019)	AlexNet + SVM	Bag of Deep Features	subject-dependent	93,80
(F. Wang et al., 2020)	Proposed CNN	EFDM	subject-dependent	90,59
(Cimtay & Ekmekcioglu, 2020)	Inception-ResNet-V2	Raw	cross-subject	78,34
(Song et al., 2020)	DGCNN	DE, PSD, DASM, RASM, DCAU	cross-subject	79,95
(Sidharth et al., 2023)	Modified ResNet50	image matrix of MPC, MSC and DE featurs	cross-subject	71,60
(Zhu et al., 2024)	JD-IRT	Deep domain-adapted features	cross-subject	83,21
This Study	IncptionNetV2 + SVM	Spectrogram	subject-dependent	86,67

The best-performing configuration in this work through DenseNet121 and InceptionV3 achieve an average validation accuracy of 86.67%, outperforming earlier approaches such as the Deep AutoEncoder by Liu et al. (2016) with 82.11% and the DNN on DE features by Zheng and Lu (2015) with 86.05%. Although some methods report higher accuracies, such as the AlexNet + SVM model by Asghar et al. (2019) at 93.80% and the EFDM-based CNN by (F. Wang et al., 2020) at 90.59%, these rely on additional feature selection techniques or custom-designed CNN architectures.

Notably, several studies have employed EEG-based emotion recognition with a cross-subject evaluation protocol on the SEED dataset. Specifically, researchers apply the Leave-One-Subject-Out (LOSO) strategy, in which models are trained on data from multiple subjects and evaluated on data from a previously unseen individual (Sidharth et al., 2023). These studies underscore the increased complexity and generalization challenge inherent in cross-subject emotion recognition as they report consistently lower performance, with average classification accuracies ranging from 71.6% in the study by Sidharth et al. (2023) to 83,21% in the work of (Zhu et al., 2024).

4.3. Spatial and Temporal Performance Analysis

To gain deeper insight into predictions and model behavior, this section examines the model performance by analyzing predictions across individual trials, temporal segments, and EEG channels. This fine-grained evaluation reveals temporal and spatial trends in prediction behavior, offering a more nuanced understanding of how emotional responses manifest in EEG data. These observations complement the main findings and help identify factors that influence prediction quality, potentially guiding future improvements in EEG-based emotion recognition systems.

4.3.1. Performance by Trial and Segment

Each subject watched all 15 video trials in three separate sessions conducted over a span of several weeks (W. L. Zheng & Lu, 2015). In this context, each complete viewing of the 15 videos by a subject is referred to as an experiment and the video trials used for validation correspond to the following emotional labels: trials 10 and 14 are labeled as positive, trials 11 and 13 as neutral, and trials 12 and 15 as negative.

Figure 4.18 (a) illustrates the average classification accuracy across individual trials, aggregated over all subjects and models prior to majority voting.

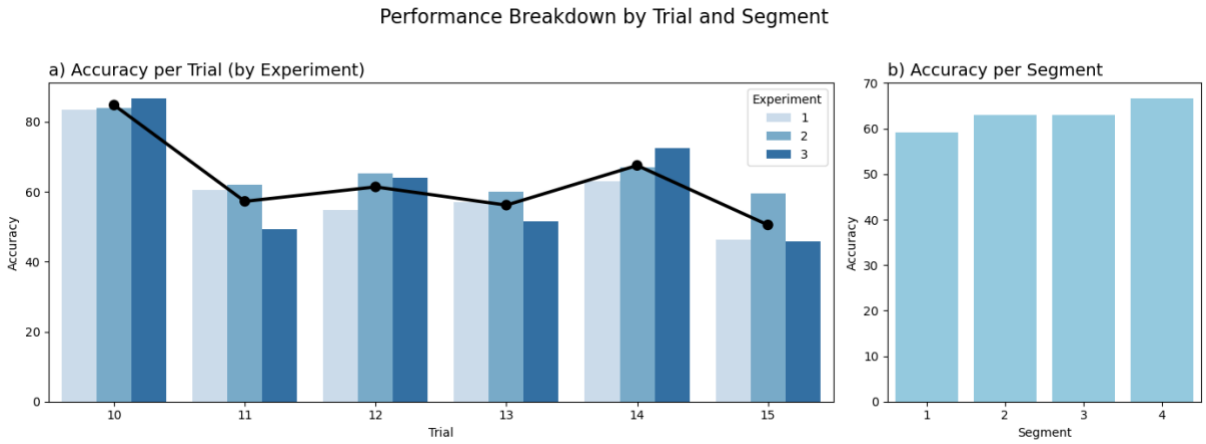

Figure 4.18: Validation accuracy by trial and segment

The results align with earlier findings, highlighting that positive emotions are more reliably classified, while neutral and negative emotions pose greater challenges. Trial 10 exhibits the highest overall validation accuracy, suggesting that its associated stimulus may have been especially effective in eliciting strong, distinguishable positive responses in the EEG signals. Conversely, trial 15, associated with a negative video, yields the lowest average accuracy, highlighting the difficulty of decoding negative emotional states from EEG data. Trials 11 to 13, which include both neutral and negative content, result in comparable mid-range performance.

A further temporal trend emerges when comparing trials over time. For positively labeled trials, classification accuracy tends to improve in later experimental sessions. This may indicate a cumulative emotional engagement effect, where subjects exhibit stronger or more consistent emotional responses to positive stimuli in repeated viewings. In contrast, a slight decline in performance on neutral trials over time may reflect habituation or a shift toward a more detached experience after the initial exposure.

Figure 4.18 (b) illustrates model performance across individual segments. The results show a consistent increase in performance across subsequent one-minute segments, suggesting that emotional intensity may build over time as subjects engage with the video. This trend leads to the highest classification accuracy in the final segment of each trial. Such a temporal build-up could be an important factor to consider in future EEG-based emotion recognition models, where segment-level weighting might enhance predictive performance. It may also inform the design of future datasets by emphasizing the importance of capturing emotional build-up over time.

4.3.2. Performance by Channel

Analyzing per-channel predictions prior to applying majority voting enables a more fine-grained evaluation of how individual EEG channels contribute to the overall classification performance. Figure 4.19 (a) shows the spatial distribution of the 12 selected channels, while Figure 4.19 (b) presents the average accuracy of all subjects across these channels.

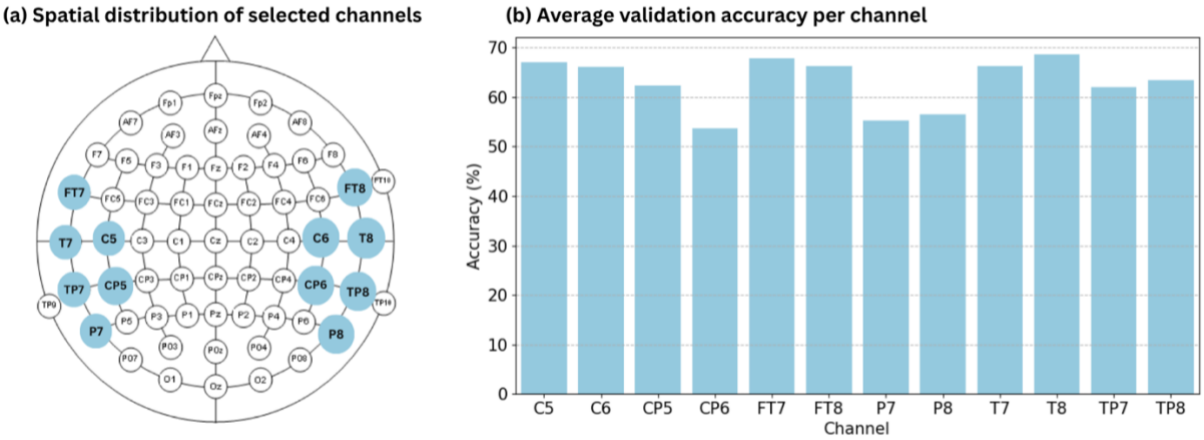

Figure 4.19: Spatial layout and validation accuracy of selected channels

The results reveal considerable variability in classification accuracy across channels, ranging from 52% for CP6 to 69% with FT7. This variability underscores the importance of combining channel-level predictions to achieve more stable and robust classification. While individual channel accuracies are relatively low and inconsistent, their aggregation yields a much higher validation accuracy as reported previously.

The spatial location of electrodes also appears to influence performance. For instance, P7 and P8, located at the posterior cortex, exhibit lower performance relative to other channels. In contrast, FT7 and FT8, situated in the frontal region, demonstrate significantly higher accuracy. This suggests that signals from the frontal lobe may carry more salient features for emotion recognition than those from the parietal lobe. This pattern is further supported by the performance of neighboring electrodes. T7 and T8 show superior performance compared to TP7 and TP8, which are positioned just below. Similarly, C5 and C6 considerably outperform their lower neighbors, CP5 and CP6.

These findings support the feasibility of performing emotion recognition using a reduced number of EEG channels. Furthermore, the overall relatively good performance while using only 12 of the 62 available channels suggests that future studies and practical applications can be designed around more compact consumer grade EEG devices, such as the Emotiv EPOC, which have shown promising results in related studies (Erat et al., 2024).

To explore the regional specificity of emotional processing, Figure 4.20 presents per-channel performance for a single subject, broken down by emotion label. This analysis reveals how different brain regions contribute unequally to the classification of emotional states. For this subject, electrodes located in the parietal and temporal lobes (P7, P8, T7, T8, TP7, TP8) perform particularly well on positive emotions, with T8 reaching 98%. However, these same regions perform poorly on negative emotions, with accuracies between 31% and 45%.

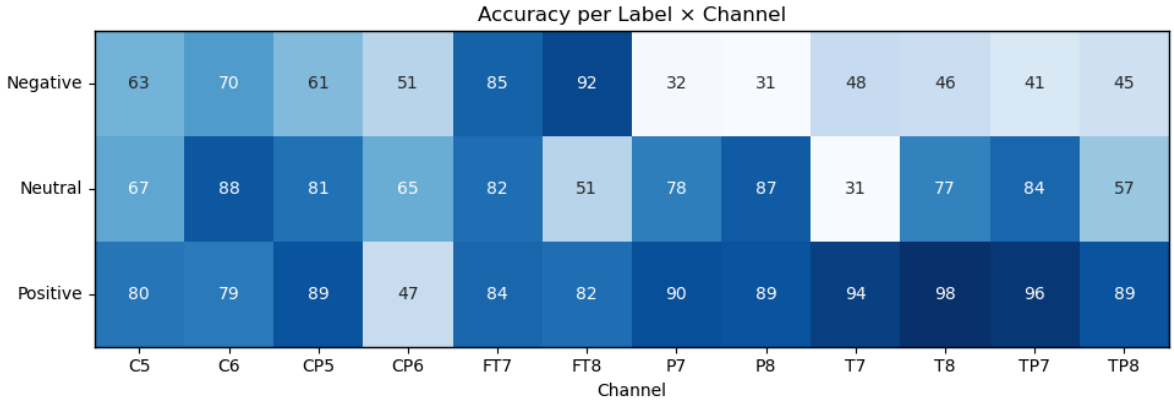

Figure 4.20: Channel-wise validation accuracy across emotional categories

In contrast, the frontotemporal electrodes (FT7, FT8) exhibit strong performance on negative emotions, each reaching accuracy above 85%, indicating their potential relevance for detecting negative emotional states. The central region (C5, C6, CP5, CP6) shows more balanced but generally moderate performance across all labels.

These findings reinforce the notion that emotions are differentially represented across cortical regions, and that the neural correlates of emotional states may vary not only by emotion type but also by anatomical location. This highlights the value of region-specific analysis in EEG-based emotion recognition. Identifying channels with consistently high performance such as FT7/FT8 for negative and T8/TP7 for positive emotions can inform future research, particularly in refining channel selection strategies for more efficient and effective emotion classification.

5. CONCLUSIONS

This study systematically investigated the use of pretrained CNNs to classify positive, neutral, and negative emotional states from EEG signals using the SEED dataset. It focused on evaluating the impact of preprocessing configurations, spectrogram-based feature representations, and two strategies for leveraging pretrained models across four CNN architectures: MobileNetV2, ResNet50, InceptionV3, and DenseNet121.

In relation to **RQ1**, the results demonstrated that both preprocessing and image generation configurations significantly influence classification performance, with accuracy increasing from 66.7% to 83.3% between the lowest- and highest-performing settings in the pre-evaluation. Among the preprocessing configurations, STFT-based spectrograms yielded the best results when applying a bandpass filter of 1–90 Hz, suggesting that high frequency components contain valuable emotional information. Notably, additional artifact removal techniques such as ATAR and wavelet-based denoising did not enhance performance and may have inadvertently suppressed informative signal components, particularly when working with spectrogram representations.

Image configuration choices were found to be equally impactful. Percentile-based color scaling, particularly the 0.1st to 99.9th percentiles, produced visually more discriminative spectrograms and led to improved model performance, while smoothing via interpolation had a negative effect. Finally, using 60-second segments provided the best temporal balance for emotion recognition, with performance further improved by applying a 25% overlap to increase data availability. Together, these choices enhanced signal quality and optimized the visual encoding of emotional EEG patterns.

Addressing **RQ2**, per-channel spectrograms ultimately yielded higher overall performance compared to stacked spectrograms. Although stacked spectrograms, created by horizontally concatenating per-channel spectrograms, preserve spatial dynamics across EEG channels and initially showed improved accuracy for negative and neutral emotions, the significant reduction in training data limited generalization capability. These findings suggest that while stacked spectrograms are promising in capturing spatial relationships, their effectiveness may depend on larger datasets and represent a valuable direction for future work.

In response to **RQ3**, using the pretrained CNN as a static feature extractor followed by SVM classification, outperformed transfer learning where the final layers of the network are adapted and fine-tuned in terms of validation accuracy, F1 score, and computation time. The highest average validation accuracy of 86,67%, was achieved by InceptionV3 and DenseNet121 using the feature extraction method, underscoring their strong capability to learn informative representations from EEG spectrograms. MobileNetV2, InceptionV3, and DenseNet121 showed better performance using this strategy, with respective gains of 5%, 9%, and 6%. Additionally, computation time decreased significantly, with DenseNet121 dropping from approximately 73 minutes to 5 seconds per subject.

Compared to a traditional SVM baseline trained on PSD features, all pretrained models achieved superior performance, highlighting the benefits of deep feature representations for EEG-based emotion recognition.

Beyond model comparisons, the study also confirmed that positive emotions were the most consistently and accurately classified across subjects and models. The majority of misclassifications involved confusion between negative and neutral emotions, with a notable tendency to misclassify neutral instances as negative. Performance varied significantly across subjects, with noticeable differences in the distribution of misclassification errors, highlighting the individual variability of emotional representation in EEG signals.

The study further found that the specific video stimuli used in the trials had a considerable impact on emotion induction, with validation accuracy differing notably between trials. Performance tended to improve in later video segments, possibly reflecting a temporal buildup of emotional intensity over time.

In addition, the location of EEG channels was shown to significantly influence classification performance. Channels situated in central and frontal regions of the cortex (C5, C6, T7, T8, FT7, FT8) consistently outperformed those located in posterior regions (CP5, CP6, TP7, TP8, P7, P8). These findings may inform future research, where selecting the most informative electrodes could help reduce hardware complexity and computational demands in real-world applications such as affective brain–computer interfaces.

In summary, the results underscore the importance of informed design choices in preprocessing, image generation, spectrogram representation, and pretrained model adaptation strategies for EEG-based emotion recognition. Collectively, they provide a robust methodological foundation for future work in the field.

6. LIMITATIONS AND FUTURE WORKS

One key limitation of this study is that the pre-evaluation of preprocessing and image generation parameters was performed using data from only a single subject, due to computational constraints. While the findings offer valuable guidance, it is likely that optimal configurations may vary across individuals. Evaluating preprocessing strategies across multiple subjects would allow for a more robust and generalizable selection of parameters.

It would also be interesting to evaluate stacked spectrograms with shorter segments, as the resulting increase in dataset size from this compact representation may lead to improved performance and better model generalization.

In this work, the evaluation was performed in a subject-dependent setting, with models trained and tested individually for each participant. While this approach reflects practical use cases for personalized models, it limits conclusions about cross-subject generalizability.

Another limitation of this study is the use of a single dataset, where emotions were elicited solely through film clips. Although the SEED dataset is widely used in research, passive video stimuli may lead to uneven emotional engagement. To develop models that generalize better to real-life conditions, future studies could benefit from open-environment EEG datasets that capture more naturalistic settings, including movement, noise, and varying attention levels (X. Li et al., 2022).

Future work could also explore CNN architectures pretrained on EEG-specific data, such as EEGNet, which is designed to process raw EEG signals directly (Lawhern et al., 2018). Although studies in domains like motor and mental imagery have shown that CNNs pretrained on image data often outperform EEGNet when using spectrogram representations (Sadiq et al., 2022), it would be interesting to investigate whether the same applies to emotion recognition.

Finally, incorporating multimodal physiological data represents a promising direction. Combining EEG with complementary biosignals such as ECG, EMG, or blood pressure could enhance emotion recognition and provide deeper insight into how emotional states are generated and regulated (H. Liu et al., 2024).

EEG-based emotion recognition offers promising applications in mental health and assistive technologies, but it also raises important ethical concerns. The sensitive nature of EEG data requires strict safeguards to protect privacy and prevent unintended inference of emotional states. Real-world deployment must ensure informed consent and transparency to avoid misuse, such as emotional profiling or surveillance. Responsible application of these technologies is essential to uphold ethical standards in neuroscience and AI.

BIBLIOGRAPHICAL REFERENCES

Abbas, Q., Baig, A. R., & Hussain, A. (2023). Classification of Post-COVID-19 Emotions with Residual-Based Separable Convolution Networks and EEG Signals. *Sustainability 2023, Vol. 15, Page 1293*, *15*(2), 1293. https://doi.org/10.3390/SU15021293

Abdulwahhab, A. H., Abdulaal, A. H., Thary Al-Ghrairi, A. H., Mohammed, A. A., & Valizadeh, M. (2024). Detection of epileptic seizure using EEG signals analysis based on deep learning techniques. *Chaos, Solitons & Fractals*, *181*, 114700. https://doi.org/10.1016/J.CHAOS.2024.114700

Aftanas, L. I., Reva, N. V., Savotina, L. N., & Makhnev, V. P. (2006). Neurophysiological correlates of induced discrete emotions in humans: an individually oriented analysis. *Neuroscience and Behavioral Physiology*, *36*(2), 119–130. https://doi.org/10.1007/S11055-005-0170-6

Alam, T. S., Jowthi, C. B., & Pathak, A. (2024). Comparing pre-trained models for efficient leaf disease detection: a study on custom CNN. *Journal of Electrical Systems and Information Technology 2024 11:1*, *11*(1), 1–26. https://doi.org/10.1186/S43067-024-00137-1

Arshed, M. A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M., & Shafi, M. (2023). Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models. *Information 2023, Vol. 14, Page 415*, *14*(7), 415. https://doi.org/10.3390/INFO14070415

Asghar, M. A., Khan, M. J., Fawad, Amin, Y., Rizwan, M., Rahman, M., Badnava, S., & Mirjavadi, S. S. (2019). EEG-Based Multi-Modal Emotion Recognition using Bag of Deep Features: An Optimal Feature Selection Approach. *Sensors 2019, Vol. 19, Page 5218*, *19*(23), 5218. https://doi.org/10.3390/S19235218

Aslan, M., Baykara, M., & Alakus, T. B. (2024). LieWaves: dataset for lie detection based on EEG signals and wavelets. *Medical and Biological Engineering and Computing*, *62*(5), 1571–1588. https://doi.org/10.1007/S11517-024-03021-2/TABLES/2

Bagherzadeh, S., Maghooli, K., Shalbaf, A., & Maghsoudi, A. (2023). Emotion Recognition Using Continuous Wavelet Transform and Ensemble of Convolutional Neural Networks through Transfer Learning from Electroencephalogram Signal. *Frontiers in Biomedical Technologies*, *10*(1), 47–56. https://doi.org/10.18502/fbt.v10i1.11512

Bagherzadeh, S., Shalbaf, A., Shoeibi, A., Jafari, M., Tan, R.-S., & Rajendra Acharya, U. (2024). Developing an EEG-Based Emotion Recognition Using Ensemble Deep Learning Methods and Fusion of Brain Effective Connectivity Maps. *IEEE Access*, *12*, 50949–50965. https://doi.org/10.1109/ACCESS.2024.3384303

Bajaj, N. (2021). Wavelets for EEG Analysis. *Wavelet Theory*. https://doi.org/10.5772/INTECHOPEN.94398

Bajaj, N., Requena Carrión, J., Bellotti, F., Berta, R., & De Gloria, A. (2020). Automatic and tunable algorithm for EEG artifact removal using wavelet decomposition with applications in predictive modeling during auditory tasks. *Biomedical Signal Processing and Control*, *55*, 101624. https://doi.org/10.1016/J.BSPC.2019.101624

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). *Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks*. http://arxiv.org/abs/1511.06448

Bisina, K. V., & Azeez, M. A. (2017). Optimized estimation of power spectral density. *Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017*, *2018-January*, 871–875. https://doi.org/10.1109/ICCONS.2017.8250588

Cecotti, H., & Graeser, A. (2008). Convolutional Neural Network with embedded Fourier transform for EEG classification. *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2008.4761638

Cimtay, Y., & Ekmekcioglu, E. (2020). Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition. *Sensors (Basel, Switzerland)*, *20*(7). https://doi.org/10.3390/S20072034

Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*(3), 425–455. https://doi.org/10.1093/BIOMET/81.3.425

Duan, R. N., Zhu, J. Y., & Lu, B. L. (2013). Differential entropy feature for EEG-based emotion classification. *International IEEE/EMBS Conference on Neural Engineering, NER*, 81–84. https://doi.org/10.1109/NER.2013.6695876

Ein Shoka, A. A., Dessouky, M. M., El-Sayed, A., & Hemdan, E. E. D. (2023). EEG seizure detection: concepts, techniques, challenges, and future trends. *Multimedia Tools and Applications*, *82*(27), 1. https://doi.org/10.1007/S11042-023-15052-2

Erat, K., Şahin, E. B., Doğan, F., Merdanoğlu, N., Akcakaya, A., & Durdu, P. O. (2024). Emotion recognition with EEG-based brain-computer interfaces: a systematic literature review. *Multimedia Tools and Applications*, *83*(33), 79647–79694. https://doi.org/10.1007/S11042-024-18259-Z/FIGURES/36

Gao, Z., Tian, Y., Lin, S.-C., & Lin, J. (2025). *A CT Image Classification Network Framework for Lung Tumors Based on Pre-trained MobileNetV2 Model and Transfer learning, And Its Application and Market Analysis in the Medical field*. https://arxiv.org/pdf/2501.04996

Goshvarpour, A., & Goshvarpour, A. (2019). EEG spectral powers and source localization in depressing, sad, and fun music videos focusing on gender differences. *Cognitive Neurodynamics*, *13*(2), 161–173. https://doi.org/10.1007/S11571-018-9516-Y/TABLES/3

Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology*, *14*(2), 233–238. https://doi.org/10.1016/j.conb.2004.03.010

Hazarika, N., Chen, J. Z., Tsoi, A. C., & Sergejew, A. (1997). Classification of EEG signals using the wavelet transform. *Signal Processing*, *59*(1), 61–72. https://doi.org/10.1016/S0165-1684(97)00038-8

Houssein, E. H., Hammad, A., & Ali, A. A. (2022). Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. *Neural Computing and Applications 2022 34:15*, *34*(15), 12527–12557. https://doi.org/10.1007/S00521-022-07292-4

Huang, H., Xie, Q., Pan, J., He, Y., Wen, Z., Yu, R., & Li, Y. (2021). An EEG-Based Brain Computer Interface for Emotion Recognition and Its Application in Patients with Disorder of Consciousness. *IEEE Transactions on Affective Computing*, *12*(4), 832–842. https://doi.org/10.1109/TAFFC.2019.2901456

Hussain, M., Bird, J. J., & Faria, D. R. (2019). A Study on CNN Transfer Learning for Image Classification. *Advances in Intelligent Systems and Computing*, *840*, 191–202. https://doi.org/10.1007/978-3-319-97982-3_16

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59

Katsigiannis, S., & Ramzan, N. (2018). DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics*, *22*(1), 98–107. https://doi.org/10.1109/JBHI.2017.2688239

Khosla, A., Khandnor, P., & Chand, T. (2020). A comparative analysis of signal processing and classification methods for different applications based on EEG signals. *Biocybernetics and Biomedical Engineering*, *40*(2), 649–690. https://doi.org/10.1016/J.BBE.2020.02.002

Kim, Y., Kwon, G., Kim, J., Hwang, D. U., Son, E. J., Oh, S. H., & Kim, W. (2022). Emotion recognition while applying cosmetic cream using deep learning from EEG data; cross-subject analysis. *PLOS ONE*, *17*(11), e0274203. https://doi.org/10.1371/JOURNAL.PONE.0274203

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis; Using physiological signals. *IEEE*

Transactions on Affective Computing, *3*(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

Krishnapriya, S., & Karuna, Y. (2023). Pre-trained deep learning models for brain MRI image classification. *Frontiers in Human Neuroscience*, *17*, 1150120. https://doi.org/10.3389/FNHUM.2023.1150120/BIBTEX

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/AACE8C

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, 253–256. https://doi.org/10.1109/ISCAS.2010.5537907

Li, J., Zhang, Z., & He, H. (2018). Hierarchical Convolutional Neural Networks for EEG-Based Emotion Recognition. *Cognitive Computation*, *10*(2), 368–380. https://doi.org/10.1007/S12559-017-9533-X/FIGURES/8

Li, W., Fan, L., Shao, S., & Song, A. (2024). Generalized Contrastive Partial Label Learning for Cross-Subject EEG-Based Emotion Recognition. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1–11. https://doi.org/10.1109/TIM.2024.3398103

Li, W., Huan, W., Hou, B., Tian, Y., Zhang, Z., & Song, A. (2022). Can Emotion Be Transferred? - A Review on Transfer Learning for EEG-Based Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(3), 833–846. https://doi.org/10.1109/TCDS.2021.3098842

Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., & Hu, B. (2017). Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, 352–359. https://doi.org/10.1109/BIBM.2016.7822545

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., & Hu, B. (2018). Exploring EEG features in cross-subject emotion recognition. *Frontiers in Neuroscience*, *12*(MAR), 294333. https://doi.org/10.3389/FNINS.2018.00162/BIBTEX

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., Zhao, Z., Kumar, N., & Marttinen, P. (2022). EEG based Emotion Recognition: A Tutorial and Review. *ACM Computing Surveys*, *55*(4). https://doi.org/10.1145/3524499

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Liu, H., Lou, T., Zhang, Y., Wu, Y., Xiao, Y., Jensen, C. S., & Zhang, D. (2024). EEG-Based Multimodal Emotion Recognition: A Machine Learning Perspective. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1–29. https://doi.org/10.1109/TIM.2024.3369130

Liu, W., Zheng, W.-L., & Lu, B.-L. (2016). *Multimodal Emotion Recognition Using Multimodal Deep Learning*. https://arxiv.org/pdf/1602.08225

Lu, Y., Yang, L., Worrell, G. A., Brinkmann, B., Nelson, C., & He, B. (2012). Dynamic imaging of seizure activity in pediatric epilepsy patients. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *123*(11), 2122. https://doi.org/10.1016/J.CLINPH.2012.04.021

Mandhouj, B., Cherni, M. A., & Sayadi, M. (2021). An automated classification of EEG signals based on spectrogram and CNN for epilepsy diagnosis. *Analog Integrated Circuits and Signal Processing*, *108*(1), 101–110. https://doi.org/10.1007/S10470-021-01805-2/TABLES/6

Martišius, I., & Damaševičius, R. (2016). A Prototype SSVEP Based Real Time BCI Gaming System. *Computational Intelligence and Neuroscience*, *2016*(1), 3861425. https://doi.org/10.1155/2016/3861425

Patil, A. U., Lin, C., Lee, S. H., Huang, H. W., Wu, S. C., Madathil, D., & Huang, C. M. (2023). Review of EEG-based neurofeedback as a therapeutic intervention to treat depression. *Psychiatry Research: Neuroimaging*, *329*, 111591. https://doi.org/10.1016/J.PSCYCHRESNS.2023.111591

Pusarla, A. N., Singh, B. A., & Tripathi, C. S. (2022). Learning DenseNet features from EEG based spectrograms for subject independent emotion recognition. *Biomedical Signal Processing and Control*, *74*, 103485. https://doi.org/10.1016/J.BSPC.2022.103485

Raghu, S., Sriraam, N., Temel, Y., Rao, S. V., & Kubben, P. L. (2020). EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Networks*, *124*, 202–212. https://doi.org/10.1016/J.NEUNET.2020.01.017

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 512–519. https://doi.org/10.1109/CVPRW.2014.131

Rguibi, Z., Hajami, A., Zitouni, D., Elqaraoui, A., & Bedraoui, A. (2022). CXAI: Explaining Convolutional Neural Networks for Medical Imaging Diagnostic. *Electronics 2022, Vol. 11, Page 1775*, *11*(11), 1775. https://doi.org/10.3390/ELECTRONICS11111775

Rozgic, V., Vazquez-Reina, A., Crystal, M., Srivastava, A., Tan, V., & Berka, C. (2014). Multi-modal prediction of PTSD and stress indicators. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 3636–3640. https://doi.org/10.1109/ICASSP.2014.6854279

Sadiq, M. T., Aziz, M. Z., Almogren, A., Yousaf, A., Siuly, S., & Rehman, A. U. (2022). Exploiting pretrained CNN models for the development of an EEG-based robust BCI framework. *Computers in Biology and Medicine*, *143*, 105242. https://doi.org/10.1016/J.COMPBIOMED.2022.105242

Samal, P., Mohammad, ·, Hashmi, F., Hashmi, M. F., Samal, P., & Hashmi, M. F. (2024). Role of machine learning and deep learning techniques in EEG-based BCI emotion recognition system: a review. *Artificial Intelligence Review 2024 57:3*, *57*(3), 1–66. https://doi.org/10.1007/S10462-023-10690-2

Samek, W., Meinecke, F. C., & Muller, K. R. (2013). Transferring subspaces between subjects in brain - Computer interfacing. *IEEE Transactions on Biomedical Engineering*, *60*(8), 2289–2298. https://doi.org/10.1109/TBME.2013.2253608

Schmidt, L. A., & Trainor, L. J. (2001). Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. *Cognition & Emotion*, *15*(4), 487–500. https://doi.org/10.1080/02699930126048

Shen, M., Yang, F., Wen, P., Song, B., & Li, Y. (2024). A real-time epilepsy seizure detection approach based on EEG using short-time Fourier transform and Google-Net convolutional neural network. *Heliyon*, *10*(11), e31827. https://doi.org/10.1016/j.heliyon.2024.e31827

Sidharth, S., Samuel, A. A., Ranjana, H., Panachakel, J. T., & Parveen K, S. (2023). Emotion detection from EEG using transfer learning. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. https://doi.org/10.1109/EMBC40787.2023.10340389

SJTU BCI Lab. (2024, September). *SEED Dataset*. Shanghai Jiao Tong University. https://bcmi.sjtu.edu.cn/home/seed/

Song, T., Zheng, W., Song, P., & Cui, Z. (2020). EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, *11*(3), 532–541. https://doi.org/10.1109/TAFFC.2018.2817622

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the Inception Architecture for Computer Vision* (pp. 2818–2826).

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Thodoroff, P., Pineau, J., & Lim, A. (2016). *Learning Robust Features using Deep Learning for Automatic Seizure Detection*. https://arxiv.org/abs/1608.00220v1

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., & Coleman, S. (2020). Emotion recognition with convolutional neural network and EEG-based EFDMs. *Neuropsychologia*, *146*. https://doi.org/10.1016/j.neuropsychologia.2020.107506

Wang, X., Ren, Y., Luo, Z., He, W., Hong, J., & Huang, Y. (2023). Deep learning-based EEG emotion recognition: Current trends and future perspectives. *Frontiers in Psychology*, *14*, 1126994. https://doi.org/10.3389/FPSYG.2023.1126994/BIBTEX

Wang, Y., Huang, Z., McCane, B., & Neo, P. (2018). EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*. https://doi.org/10.1109/IJCNN.2018.8489715

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., Aggarwal, C. C., Pei, J., & Zhou, Y. (2024). *A Comprehensive Survey on Data Augmentation*. https://arxiv.org/pdf/2405.09591

Wang, Z., Wang, Y., Zhang, J., Hu, C., Yin, Z., & Song, Y. (2022). Spatial-Temporal Feature Fusion Neural Network for EEG-Based Emotion Recognition. *IEEE Transactions on Instrumentation and Measurement*, *71*. https://doi.org/10.1109/TIM.2022.3165280

Yap, H. Y., Choo, Y. H., Mohd Yusoh, Z. I., & Khoh, W. H. (2023). An evaluation of transfer learning models in EEG-based authentication. *Brain Informatics*, *10*(1), 1–20. https://doi.org/10.1186/S40708-023-00198-4/FIGURES/9

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8689 LNCS*(PART 1), 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, Y., Zhang, S., & Ji, X. (2018). EEG-based classification of emotions using empirical mode decomposition and autoregressive model. *Multimedia Tools and Applications*, *77*(20), 26697–26710. https://doi.org/10.1007/S11042-018-5885-9/TABLES/3

Zheng, W. L., & Lu, B. L. (2015). Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175. https://doi.org/10.1109/TAMD.2015.2431497

Zheng, W.-L., & Lu, B.-L. (2016). Personalizing EEG-Based Affective Models with Transfer Learning. *International Joint Conference on Artificial Intelligence*.

Zhong, P., Wang, D., & Miao, C. (2022). EEG-Based Emotion Recognition Using Regularized Graph Neural Networks. *IEEE Transactions on Affective Computing*, *13*(3), 1290–1301. https://doi.org/10.1109/TAFFC.2020.2994159

Zhu, L., Yu, F., Huang, A., Ying, N., & Zhang, J. (2024). Instance-representation transfer method based on joint distribution and deep adaptation for EEG emotion recognition. *Medical and Biological Engineering and Computing*, *62*(2), 479–493. https://doi.org/10.1007/S11517-023-02956-2/FIGURES/9

Zhuang, N., Zeng, Y., Yang, K., Zhang, C., Tong, L., & Yan, B. (2018). Investigating Patterns for Self-Induced Emotion Recognition from EEG Signals. *Sensors (Basel, Switzerland)*, *18*(3). https://doi.org/10.3390/S18030841

APPENDIX A

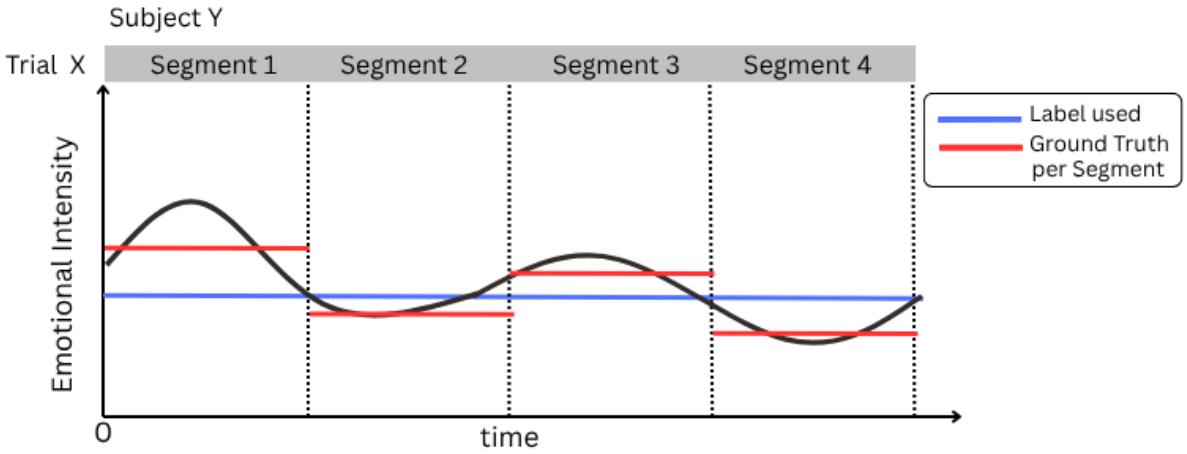

Figure A.1: Variability of emotional ground truth. Adapted from (Y. Wang et al., 2018)

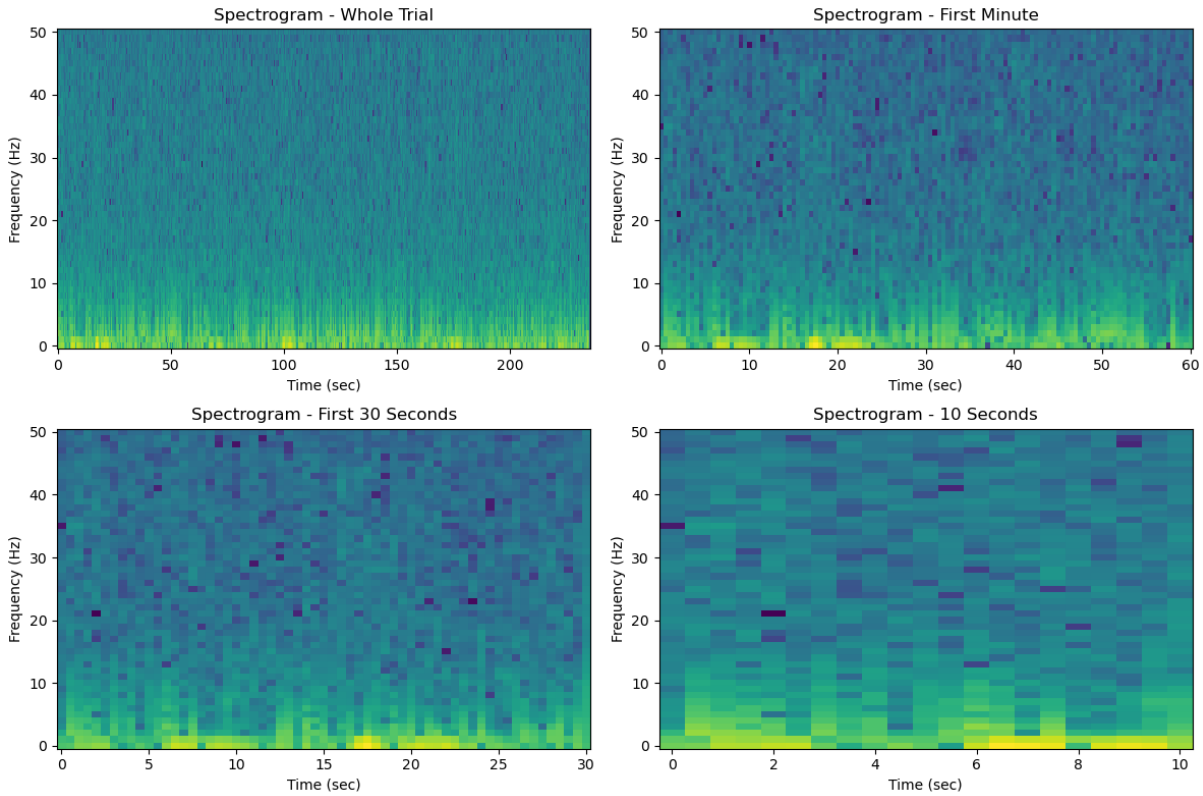

Figure A.2: Comparison of spectrograms over different time segments (entire trial, 1 minute, 30 seconds, and 10 seconds)

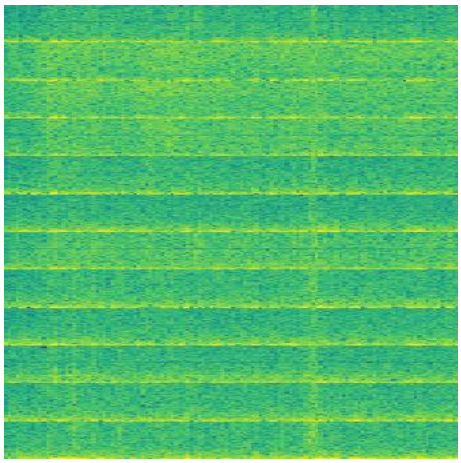

Figure A.3: Multi-channel stacked spectrogram composed of 12 EEG channels

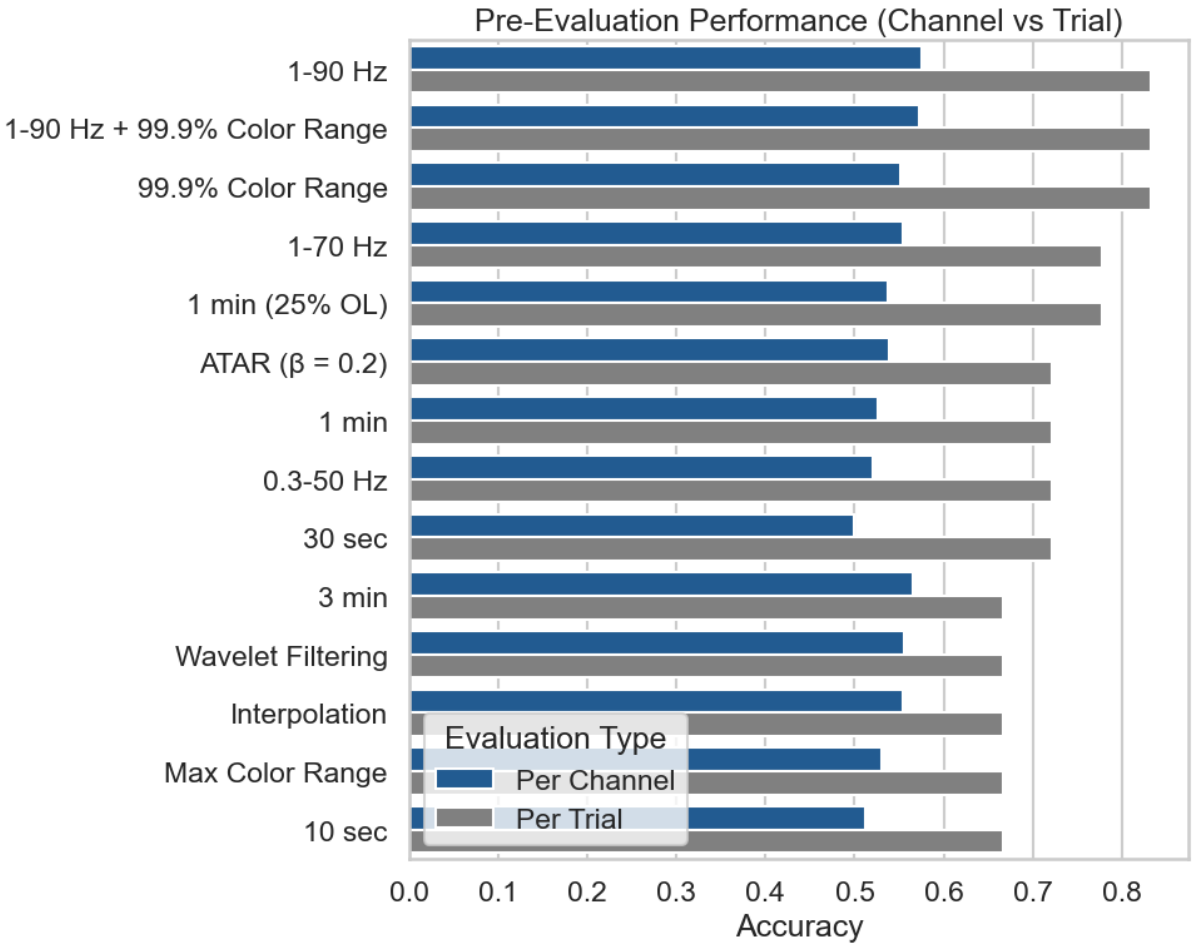

Figure A.4: Validation accuracy across all configurations during the pre-evaluation phase

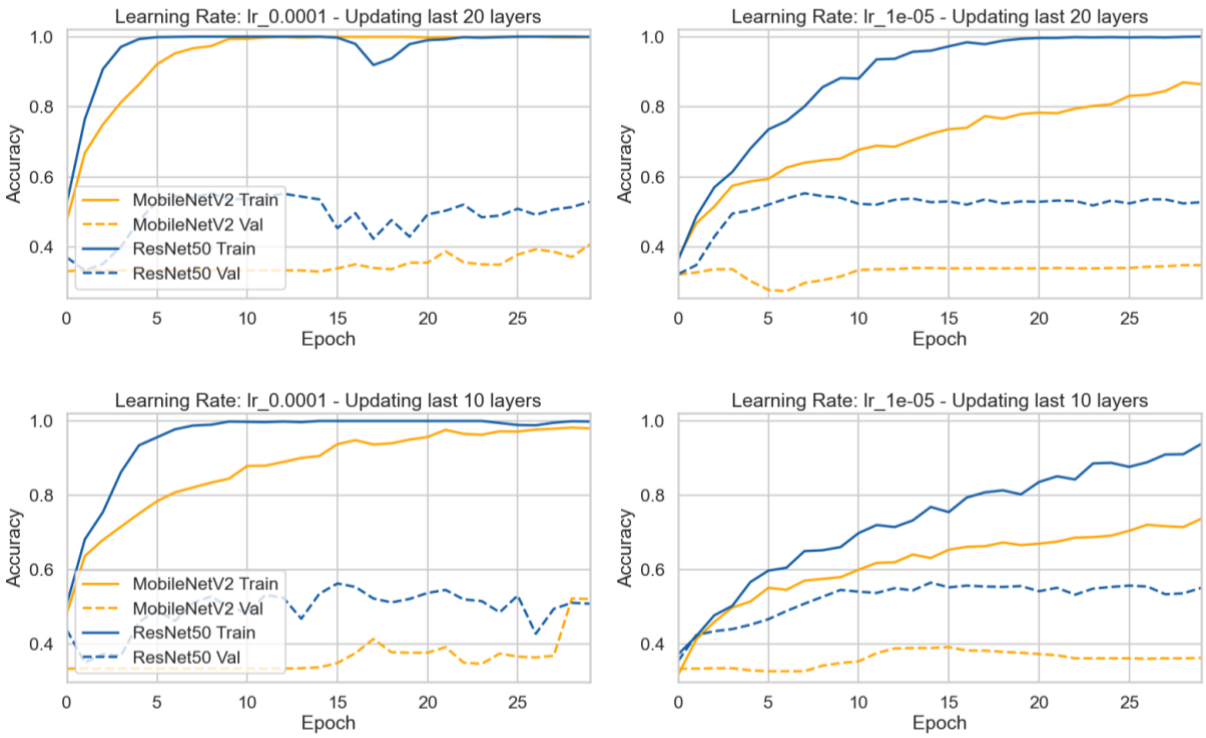

Figure A.5: Impact of trainable layers and learning rate on model convergence

Table A.1: Summary of Optimal Hyperparameters for Stacked Spectrogram Configurations with SVM Classification

Hyperparameter	**MobileNetV2**	**ResNet50**	**InceptionV3**	**DenseNet121**
C	10	0.001	0.001	1
Kernel	Radial-basis function	Linear	Linear	Linear

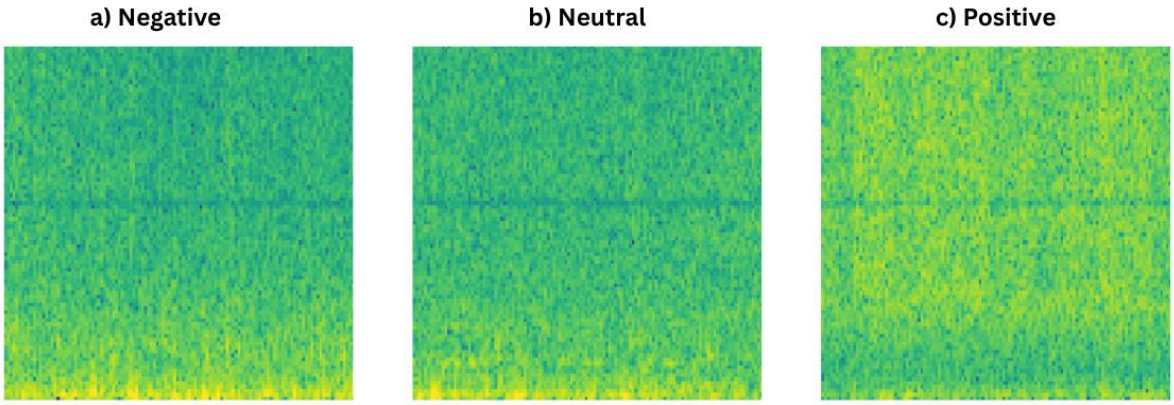

Figure A.6: Per-channel spectrogram examples for negative, neutral and positive emotions

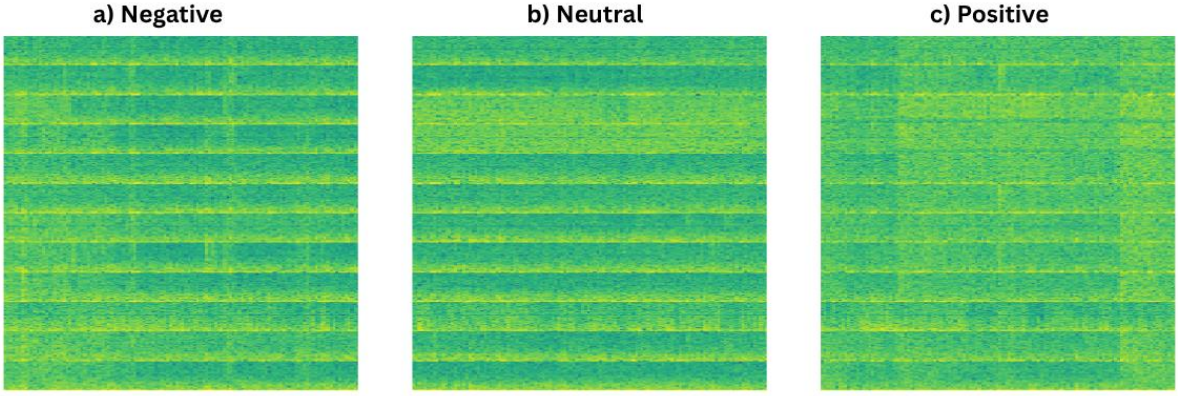

Figure A.7: Stacked spectrogram examples for negative, neutral and positive emotions

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa