Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Deep Automatic Understanding of Music Liking Through EEG

Author:
Jennifer Jiang

Supervisor:
Dr. Stefanos Zafeiriou

Submitted in partial fulfilment of the requirements for the MSc degree in MSc Computing Science of Imperial College London

September 2019

Abstract

Recent advances in biosensors technology and mobile electroencephalographic (EEG) interfaces have opened new application fields for cognitive monitoring. Recently, deep learning (DL) has also shown great promise in helping make sense of EEG signals. This work aims to assess the potential of deep learning methods for the automatic classification of subjective music preference using EEG signals. EEG data from 22 subject during a passive music listening task was collected and used as training data for five different Convolutional Neural Network (CNN) architectures and two baseline models using traditional machine learning classifiers Support Vector Machines (SVMs) and *k*-Nearest Neighbour (*k*NN). The results show that the baseline model SVM achieved best accuracy at 45.13% and that deep learning approaches require bigger EEG datasets for the classification of subjective music liking.

Acknowledgments

I would like to give my thanks to:

- My supervisor, Stefanos Zafeiriou, and Panagiotis Tzirakis, for giving guidance and feedback on the project.
- Dr. Dimitrios Adamos, for overseeing the EEG experimental sessions and pre-processing the collected data.
- Kendeas Theofanous for being my partner in EEG experimental sessions.
- Finally, thanks to all who agreed to participate in the experiment.

Contents

Chapter 1

Introduction

A Brain-Computer Interface (BCI) is a system that translates activity patterns of the human brain into messages or commands to communicate with the outer world. Traditionally, BCI systems are used for neuroprosthetics and building communication channels for patients with disabling neurological disorders. However, recent advances in commercial biosensor technologies and deep learning methods have opened up innovations of BCIs outside of the clinical domain, giving rise to potential novel applications of non-invasive BCIs in everyday life.

Electroencephalography (EEG), the measure of electrical fields produced by the active brain, is the most commonly used non-invasive neuroimaging technique. In clinical settings, EEG finds its use in sleep stage monitoring and epilepsy diagnosis, often requiring clinician's expertise to operate. Low-cost wireless consumer EEG devices, such as the Emotiv or Neurosky headsets [32], have shown potential for applications in real-time cognitive and emotional monitoring. Among them, the convenient bio-assessment of music-liking level emerges as a technological achievement that would majorly enhance current music recommendation systems.

Previous research have explored using EEG signals to classify music liking preference. Two recent works are of particular relevance to this project. Hadjidimitriou et al. [39] used different feature extraction methods and traditional machine learning methods to classify liked or disliked music across different subjects. Adamos et al. [5] identified EEG-specific biomarkers for music appreciation and implemented a real-time music recommendation system on a per-subject personalised basis. Both serve as important proof-of-concept studies on interfacing commercial EEG devices with a music recommendation system based on subjective liking.

EEG signals are characterised as high-dimensional, non-stationary time series data with low signal-to-noise ratio. The classification of EEG signals is a challenging task, traditionally involving hand-crafted features, and requiring heavy investments in both expertise and time. Deep learning, in particular, Convolutional Neural Networks (CNNs), is emerging as a promising tool for EEG classification as its end-to-end nature bypasses the need for heavy feature engineering. In the last year alone, three comprehensive reviews have examined the use of deep learning in recent EEG

studies of different applications, including emotion recognition, cognitive monitoring, and BCIs; all have concluded that deep learning methods achieve state-of-the art results, and there is much potential in the field [91] [29] [126]. In the latest BCI Kaggle competition, two of the top three winning teams used deep learning approaches [110]. A recent high-impact Nature paper reported use of a deep learning architecture to successfully synthesise speech from neural decoding of spoken sentences [11]. Commercially, companies such as Neuralink [80] are trying to bring Artificial Intelligence(AI)-powered lightweight BCI devices.

In light of these developments, this project aims to assess the potential of deep learning methods for the automatic classification of subjective music preference using EEG signals. The rest of this report is structured as follows: Chapter 2 outlines the characteristics and acquisition of EEG signals and the neural basis of music liking, followed by a summary of deep learning approaches and a literature review of its recent use in EEG classification. Chapter 3 details the experimental protocols used to collect the EEG data used in this project, and the implementation of deep learning architectures. Finally, the results are reported and evaluated in Chapter 4 and Chapter 5.

Chapter 2

Background

2.1 EEG

2.1.1 Measuring Brain Activity with EEG

Electroencephalography (EEG), also known as brainwaves, is the most commonly used non-invasive neuroimaging technique for measuring brain activity. Specifically, EEG monitors voltage fluctuations produced by the summation of pyramidal neurons in the outer cortical layers of the brain (Figure 2.1). When neurons fire, neurotransmitters (e.g. dopamine) are released across the synapse which causes a voltage change across the cell membrane, generating a subtle electrical field called the post-synaptic potential, which lasts tens to hundreds of milliseconds and are often on the order of tens of μV when it reaches the scalp. This measured potential reflects neuronal activity and can be used to study a wide range of brain processes.

The amplitude of EEG signals are measured by electrodes placed on the scalp. The recording equipment is usually a cap or headset, and compared to other neuroimaging methods, is more portable, accessible, and low-cost. The electrode locations are highly correlated spatially, meaning the more electrodes, the more spatial information captured. These locations are standardised by the 10-20 international system or the intermediate 10% electrode locations, which divides the scalp into 10% and 20% intervals [73] (Figure 2.2). Existing EEG collection systems can have as many as 64 electrodes (BCI 2000 system) and as few as 1 electrode (Mindware headset).

EEG signals have excellent temporal resolution: ionic currents change rapidly and events occurring in milliseconds can be captured. On the other hand, EEG signals suffer from poor spatial resolution. This is due to the limited number of electrodes, and the fact that the electric fields generated by the brain are obstructed by tissues, such as the skull, between the source and the sensor.

EEG signals typically contain several non-overlapping frequency bands, which result from oscillations of locally synchronised neuronal activity. This is referred to as brain rhythms and are strongly correlated with distinct behavioural states [67].

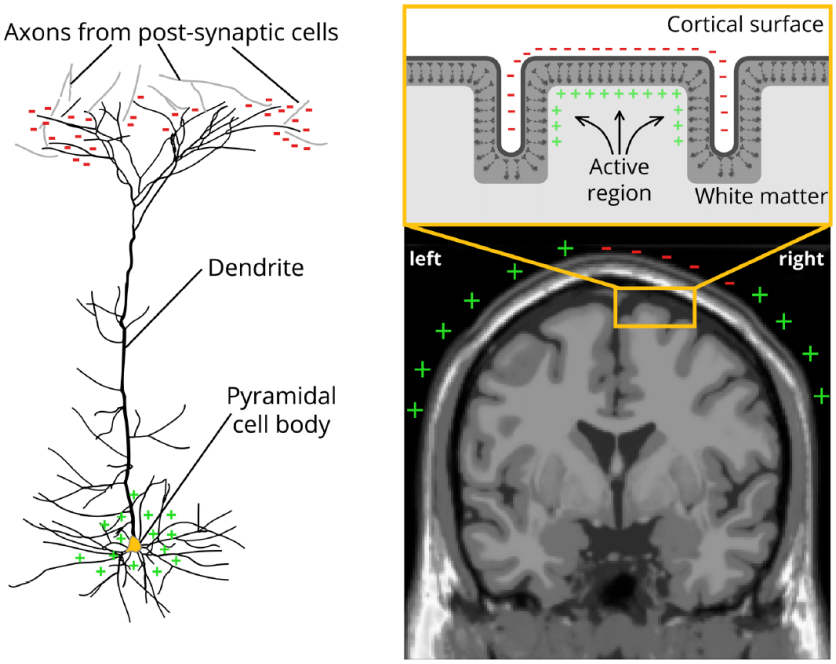

Figure 2.1: Left: Illustration of a pyramidal neuron with cell body and dendrites. Neurons typically consist of a cell body and dendrites that connect to each other through synapses. The neuron receives a signal from axons of another neuron across the synapse, generating a potential difference (red minus signs). Pyramidal neurons are large neurons present throughout the grey matter of the cortical layer and are always oriented perpendicular to the cortical surface. This unique characteristic allows the stable detection of their synchronised activity on the surface of the skull. **Right**: Illustration of perpendicular pyramidal cells in the grey matter (grey line above white matter) generating a summation of potential difference which can be measured on the surface of the skull. Images taken from [52]

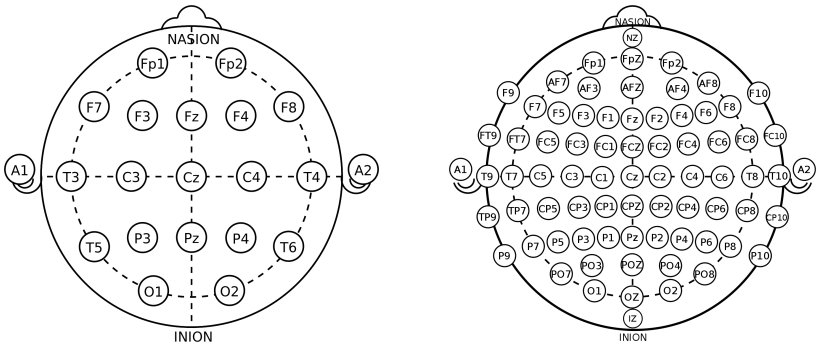

Figure 2.2: Standardised systems of EEG electrode locations. **Left**: International 10-20 System. **Right**: Intermediate 10% Electrode Positions (American Electroencephalography Society). Lettering on electrodes are abbreviations of brain areas electrodes are located over. For example, 'Fp'= pre-frontal, 'F' = frontal, 'C' = central, 'P' = parietal, 'T' = temporal, 'O' = occipital. Images Courtesy of Creative Commons License [27] [26].

Table 2.1 and Figure 2.3 show the visualisation of different rhythms, denoted Delta, Theta, Alpha, Beta, Gamma by increasing frequency ranges, and their associated brain states.

Rhythm	Frequency (Hz)	Amplitude	Brain State	Location
Delta	0.5-4	High	Deep sleep pattern	Frontal and posterior
Theta	4-8	High	Light sleep pattern	Entorhinal cortex, hippocampus
Alpha	8-12	Medium	Eyes closed, relaxed state	Posterior regions
Beta	12-30	Low	Active thinking	Frontal
Gamma	Above 30	Low	Cross-modal sensory processing	Somatosensory cortex

Table 2.1: EEG rhythm bands and corresponding characteristics.

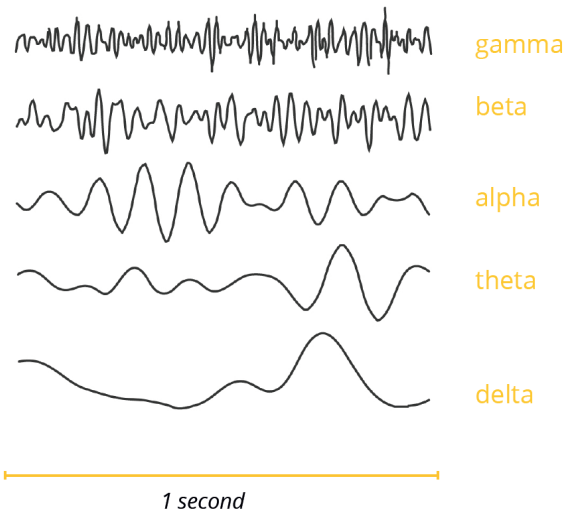

Figure 2.3: Example visualisation of different brain rhythm frequencies. Figure adapted from [52].

There are many applications for EEG. In clinical settings, EEG is used most often to

study sleep patterns [1] and epilepsy [4]. Conditions linked to changes in electrical brain activity, such as attention deficit hyperactivity disorder (ADHD) [14], disorders of consciousness [34], and depth of anaesthesia [40], are also monitored to various extents using EEG. In neuroscience and psychology research, EEG is widely used as a tool for understanding the brain and its underlying functions. Finally, EEG is widely used in Brain-Computer Interfaces (BCI), which allow communication channels to bypass the neural pathways of the brain so that brain activity can be directly translated into directives that affect the user's environment [88].

Various types of EEG signals are utilised in BCI applications. For example, P300 is an Event-Related Potential (ERP) that can be observed as a positive deflection in waveform around 300ms after the presentation of a novel visual stimuli. It is a well-characterised signal used often in BCI spellers for disabled patients [33]. Motor Imagery (MI) is a type of spontaneous EEG based on oscillatory somatosensory rhythms (SMR), requiring the user to imagine the execution of specific physical movements without external stimuli. MI is widely-used for BCI prosthetic control, and many public datasets are used for BCI decoding and Kaggle competitions [110].

In brief, we can divide EEG-BCI into active or passive paradigms: those that require explicit time-locked stimulation and those that do not. In general, passive BCI paradigms are challenging to train, due to the lower signal-to-noise ratio (SNR) and larger inter-subject variation [84](see next section). However, with the popularity of commercial EEG headsets, ease of use has become an important criteria for BCI usability. As a result, passive BCI paradigms have gathered increasing interest for healthy users in addition to disabled users, in applications such as device control, user state monitoring, evaluation, training and education, gaming and entertainment, cognitive improvement, safety, and security [13] [123] [114].

2.1.2 EEG Processing and Classification

On the most basic level, an EEG dataset consists of a 2D matrix of real values: channels and time. For example, a 32-channel EEG signal sampled at 256Hz over a 10 second time period would have the dimensions 32×2560. These discrete values represent brain-generated potentials recorded on the scalp associated with specific task conditions. This highly structured form makes EEG data suitable for machine learning. Consequently, a number of traditional machine learning and pattern recognition algorithms have been applied to EEG data. The traditional pipeline for EEG processing typically involves preprocessing (band-pass filtering and spatial filters), feature extraction, feature selection, and classification (Figure 2.4).

At the preprocessing step, low-pass/high-pass filtering and band-pass filtering is first applied. Frequencies below 1Hz and above 50Hz are usually filtered out as they do not contain useful EEG information. A band-pass filter can then filter the signal into a frequency band of interest, for example the beta band. Spatial filters and artifact removal algorithms may then be applied. For example, Independent Component Analysis (ICA) is commonly used to remove artifacts, and Principal Component

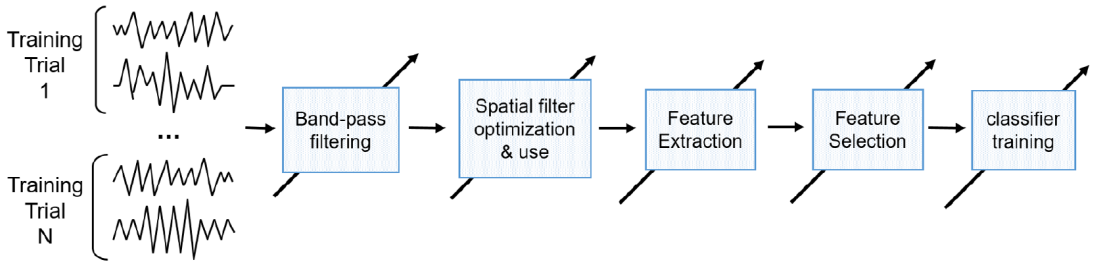

Figure 2.4: Overall process for EEG classification. Figure modified from [71]

Analysis (PCA) is commonly used reduce the dimensionality of the signals [72] [99].

Because of the high dimensional nature of EEG signals, feature extraction and selection aim to represent raw EEG signals by a smaller number of relevant values, which describe the task-relevant information contained in the signals. Three main sources of information can be extracted from EEG signals:

- **Spatial information:** describes where the information the relevant information comes from, focusing on specific brain regions. In practice, this would mean selecting specific channels of interest.
- **Spectral information:** describes how the power in a frequency band varies. In practice, this would mean calculating the power in some specific frequency band.
- **Temporal information:** describes how signal values vary over time. In practice, this would mean using EEG signal values at different time points and different time windows.

Passive EEG paradigms are based on oscillatory activity and therefore typically use spatial and spectral information, whereas active EEG paradigms are based on ERP and mostly use spatial and temporal information [58] [72] [71]. A feature vector is formed from the feature extraction stage which feeds into a classifier. Classic supervised learning methods such as Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and decision trees are common classifiers[72][71].

A few characteristics of EEG data makes it especially challenging to process. First, EEG has a low signal-to-noise ratio [21]. This can be due to both objective and subjective factors. Major objective factors include the obstruction of the skull and other tissues between cortex and scalp, environmental noise, and artifacts generated by activity-specific noise and stimulation. Various noise reduction techniques have therefore been developed to minimise the impact of these noise sources to better extract true brain activity from recorded signals. Subjective factors include the subjects mental stage and level of fatigue. These factors may impact affective and subjective tasks.

Second, EEG is a non-stationary signal [67] [35]. That is, its statistics vary across time. As a results, a classifier trained on limited amount of user data temporally

might generalise poorly to data recorded at a different time from the same individual. The is a great challenge for real-life applications of EEG classification, which often need to work on limited amounts of data.

Finally, EEG data have high inter-subject variability. This is due to the physiological differences between different individuals [88] [6] [8]. This can severely limit the ability of classifiers to generalise across subjects, which is important to many real-time applications of EEG classification.

2.2 Neural Substrate of Music Liking

The neural correlates of the subjective aesthetic appreciation of music has been studied by neuroscientists in the past two decades. Functional neuroimaging studies have shown that music strongly affect emotional states [24] [50]. Listening to pleasurable music can modulate changes to the subcortical structures of the brain involved in reward and emotions, including the ventral striatum, midbrain, amygdala, orbitofrontal cortex, and ventral medial prefrontal cortex [94][124]. The pleasant emotional response to music can also induce regional cerebral blood flow changes, giving rise to the phenomenon of ”chills” and ”shivers down the spine” when listening to liked music [22].

More recently, studies have begun to look at music perception and its induced emotions in EEG brain activity. The perception of music is reported to induce a significant increase of power in beta band over posterior brain regions when compared to resting condition [79]. An increase in gamma band was observed in trained musicians [20]. [98] reported that frontal brain EEG alpha power and asymmetric activation pattern is closely related to valence and intensity of music emotion [10], and [18] also reported changes in the alpha band modulated by emotional valence intensity. The contrast of pleasant and unpleasant music is also associated with an increase of frontal mid-line theta power [95].

Regarding the subjective evaluation of music, [38] and [83] found that higher frequency bands, in particular the gamma band over the forebrain area, were the most important for the decoding of music preference. Recently,[5] reported a biomarker for the assessment of spontaneous aesthetic brain responses during music-listening based on the concept of cross-frequency coupling (CFC) [55] and functional connectivity between different areas of the brain [119]. They reported that the beta and gamma oscillations recorded over the left prefrontal cortex, in particular the electrode location AF3, are most important for estimating the subjective aesthetic appreciation of a piece of music, and may reflect the inter-connectivity of the frontal cortex with subcortical music-rewarding dopaminergic areas.

To assess the underlying changes in electrical brain activity in response to preferred music, we followed the approach taken by Adamos et al. [5], which has the following characteristics:

- A *passive* listening paradigm is used [23]. That is, there is no task for the

subject, similar to how a real-life music recommendation system would be implemented.

- The *spontaneous* brain response of aesthetic appreciation to music is measured. This is to avoid the active cognitive processing during music listening, so that the signal measured reflects the pure response to music.
- Only the subjective liking/disliking is accounted for to avoid the limitations posed by emotional polarisation in traditional valence-arousal model related strategies [50] [8]. For example, listening to sad music can be classified as negative emotions, but this does not indicate the music is undesirable to the listener.

2.3 Deep Learning

Deep learning methods are a special case of representation learning, composed of multiple layers stacked on top of each other, where each layer learns representations from the previous layer. With recent advances, deep learning has become increasingly popular for use in feature learning of biosignals in BCI applications, such as EEG [74].

This section outlines three deep learning architectures commonly used for EEG classification tasks: Convolutional Neural Networks, Recurrent Neural Networks, and Deep Belief Networks.

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are capable of processing data with grid-like topology. A common example is images. Greyscale images are a single two-dimensional array while coloured images consist of three two-dimensional arrays. In CNNs each neuron applies the convolution to its input using an N two dimensional array, called kernel or filter bank. This operation produces an output which is referred to as feature map. For an input of a two-dimensional array, where I is the input, and K is the two-dimensional kernel, this can be represented mathematically as:

$$O(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(i,j) \tag{2.1}$$

A CNN is typically composed of successive convolutional (filtering) and pooling (subsampling) layers with a nonlinear activation functions applied before or after pooling, followed by one or more fully connected layers. Common nonlinear activation functions used are ReLU, which computes $f(x) = max(0, x)$; Leaky ReLU, which computes $f(x) = max(ax, x)$; and Maxout, which is a generalisation of ReLU

and leaky ReLU. The most popular pooling approach is max-pooling, which computes the maximum of non-overlapping kernels. Like regular neural networks, CNNs are trained using iterative optimisation with back-propagation algorithm, usually stochastic gradient descent (SGD). An example of a CNN architecture is shown in the Figure 2.5.

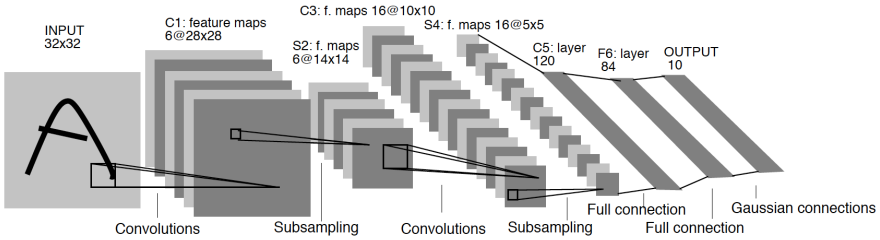

Figure 2.5: Example of CNN architecture used for document recognition [64]

2.3.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) take sequential information into account and processes sequence of values. In its simplest form, this recurrence is represented by:

$$\boldsymbol{s}_t = f_{\boldsymbol{W}}\left(\boldsymbol{s}_{t-1}, \boldsymbol{x}\right)$$

where $\boldsymbol{s}_t$ is the the state of the system at time t and represents the dynamics of the inputs x up until time t, $\boldsymbol{W}$ are the weight matrices to be learned and the function f describes the relationship between input x and the state $\boldsymbol{s}_{t-1}$. Figure 2.3.2 visualises how RNNs unfold in time, and the equation is below:

$$\begin{aligned} \boldsymbol{s}_t &= \tanh\left(\boldsymbol{s}_{t-1} * \boldsymbol{W} + \boldsymbol{x}_t * \boldsymbol{U}\right) \\ \boldsymbol{o}_t &= \boldsymbol{U}\boldsymbol{s}_t \end{aligned} \tag{2.2}$$

RNNs are trained using a variant of the back-propagation algorithm, known as backpropagation through time (BPTT). The most common RNN is the Long Short-Term Memory (LSTM) network [47]. An LSTM consists of four gates, the input gate i_t, the forget gate f_t, the output gate o_t and the cell gate s_t. The input gate decides the amount of information that is passed from the previous time step $t-1$ to the current time step t. The forget gates decides the amount of information that is ”deleted” from one step to another. An example of an LSTM cell and the equations for the cell is seen in Figure **??** and equation 2.3.

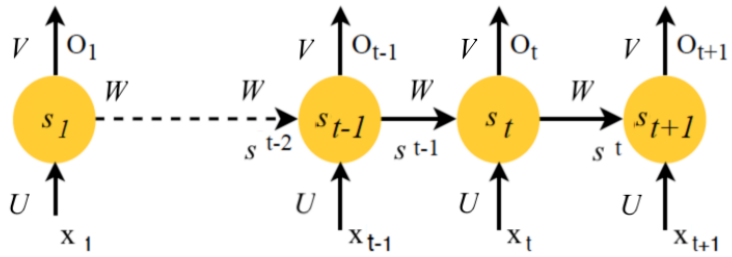

Figure 2.6: The recurrent procedure of the RNN model unfolding in time. For a specific node in time range $[1, t+1]$, the node at time t receives two inputs variables (x_{t} denotes input at time t and $s_{\text{t-1}}$ denotes the hidden state at time $t-1$) and exports two variables (output O_{t} and the hidden state s_{t} at time t). The input x_{t} is fed to the cell $s_{\text{t-1}}$ at time t and processed using the matrix U. The previous time steps are taken into account to produce the output O_{t} using matrix V. The parameters U, V, and W are the same for all time steps. Figure modified from [126].

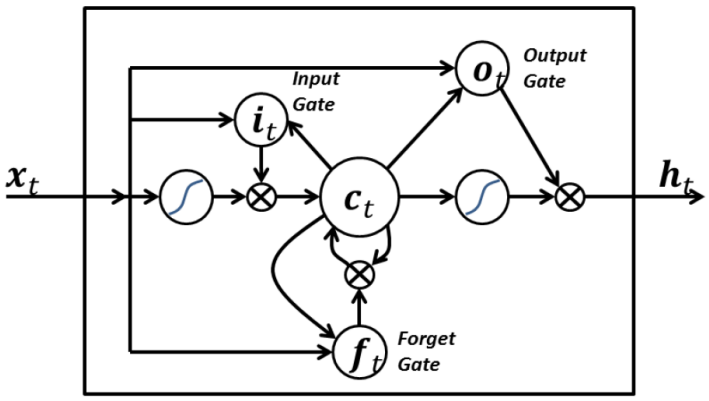

Figure 2.7: Example of an LSTM cell containing four gates: input, output, cell, and forget [47].

$$
\begin{aligned}
i_t &= \sigma\left(s_{t-1}W_{is} + x_t U_{xi} + b_i\right) \\
f_t &= \sigma\left(s_{t-1}W_{fs} + x_t U_{xf} + b_f\right) \\
s_t &= f_t s_{t-1} + i_t \tanh\left(U_{xs}x_t + W_{sh}h_{t-1} + b_s\right) \\
o_t &= \sigma\left(s_{t-1}W_{os} + x_t U_{xo} + b_o\right) \\
h_t &= o_t \tanh\left(s_t\right)
\end{aligned}
\tag{2.3}
$$

2.3.3 Deep Belief Networks

Deep Belief Networks (DBN) are neural networks consisting of a visible and hidden layer composed of stacked Restricted Boltzmann Machines (RBMs).

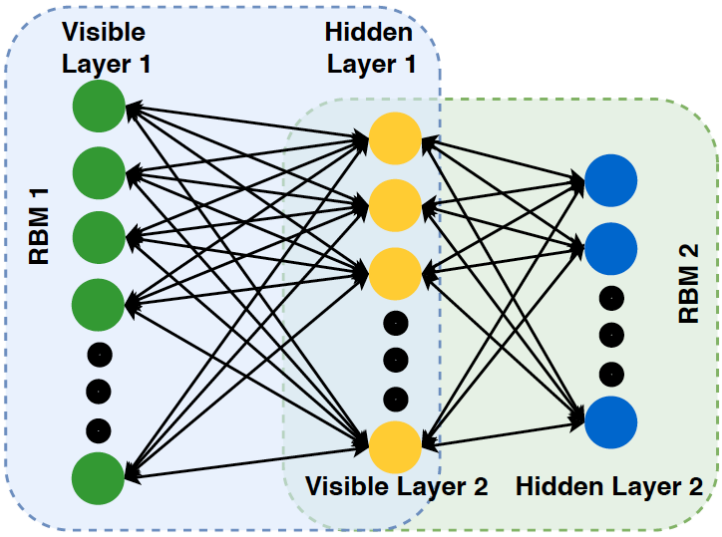

Figure 2.8: A 3-layer DBN composed of RBMs. In this illustration, there are two RBM components with the hidden layer of the first RBM working as the visible layer of the second RBM. The last hidden layer is the encoded representation [126].

DBNs are energy based models, which means that the probability distribution in each layer is defined through an energy function. A basic assumption in RBMs is that the neurons in the same layer are conditionally independent given the layer they are connected to. The joint probability distribution between the visible and hidden layer is defined as follows:

$$P(v,h) = \frac{1}{Z}\exp(-E(v,h))$$

where $Z = \sum_{v,h} e^{-E(v,h)}$ is a normalisation constant. Assuming a Gaussian RBM, the energy function is defined as follows:

$$E(v,h) = \frac{1}{2\sigma^2}\sum_i v_i^2 - \frac{1}{\sigma^2}\left(\sum_i c_i v_i + \sum_j b_j h_j + \sum_{i,j} v_i W_{i,j} h_j\right)$$

Where $c \in \mathbb{R}^D$ and $b \in \mathbb{R}^K$ are biases for the visible and hidden layers respectively, $W \in \mathbb{R}^{D\times K}$ are the weights between visible and hidden layer and σ is a hyperparameter. To learn the probability distribution of the input data, RBMs are usually trained according to a procedure called contrastive divergence learning. This learning procedure is based on a gradient ascent of the log-likelihood of the training data. DBNs are trained greedily, meaning the bottom layer is trained first and its hyperparameters are fixed before the next layer is trained.

2.4 Related Works

2.4.1 Traditional Machine Learning Approaches

The current state-of-the-art decoding algorithms include filter bank common spatial pattern (FBCSP)/ common spatial pattern (CSP) [56] [83], riemannian geometry classifier [28], discrete wavelet transform (DWT) [77][37] [116], as well as traditional machine learning algorithms Support Vector Machines (SVMs) [70] [87] [112], *k*-Nearest Neighbours (*k*NN) [38] [39], and Linear Discriminant Analysis (LDA) [71].

Pan et al. [83] used Common Frequency Pattern, a spectral-based version of CSP, for feature extraction and linear SVM for classification of binary music preference ("like" and "dislike") using 2-channel frontal EEG signal from 12 subjects. They achieved an average accuracy of 74.77% and through frequency band optimisation, found the gamma band to be most important for EEG music preference tasks. Similarly, Tseng et al.[112] classified music preference with an online device at 85.7% accuracy rate using SVMs and features extracted using Fast Fourier Transform (FFT).

Lin et al. [70] used the power spectrum asymmetry index of EEG signals from 32-channels as feature extraction and SVM as classifier to identify 4 classes of emotions from music listening. They reached 82.29% accuracy with 26 subjects with the best performance from alpha and beta bands. The same authors also used short-time Fourier transform to extract power spectral density (PSD) over time and compared two different schemes of SVM classifiers, multi-class SVM and hierarchical binary SVMs. The study found that hierarchical binary SVMs outperformed one-step multi-class SVMs by 10% [70].

However, simple Fourier transforms might not account for the non-stationary nature of EEG signals. EEG classification using both the time and frequency domains relies on the analysis of spectral power at specific time windows that span the entire duration of the measurement period. Hadjidimitriou et al. [39] used three different feature extraction methods in the time-frequency domain: spectrogram, Hilbert-Huang spectrum, and Zhao-Atlas-Marks transform. Using SVMs and *k*NN to classify liked or disliked music in 9 subjects, they found that *k*NNs performed the best with 86.5% accuracy. The same authors refined the same study to account for the degree of familiarity of the music listened to, and found that familiar and liked music had the highest classification accuracy, at $91.02 \pm 1.45\%$ with *k*NN [38].

Moon et al. [76] used short-time Fourier transform (STFT) and asymmetry scores in the time-frequency domain to classify preference of music videos in 4 classes. They tested SVMs, LDA, and *k*NN as classifiers and also found that *k*NN performed best, at 97.39%. To address inter-subject variability, [37] recently used flexible analytic wavelet transform, a variant of DWT, to extract channel-specific information in frequency sub-bands. Random forests and SVMs were used as classifiers on the public Dataset For Emotion Analysis using Physiological Signals (DEAP), achieving 79.99%.

Recently, Adamos et al.[5] identified EEG-specific biomarker for music liking in the high-beta and low-gamma bands of the left prefrontal cortex (AF3 electrode location) during passive listening. Hilbert transform, asymmetry index, and cross-frequency coupling were used as a composite feature extractor, and regression Extreme Learning Machines (ELM), a type of feedforward network, was used to predict music liking on a scale of 1 to 4 across 14 subjects. In an updated work they proposed an additional NeuroPicks system with non-linear dynamics VQ-scheme feature extraction method on a per-subject basis [57]. Both methods were on an online music recommendation BCI system. These serve as important proof-of-concept studies on interfacing commercial EEG devices with a music recommendation system based on subjective liking in real-time.

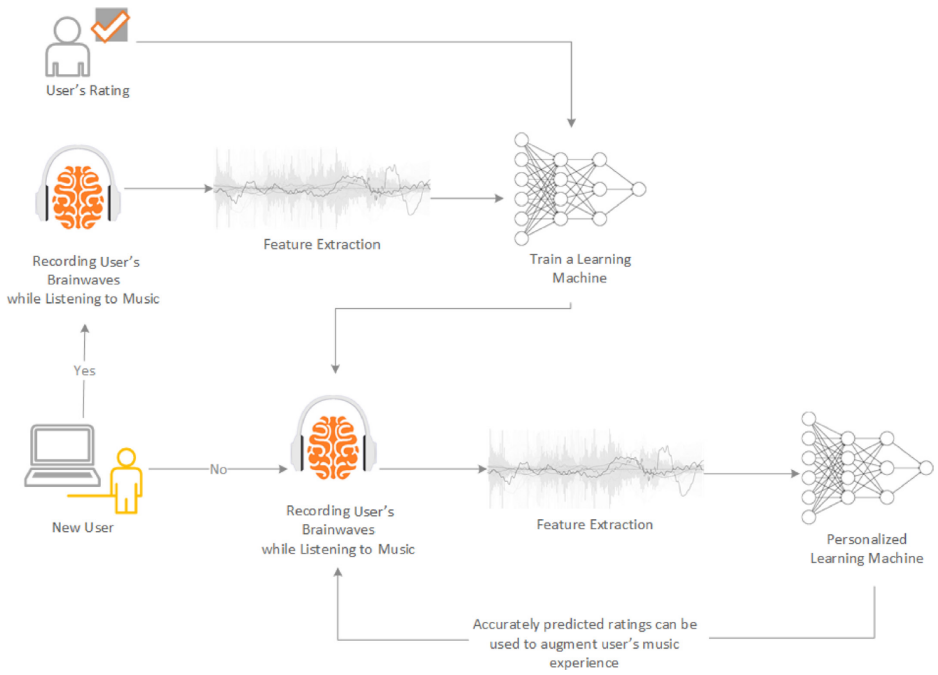

Figure 2.9: Processing flowchart of a real-time music recommendation system BCI from Adamos et al. [57].

Although FBCSP and riemannian geometry classifiers are the current state-of-the-art for non-DL EEG studies, these have yet to be implemented for the classification of emotions and preference in music listening.

2.4.2 Deep Learning Approaches

A major limitation in using deep learning for EEG classification tasks is that most EEG datasets are limited, making such data less adequate for training large-scale networks with millions of parameters. As it is often demonstrated, the advantages of deep neural networks over traditional machine-learning techniques become more apparent when the dataset size becomes very large [125]. Nonetheless, shallow deep learning models with reduced parameters have shown promise for EEG classification in the literature.

The most popular deep learning approach used by EEG classification tasks is CNN [91][29]. CNNs are a popular choice for EEG classification tasks as it is an end-to-end paradigm — a CNN combines feature extraction and classification in one step, bypassing the traditionally laborious EEG feature extraction process. In addition, many papers and reviews have reported that CNNs performed just as well, if not better, on raw data than on preprocessed data, increasing the appeal of using CNNs for real-time applications [97] [41] [78] [91]. As EEG signals are highly dynamic, non-linear time series data, RNNs, in particular LSTMs, also appear to be ideal for decoding the temporal characteristics and sequential nature of EEG data. However, recent works have also shown that CNNs are capable of learning hierarchical features in time series data [16] [48] [42].

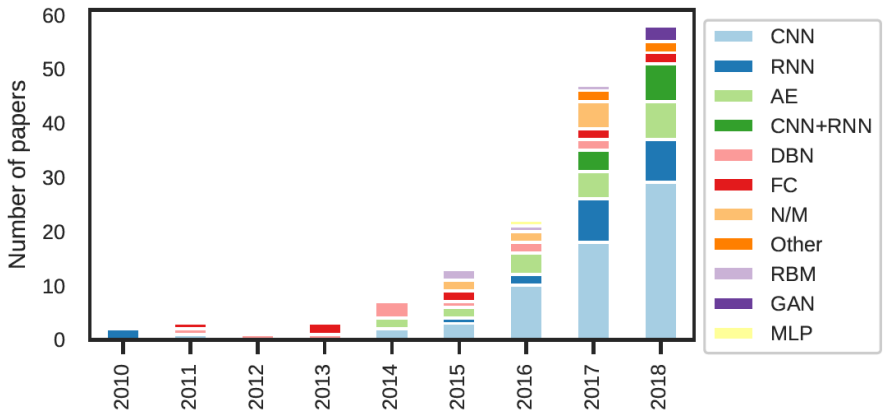

Figure 2.10: Number of EEG classification papers and the deep learning architectures employed, as systematically reviewed by Roy et al. [91]. AE = auto-encoder, FC = fully-connected, GAN = generative adversarial network, MLP = multi-layer perceptron.

As [29] reviewed in recent EEG studies (Figure 2.14), the input into CNNs and RNNs can be either raw or normalised time series signals values, or images derived from spectrograms or 2D/3D grids. CNNs have been used in several papers for EEG-based classification tasks, such as sleep-stage scoring [113][102], speech decoding [11], visualising dream imagery [48], emotion state estimation [121] [86] [111] [93] [68], cognitive load classification [17] [45] [54] [41] [9], biometric identification [115] [81], classifying and diagnosing neurological diseases [51] [3] [2] [12] [96] [82] [92], BCI applications such as P300 or motor imagery signals [63][31]

[97] [78] [109], and the classification of rhythm and tempo perception in music [107] [106] [104] [122]. RNNs have also been used for user-identification [108], emotion recognition [7][69], attention monitoring [43], cognitive load classification [61][45], sleep [62], and prediction of epileptic seizures [113] (For a more comprehensive summary of applications, see [91] [29]).

To the best of this author's knowledge, deep learning has not been used for the classification of liking preference in music listening. A few research groups, however, have worked on the classification of music-related events. Stober et al. used shallow 1 or 2 layer CNNs to classify the perception of 13 different types of rhythms, achieving mean accuracy of 24.4% [106]. In another study, they used convolutional auto-encoders (CAE) to classify audio events [107]. To address the limited EEG data, cross-trial encoding was used to attempt to capture invariance between trials and subjects. The group also established the *OpenMIIR* dataset [105], which was used to measure their model performances. Yu et al. [122] used a DenseNet as feature extractor for the classification of audio events in a hybrid model combined with audio data. For an 8-class problem their accuracy was 60% using EEG data only, and 81% when combined with audio data.

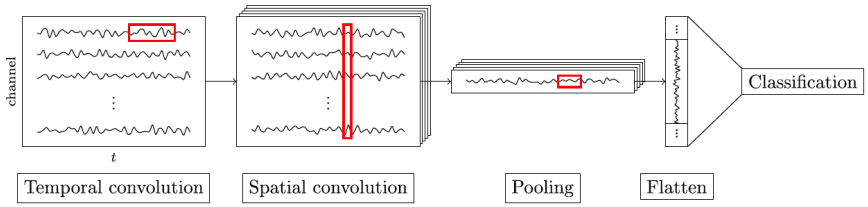

Figure 2.11: Illustration of general architecture from [31], inspired by [97]

Lawhern et al. [63] developed a CNN model, *EEGNet*, which classifies EEG signals for different BCI tasks such as P300, movement-related cortical potentials (MRCP), error-related negativity response (ERN), and sensorimotor rhythms (SMR). Their network used depthwise and separable convolutions, inspired by FBCSP, to learn spatial and temporal patterns in EEG time series. The performance was comparable to FBCSP, but the results were prone to inter-subject variability. Nonetheless, their results showed that a relatively simple CNN architecture was sufficient to generalise across different BCI tasks.

Schirrmeister et al. [97] investigated the design choices of CNN architectures for decoding motor imagery movements from raw EEG data. They presented two models, a *DeepConvNet* and a *ShallowConvNet* model. Using 1D temporal convolution in the first layer, and filters in the second layer that operate spatially in 2D across the learned temporal filters, batch normalisation and exponential linear units (ELU) activation, the authors obtained accuracies of 71.90% and 70.10% (4 classes) on BCI dataset IVa for DeepConvNet and ShallowConvNet, respectively. They found that the 5-layer DeepConvNet performed best on raw data and statistically better than FBCSP. The approach of using spatial and temporal filters was also used in [78].

Alhagry et al. [7] extracted temporal features using an RNN for classification on the 32-subject DEAP dataset. Their RNN consists of two LSTM layers, a dropout

layer, and a fully-connected (FC) layer, achieving 85.65% accuracy. Sun et al. [108] used 1D convolutional LSTM for user-identification with 4 layers of convolution followed by an FC layer, and 2 layers of LSTM followed by another FC layer. With a public dataset of 109 subjects the authors were able to obtain high accuracies using only 16 channel normalised EEG data. However, user-identification is a task that only needs to be applied on a per-subject basis, thus less challenging to train as it does not account for the high inter-subject variability of EEG signals.

Salama et al. [93] and Wei et al. [118] used 3D CNNs for emotion recognition on DEAP dataset and seizure detection, respectively. [93] directly transformed normalised time series data windows into 3D inputs into a 6-layer CNN, achieving 87.44% for 2 classes. [118] transformed raw EEG data into 2D images, then fused into 3D images according to the adjacent degree of electrodes. This was then fed into a 4-layer CNN, and a 90% accuracy was achieved for 3 classes. However, [118] had small number of subjects (13).

Aoe et al. [12] used raw resting-state 106-channel magnetoencephalography (MEG) signals, which measures magnetic fields of brain activity instead of electric fields, to classify 3 different neurological diseases using *MNet*. Using large kernels in the first convolution to learn global features in the temporal and spatial domain and FFT-computed power spectrum of the frequency bands as additional input into the FC layer, the authors achieved an accuracy of 70.7% $\pm$10.6.

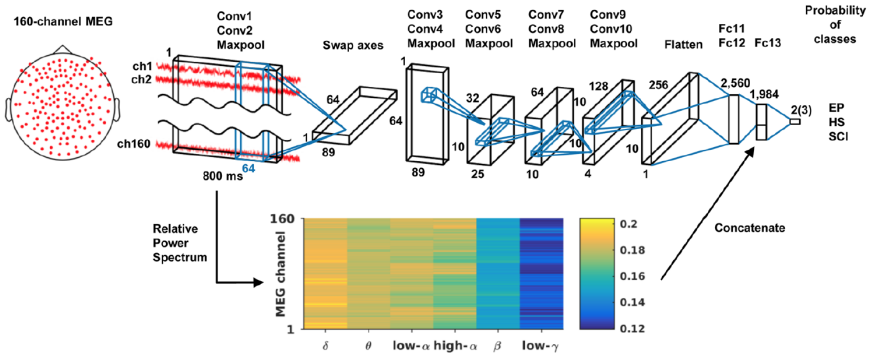

Figure 2.12: Illustration of architecture from [12]

Although using spectral representations such as mel-spectrograms as input into a CNN is commonly used for audio speech recognition tasks [36] [66] [127] [15], this is an interesting and relatively novel method for EEG signal classification tasks. Qiao et al. [86] used CNNs for emotion valence recognition on the DEAP dataset. On preprocessed EEG data, STFT was used to estimate the PSD of relevant frequency bands in manually selected critical channels. The resulting 3-channel coloured 2D images were then inputted into a two-layer convolution network, achieving an accuracy of 87.27%. [51] used 2D spectral representations obtained by computing windowed periodograms. The grey-scale 2-channel images were inputted into a 1 layer CNN with max-pooling to predict dementia stages in patients. [92] used stacked multichannel spectrograms generated from preprocessed EEG data as input into a 4-layer CNN or RNN, treating the task as an image or audio classification problem.

Bashivan et al. [17] transformed raw EEG activities into a sequence of topology-preserving multi-spectral images and used as input into a recurrent-CNN to classify cognitive load. FFT-computed spectrogram representations for relevant frequency bands were computed. 3D electrode locations over the scalp were then mapped into a 2D image using Azimuthal Equidistant Projection (AEP) to preserve relative distance between the electrodes. The images were stacked together to form an image with three colour channels and used as input into a CNN to learn the spatial representation of the data, then through an LSTM layer to learn the temporal representation of the data. A test error of 8.89% was achieved for the classification of 4 classes.

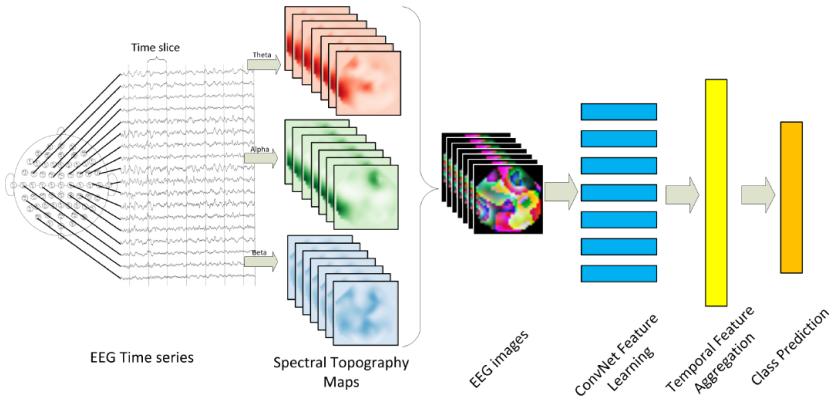

Figure 2.13: Illustration of approach from [17]

Similarly, Jiao et al. [54] transformed 64-channel raw EEG data into multi-spectral image series via topology preserving 2D or 3D projections using FFT and AEP for input into a 4-layer recurrent CNN network. They observed that for multi-channel input, the classification performance decreased when the electrode locations were randomly shuffled, indicating that spatial information between electrodes contain information that helps decoding. Kuanar et al. [61] used the same general architecture but with 22 subjects and 9-layers of convolution. They evaluated both LSTM and bidirectional-LSTM cells, achieving an average accuracy of 92.5% across the 4 classes.

However, it should be noted that the mental load experiments were ERP-based in the aforementioned studies [17] [54] [61], which is considered a easier classification task relative to passive BCI paradigms. Furthermore, the transformation of EEG channels into multiple topology preserving images could be computationally expensive to implement in real-time. [61] was also done on a per-subject basis, essentially avoiding the challenge of inter-subject variability.

DBNs are capable of achieving high accuracies on public BCI datasets, but the drawback is that the features often need to be calculated and transformed prior to input into the network [128] [29]. In particular, DBNs are said to be most effective for PSD features and frequency domain signals. The DBN network by Xu et al. [120]

uses a 3-layer DBN and takes PSD features as input has the highest classification accuray of 89% on the DEAP dataset. Plis et al. [85] showed that adding several RBM layers to a DBN with supervised pre-training results in networks can achieve considerable accuracy increases compared to other classifiers. Ren and Wu [89] used a convolutional DBN on EEG data processed with PCA and Fourier-transformed signals. They achieved an accuracy of 87.33% on public BCI datasets. Hajinoroozi et al. [41] and Jiao et al.[54] both used hybrid architectures made of channel-wise CNNs and RBMs.

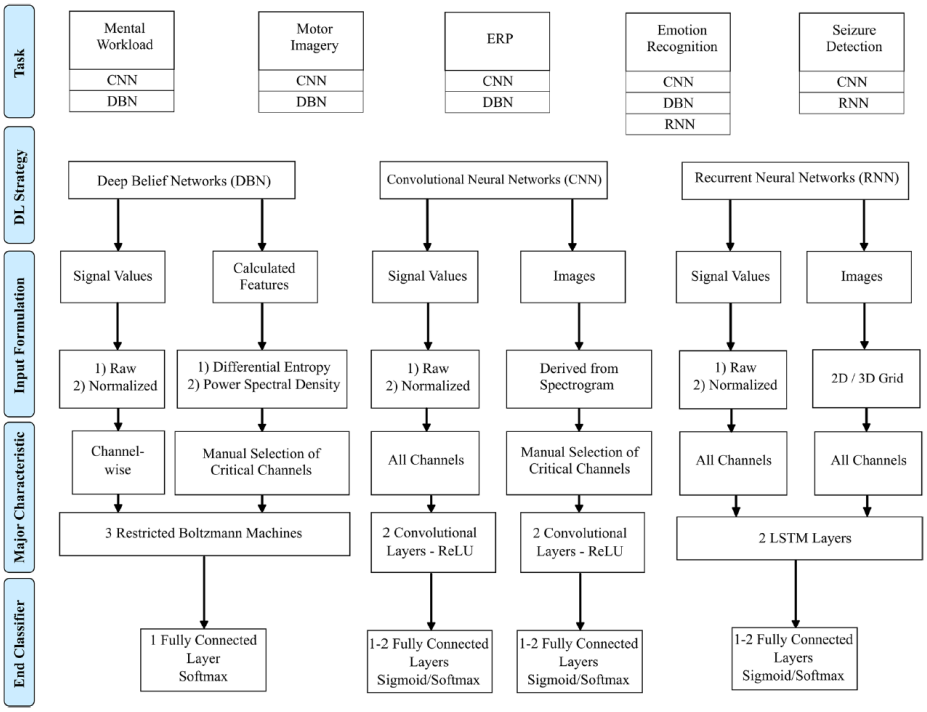

Figure 2.14: Recommendation of deep learning architectures for different EEG classification tasks from [29]

Chapter 3

Methods

3.1 Experimental Data

3.1.1 Participants

A total of 22 healthy participants (5 females, 17 males; average age 28.8) were recruited from Imperial College London across the duration of this project. Participants were master's students, PhD students, or researchers who listened to music on a daily basis. Prior to the experiment session, all participants were asked to register and fill in demographic information on the myBrainTunes project website hosted on the Imperial Department of Computing intranet (https://mybraintunes.doc.ic.ac.uk), and to read through ethics declaration and sign an informed consent form. They were also asked if they had consumed caffeine prior to the experiment, as this can sometimes affect EEG activity [30].

3.1.2 Data Acquisition

Prior to the experiment session, participants were asked to listen to excerpts from 60 songs and assigned a numerical rating from 0 to 5 based on their preference (0: do not like at all; 1: do not like; 2: undecided; 3: like 4: like very much; 5: one of my favourite songs). They were also asked if they were familiar with the song. The songs were taken from most popular tracks on Spotify of the music genres pop, rock, electronic, and ballads. (A list of musical excerpts used is provided in the Appendix)

The EEG data was acquired with the Emotiv Epoc+ wireless headset (Emotiv Systems, Inc., San Francisco, CA). The headset includes 14 active saline-based electrodes referenced to the left and right mastoid, with electrode placements AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4, as shown in Figure 3.2. Signals were digitised at the sampling frequency of 256 Hz, and band-pass filtered to an effective bandwidth of 0.16 — 45 Hz. All recording sessions were carried out in a professional studio environment with dim lights to avoid visual distractions.

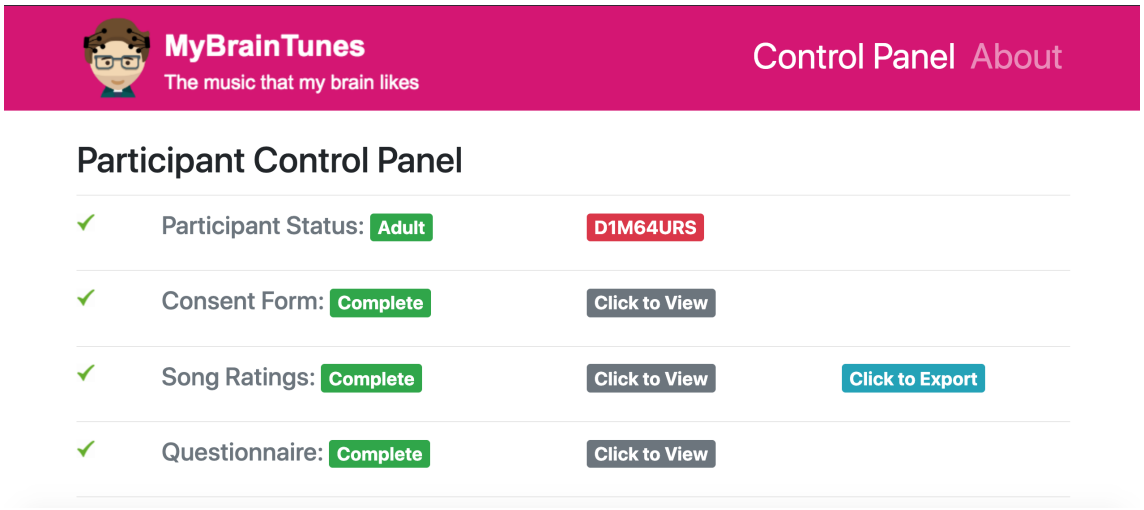

Figure 3.1: Screenshot of the control panel for the website mybraintunes.doc.ic.ac.uk. Participants were asked to fill out consent form, song ratings, and questionnaire.

Open source software *OpenSesame* [75] and Emotiv's EEG reocrding software (Testbench) were used to synchronise and automate the visual and audio setup of the experiment. Participants listened to music excerpts through in-ear earphones played through Genelec (Genelec, Finland) audio system.

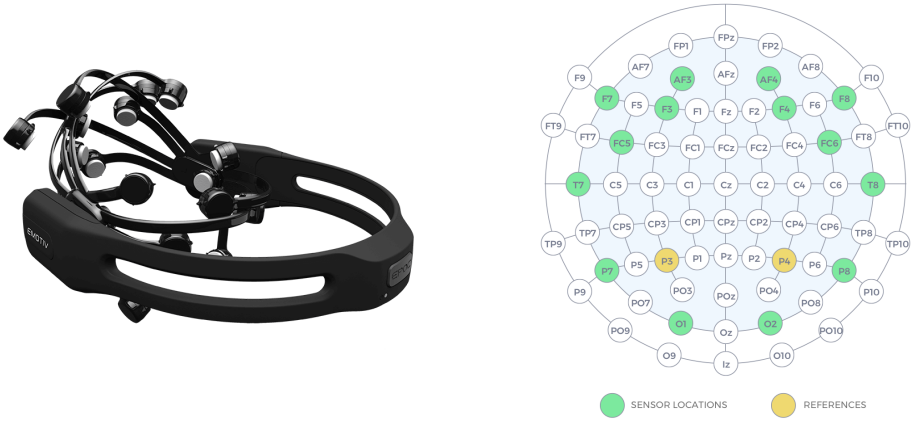

Figure 3.2: The Emotiv Epoc 14+ and the electrode locations in 10-20 system. Images taken from Emotiv Website [32].

3.1.3 Experiment

Participants were told to sit comfortably in a chair at a natural distance from the screen. As no EEG electrode caps were used, special care was taken to ensure electrodes were in the right positions for every head shape. Subjects were told to avoid excessive head movements, swallowing, minimise blinking, and to keep their eyes open and focus on the dot on the screen during the recording session. This is to

reduce muscular artifacts from movements and ocular artifacts from eyeblinks. As the electrodes are referenced to the mastoid (bone behind ear), movements such as jaw-clenching and swallowing would disrupt signal recordings. A test audio was then played so that the participant may adjust the volume of music to their comfort. As we wanted to measure the passive brain response of listening to liked music, participants were also told to enjoy the music experience during the experiment.

The recording sessions were divided into two 20 minute sessions, where each session contained excerpts from 30 songs that the participant had previously rated in the questionnaire. The music excerpts were played in random order. A resting period of 10 seconds was recorded at the beginning of each session to establish a reference for resting state baseline activity. During the recording session, a green or red dot was displayed onscreen for the participant to fixate their eyes on and to indicate the onset of music playing. A red dot indicates no music is playing and a green dot, preceded by a countdown, indicates music playing. Participants were given 10 second resting intervals in between 30 seconds of music playing (see Figure 3.3). The total number of trials was equal the total number of excerpts, i.e. 60 per subject so $22 \times 60 = 1320$ trials in total. Only the relevant 30 seconds of music listening was used for training.

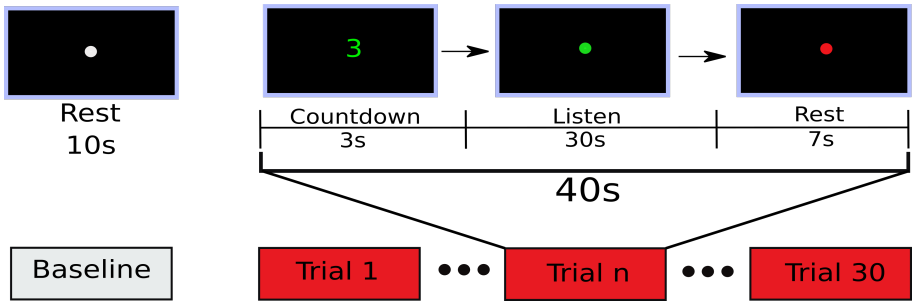

Figure 3.3: Experimental design for one session. There were two sessions and 60 trials in total for every subject.

3.1.4 Preprocessing

The recorded signals were digitised at 256 samples/s and band-pass filtered to 0.16 - 45 Hz. Independent Component Analysis (ICA) [99] was then used to clean the EEG data and remove artifacts. For each continuously recorded dataset, components that were associated with artifact activity from eyes, muscle and cardiac interference were identified and zeroed, and the estimated mixing matrix was used to reconstruct the multi-channel signal from the rest of ICs.

A filter bank was also applied to filter the signals from each sensor $x(t)$ into relevant frequency bands of standard brain rhythms. To increase the frequency resolution, the beta rhythm band was divided into beta_{low} and $\text{beta}_{\text{high}}$. The frequency ranges were thus, Delta: 1-4Hz, Theta: 4-8Hz, Alpha: 8-13Hz, Beta_{low}: 13-21Hz, $\text{Beta}_{\text{high}}$: 21-30Hz, Gamma: 30-45Hz.

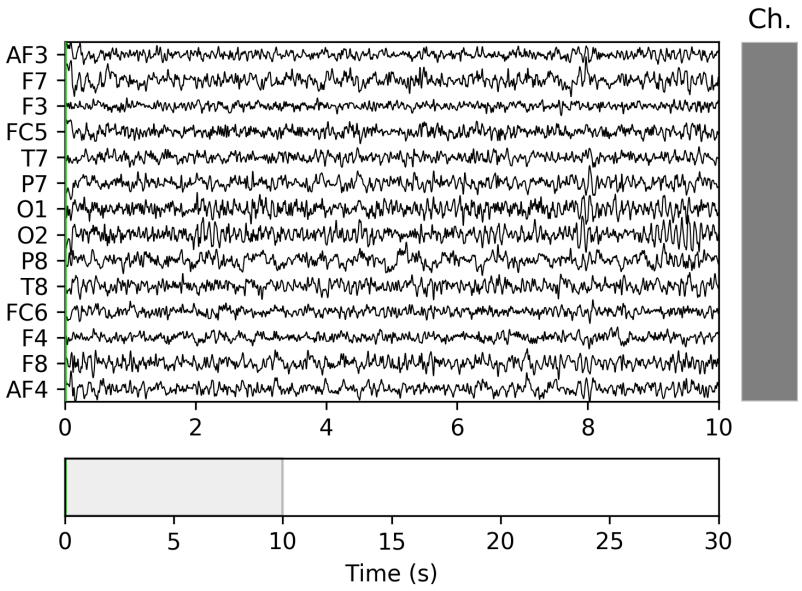

Figure 3.4: Example of raw data recorded from 14-channels from a single trial from one subject.

Additionally, the energy content envelope of the frequency bands were calculated by means of Hilbert transform, which transforms values into the frequency domain by computing a phase shift of 90 degrees to every Fourier component of a function $u(t)$ of real values. The envelope of the filtered activity $A_{\text{rhythm}}(t)$ represents the momentary strength of the associated oscillatory activity, and its relative contribution is derived by normalising with the total signal strength (summed from all brain rhythms).

As such, the implementation that follows makes use of three sets of data: the unprocessed *raw* data, the ICA-cleaned and filtered frequency bands (referred to in the following as *filtered*), and the envelope of the filtered frequency bands representing their relative strength (*envelope*). Finally, all trials were additionally normalised to zero mean and range [-1, 1].

3.2 Data Augmentation

Deep learning models require large amounts of data to learn discriminative features and prevent overfitting. In computer vision tasks, methods such as flipping, stretching, rotating, and adding noise are common ways to enlarge the dataset. For EEG time series data, a common strategy is to use overlapping crops generated by sliding a fixed-sized window over each EEG trial, and has been shown to increase CNN classification performance [97]. Formally, given an original trial $\boldsymbol{X}^i \in \mathbb{R}^{\mathrm{E}\cdot\mathrm{T}}$ with E electrodes and T timesteps, the sliding window generates a set of crops with crop size T' as timeslices of a given trial i:

$$\boldsymbol{C}^i = \left\{ \boldsymbol{X}^i_{[1,E],[t,t+T']} | t \in [1, T - T'] \right\} \tag{3.1}$$

Where all $T - T'$ crops are then used as new training data examples and receives the same label y^i as the original trial.

This cropping strategy aims to force the model into learning features that are present in all crops of the trial, as the model can no longer rely on the global temporal structures between the original trials. In this project, a sliding window frame size of 1280 sample points was used, i.e. 5 seconds given sampling frequency of 256Hz. The overlap between trials was 50% of the frame size, i.e. 2.5 seconds. This was chosen as previous works have determined that the impact of music-related emotional events happen on the scale of milliseconds and overlapping frame sizes between 3 to 5 seconds yielded the best accuracy for classification tasks [38] [39]. In total, this yields 5 new training samples per trial, increasing the training set by a factor of 5. Frame sizes of 1 second, 3 seconds, and 10 seconds were also tested, but did not yield a significant difference. Additionally, Gaussian noise was also added to the input data to increase the robustness of training.

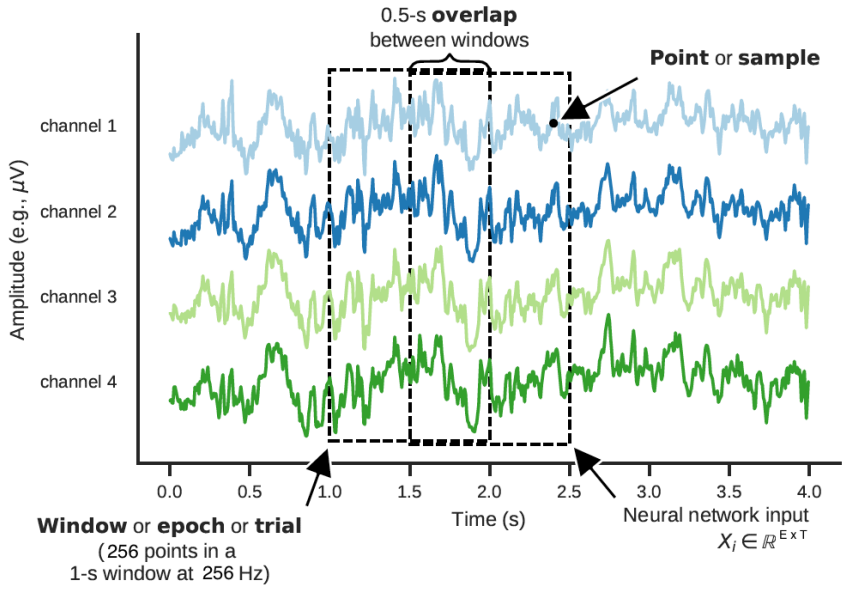

Figure 3.5: Illustration of cropped windowing approach. Figure modified from [91]

3.3 Architecture Implementation

3.3.1 Baseline Model

To benchmark the proposed architectures, a baseline model using feature extraction and a classifier. The benchmark model implementation follows [39] [38] closely,

as both are relevant works on the classification of music preference based on time-frequency analysis. In these works data collected from nine subjects during music listening with self-reported ratings of liking on a scale of 1-4 using features extracted from the beta and gamma bands using three different time-frequency distributions (spectrograms, Hilbert-Huang spectrum, Zhao-Atlas-Marks transform). The feature vectors were then classified into two classes, "like" and "dislike", using SVMs and *NN*.

Two differences should be noted on the implementation of the original paper and this implementation: [39] used an "active" listening paradigm in their experiment, meaning participants were asked to rate the music during the experiment, as opposed to listeing to the music passively as in this project. In addition, [38] only used binary classes for their classifier (like and dislike), whereas the goal of the model in this project is to predict the numerical rating of music preference in subjects (6 classes).

For this model, spectrograms were used as the time-frequency distribution (TFD). A TFD constitutes a two-dimensional spectral representation of a signal in the time and frequency domains. A spectrogram is calculated using the short-time Fourier transform (STFT), a type of linear TF representation. STFT involves pre-windowing a signal $x(\tau)$ around a time instant t and the calculating the Fourier transform for each t. The square modulus of the STFT then defines the spectrogram, which represents the energy distribution of the signal in the TF plane:

$$SPG_x(t,f) = \left| \int_{-\infty}^{+\infty} x(\tau) h^*(\tau - t) e^{-j2\pi f \tau} d\tau \right|^2 \tag{3.2}$$

where $h*(\tau - t)$ represents the short-time analysis window, (* denoting the complex conjugate). Spectrograms were generated using the Scipy function `scipy.signal.spectrogram`, using a non-overlapping Hamming window function.

Feature extraction was carried out using the method described in [38]. Based on the concept of event-related desynchronisation (EDS) and synchronisation (ERS), event-related activity can be seen as a proportional change of spectral activity in relation to a reference period. Thus after computing the TFD, from recording channel i and experimental trial j, feature F is computed in a time window W_{n} in the frequency band f_{b}

$$F^{f_b,w_n} = \frac{A^{f_b,w_n} - R^{f_b}}{R^{f_b}} \tag{3.3}$$

Where A is the average amplitude of a spectrogram of a given freqeuncy and window, and R is the average amplitude of a spectrogram from a resting trial. Subsequently, the feature vector is constructed as:

$$FV_j^{f_b,w_n} = \left\{ F_{j,1}^{f_b,w_n}, \ldots, F_{j,i}^{f_b,w_n}, \ldots, F_{j,N_c}^{f_b,w_n} \right\}$$

where i denotes the ith channel and N_{c} is the total number of channels.

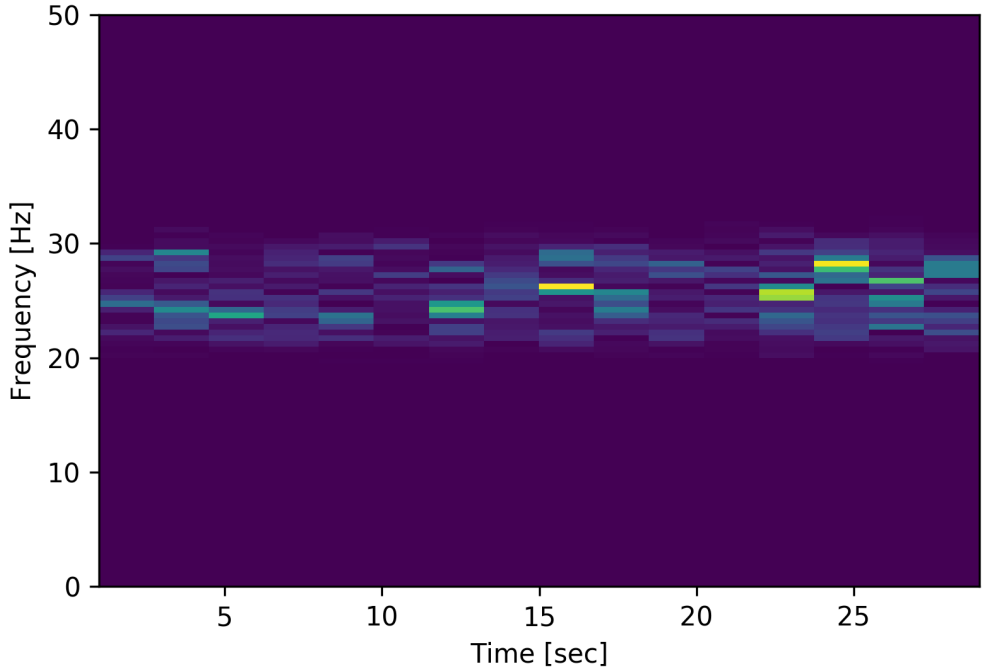

Figure 3.6: Spectrogram visualisation of the $\text{beta}_{\text{high}}$ frequency band between 21-30Hz for one trial

As the beta and gamma frequency bands were shown to be most important for music liking tasks [39] [5] [83] [70], the final feature vector was formed by concatenating feature vectors of these two bands only, giving:

$$FV_j^{w_n} = \left\{ FV_j^{beta,w_n}, FV_j^{gamma,w_n} \right\}$$

The total number of FVs computed is $\frac{30-0.5l_w}{0.5l_w}$, where l_{w} is the length of the window, 0.5 indicates a 50% overlap between window frames, and 30 is the total duration of a single trial in seconds. So for a window frame length of 1280 sample points (5 seconds), 11 FV sets were computed for every 30 second trial.

Each of the FV sets were fed into two different classifiers, *k*NN and SVM. These were implemented using the Python `sklearn` library. For *k*NN, the euclidian distance was used for the distance metric and the number of nearest neighbours *k* was set to 4 [38]. For SVM, the Gaussian radial basis function kernel was used, and hyperparameters were tuned using `GridSearchCV` with parameter values $C = \{0.1, 1, 10, 100\}$, $\gamma = \{0.001, 0.1, 1, 50\}$. For both classifiers, the entire dataset was used for training and testing using stratified 5-fold cross-validation.

3.3.2 Time Series as Input

As reviewed in Section 2.4.2 , CNNs are the most commonly used architectures in the DL-EEG literature and have shown success in learning EEG representations in various tasks. Three types of CNN architectures are presented which takes time series signal values as input: a 1D CNN, a 2D CNN, and DenseNet. As many studies reported the success of using raw data as input into CNNs [91], all architectures were experimented with using 3 types of inputs: raw, filtered, and envelope data, referred to in Section 3.1.4. All models were created using `Pytorch` with GPU acceleration.

1D Convolutional Network

Architecture As time series data have strong 1D convolution structure [65], an architecture which performs successive convolutions in the time domain and the spatial domain (electrode channels) was first created.

The input shape is $E \times T$, where E is electrode channels and T is time points in a window. 1D convolutions with a kernel size of 5 and a stride of 3 are used, followed by a ReLU activation function. Batch normalisation is then applied to reduce covariate shift in intermediate representations and improve robustness, and was shown to improve performance for DL-EEG studies [97] [53].

The number of filters is kept constant at 14, the number of electrode channels. In the fifth layer dropout [103] is used with a deactivation probability of 0.25 to reduce overfitting. Hence for every layer 14 feature maps are produced. This ensemble of operations is repeated 5 times. The axes are then swapped and 1D convolutions with a kernel size of 3 are applied to learn the representations in the spatial dimension. This is repeated for 2 layers, with a stride of 2 followed by a stride of 1, both followed by ReLU non-linearity and batch normalisation. Finally, the output is flattened and fed into 2 fully-connected layers. ReLU and dropout at 0.25 were again used, and the output dimension with `softmax` activation is 6, the number of classes. Finally, in order to recover label from the output softmax vector, the `argmax` was taken, returning the index with the largest entry as the prediction. The full architecture is shown in Table 3.1.

A number of different configurations in kernel size and stride, as well as window sizes (discussed in Section 3.2) were tested at 768, 1280, and 2560. As there is a large number of tunable parameter, exploring all possible formations can be extremely time consuming and no formal hyperparameter search was used.

Layer	**Operation**	**Stride**	**Output Shape**	**Parameters**
Input			$E \times T$	
1	14 $\times$ Conv1D (5)	3	14 $\times$ 426	994
	ReLU	-	14 $\times$ 426	-
	BatchNorm	-	14 $\times$ 426	56
2	14 $\times$ Conv1D (5)	3	14 $\times$ 141	994
	ReLU	-	14 $\times$ 141	-
	BatchNorm	-	14 $\times$ 141	56
3	14 $\times$ Conv1D (5)	3	14 $\times$ 46	994
	ReLU	-	14 $\times$ 46	-
	BatchNorm	-	14 $\times$ 46	56
4	14 $\times$ Conv1D (5)	3	14 $\times$ 14	994
	ReLU	-	14 $\times$ 14	-
	BatchNorm	-	14 $\times$ 14	56
5	14 $\times$ Conv1D (5)	3	14 $\times$ 4	994
	ReLU	-	14 $\times$ 4	-
	BatchNorm	-	14 $\times$ 4	56
	Dropout (0.25)	-	14 $\times$ 4	-
6	Swap axes	-	4 $\times$ 14	-
	4 $\times$ Conv1D (3)	2	4 $\times$ 6	52
	ReLU	-	4 $\times$ 6	-
	BatchNorm	-	4 $\times$ 6	16
7	4 $\times$ Conv1D (3)	1	4 $\times$ 4	52
	ReLU	-	4 $\times$ 4	-
	BatchNorm	-	4 $\times$ 4	16
8	Flatten	-	16	-
	FC	-	9	144
	ReLU	-	9	-
	Dropout (0.25)	-	9	-
9	Softmax	-	K	60
Total				5,590

Table 3.1: Proposed 1D CNN architecture. E is number of electrodes, T is number of time points, and K is number of classes. In this table E = 14, T = 1280 (3 seconds), K = 6. FC = Fully-Connected.

Training Training was carried out by optimising the cross-entropy loss function $\mathcal{L}(\hat{y}, y) = \sum_i y_i \log \hat{y}_i$ for multiclass classification and binary cross-entropy for binary classification. Adam, a variant of the mini-batch stochastic gradient descent algorithm, was used as the optimiser with an initial learning rate of 10^{-3} and decay rate of first and second moments of 0.9 and 0.999 respectively [59]. Adam has been shown to achieve fast convergence rates when used for CNN training and is the most popular choice of optimiser in DL-EEG studies [91]. The batch size was set to 32. The training ran for 100 epochs and was check-pointed and monitored by validation loss.

The training and validation sets were split into 80% and 20% stratified folds using `StratifiedShuffleSplit` in the `sklearn` library, so that each class is represented equally in training and validation. In addition, to address class imbalance, the loss for each class was weighted inversely proportionally to the probability of its label being chosen.

The model was trained to predict on the 6 classes of preference detailed in Section 3.1.2, or binary classifications of ”like” or ”dislike”, as some previous works have done [38] [83]. To convert to binary classes, Labels 0 and 1 were converted to ”dislike”, labels 4 and 5 were converted to ”like”, and labels 2 and 3 were discarded as they were intermediate responses. Synthetic minority over-sampling technique (SMOTE) [25] was then used to oversample the binary classes and balance the dataset. This was implemented using the `imblearn` library in Python.

The input was split into overlapping windowed frames as a means of data augmentation, as described in Section 3.2, for all 3 input datasets raw, filtered, and envelope. None-overlapping frames of sizes of 768, 1280, 2560 were also used for comparison, but no differences was found. For filtered and envelope data, a separate network was trained for every band of frequency: delta, theta, alpha, beta_{low}, $\text{beta}_{\text{high}}$, gamma. The frequency band with the best performance was chosen for the final model. The training and validation loss over 100 epochs is shown in Figure 3.7.

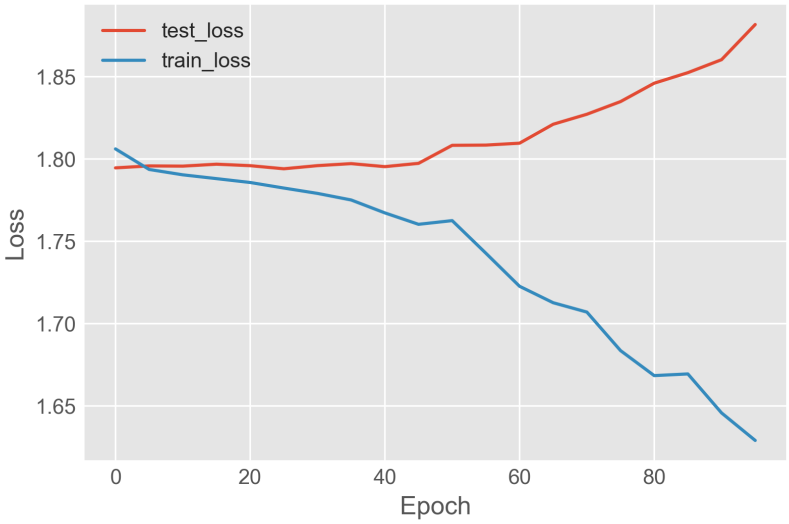

Figure 3.7: Training and testing loss for 1D CNN over 100 epochs for multiclass classification. A pattern of overfitting can be seen. Note the loss axis scale range is very close together in value for the purpose of visualisation.

2D Convolutional Network

Architecture An architecture using 2D convolutions was created based on the model described in [12] (Figure 2.12) and detailed in Table 3.2. The architecture uses a large kernel across the total number of channels in the first layer to learn global features in the spatial and temporal domain. This is similar to the use of a spatial filter in traditional EEG algorithms such as FBCSP, an approach also adopted by the architectures in [63] [97] [78] [31]. The first layer extracts spatial features with 32 filter kernels of 14×64 (for 14 electrode channels), and the second layer extracts temporal features with 64 filter kernels of 1×16. The stride is (1,2) for both layers, and an additional max-pooling operation is carried out after the second convolutional layer.

After the first two layers, the data is then treated like an image in the time and frequency domains by swapping axes, followed by 6 stacked convolutional layers with decreasing size and increasing number of filter kernels in deeper layers. This is similar to the architectures of popular ImageNet CNNs such as VGGNet and AlexNet [60] [100]. The 3rd and 4th layer each have 14 filters kernels of size 8×8 and stride (1,1). The 5th and 6th layers have 28 filters of size 1×4 with stride (1,1). The next two layers extract 28 filters of size 1×2 and stride (1,2). Max-pooling is applied every two layers and ReLU is used as the activation function in all layers. Finally, the fully-connected layer connects to 560 neurons and outputs to the softmax function to get the probability of each label.

As the amount of data in this project is limited, and the original architecture was used to accommodate EEG data with a larger number of channels and subjects, the number of layers is reduced and the number of filters is adjusted and kept at 28 for the successive layers to reduce the complexity and number of parameters in the network. To further prevent overfitting, Dropout [46] with a probability of 0.5 is used in all layers. Dropout regularisation has proved to be an effective method for reducing the overfitting in deep neural networks with millions of parameters [60] and in neuroimaging applications [85]. Batch normalisation is used in the fully-connected layers.

Layer	Operation	Stride	Output Shape	Parameters
Input			$E \times T$	
1	Unsqueeze		$1 \times 14 \times 1280$	
	$32 \times$ Conv2D (14×64)	(1,2)	$32 \times 1 \times 609$	28,704
	ReLU	-	$32 \times 1 \times 609$	-
	Dropout(0.5)	-	$32 \times 1 \times 609$	-
2	$64 \times$ Conv2D (1×16)	(1,2)	$64 \times 1 \times 297$	32,832
	ReLU	-	$64 \times 1 \times 297$	-
	MaxPool2D (1×2)	(1,2)	$64 \times 1 \times 148$	-
	Dropout (0.5)	-	$64 \times 1 \times 148$	-
3	Swap axes		$1 \times 64 \times 148$	-
	$14 \times$ Conv2D (8×8)	(1,1)	$14 \times 57 \times 141$	910
	ReLU	-	$14 \times 57 \times 141$	-
	Dropout(0.5)	-	$14 \times 57 \times 141$	-
4	$14 \times$ Conv2D (8×8)	(1,1)	$14 \times 10 \times 134$	12,558
	ReLU	-	$14 \times 10 \times 134$	-
	MaxPool2D (5×3)	(5,3)	$14 \times 10 \times 44$	-
	Dropout(0.5)	-	$14 \times 10 \times 44$	-
5	$28 \times$ Conv2D (1×4)	(1,1)	$28 \times 10 \times 41$	1,596
	ReLU	-	$28 \times 10 \times 41$	-
	Dropout (0.5)	-	$28 \times 10 \times 41$	-
6	$28 \times$ Conv2D (1×4)	(1,1)	$28 \times 10 \times 38$	1,596
	ReLU	-	$28 \times 10 \times 38$	-
	MaxPool2D(1×2)	(1,2)	$28 \times 10 \times 19$	-
	Dropout (0.5)	-	$28 \times 10 \times 19$	-
7	$28 \times$ Conv2D (1×2)	(1,2)	$28 \times 10 \times 9$	1,596
	ReLU	-	$28 \times 10 \times 9$	-
	Dropout (0.5)	-	$28 \times 10 \times 9$	-
8	$28 \times$ Conv2D (1×2)	(1,2)	$28 \times 10 \times 4$	1,596
	ReLU	-	$28 \times 10 \times 4$	-
	MaxPool2D(1×2)	(1,2)	$28 \times 10 \times 2$	-
	Dropout (0.5)	-	$28 \times 10 \times 2$	-
9	Flatten	-	560	-
	FC	-	560	314,160
	ReLU	-	560	-
	BatchNorm	-	560	2,240
	Dropout (0.5)	-	560	-
10	Softmax	-	K	3,366
Total				401,154

Table 3.2: Proposed 2D CNN architecture. E is number of electrodes, T is number of time points, and K is number of classes. In this table $E = 14$, $T = 1280$ (3 seconds), $K = 6$. FC = Fully-Connected.

Training The model was trained using the same parameters as detailed for the 1D CNN previously. Additionally, L_2 regularisation of weight decay at 0.0005 was also used for the training of the network to reduce overfitting. Training and validation loss over 100 epochs is seen in Figure 3.8. The loss did not decrease over the epochs.

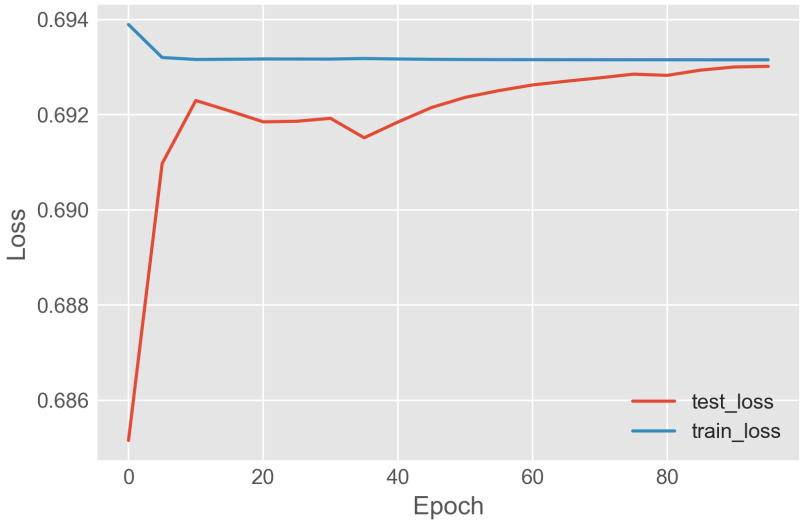

Figure 3.8: Training and test loss for 2D CNN over 100 epochs for binary classification. Note the loss axis scale range is very close together in value for the purpose of visualisation.

DenseNet

The last architecture is an implementation of Dense Convolutional Network (DenseNet) [49]. DenseNets contain layers which connect to every other layer in a feedforward fashion. For each layer, the feature maps of all preceding layers are used as inputs, and its own feature maps are used as inputs into all subsequent layers. Similar to ResNet [44], DenseNet encourages information flow through all layers in the network, but instead of through summation, DenseNet combines features by concanetating them. DenseNets have several advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. Given the limitation of data in this project, a network with reduced parameters is advantageous. DenseNet has also been previously used by [122] for EEG classification.

The implementation is based on the code and models given in the original paper [49]. The size of the convolutions are similar to the 1D CNN described before, where convolutions are done on the time domain first, then on the spatial domain. The network takes in $1 \times E \times T$ input where E is electrode channels and T is sample points in time, as before. The growth rate is set to 4, the reduction value in the bottleneck convolution layers is set to 0.5, and number of layers in each block is set to 5. The first convolution has a kernel size of 1×5 and a stride of 1×3. Each dense block contains the sequence of operations BatchNorm-ReLU-Conv, where the kernel size is 1×5 and the stride is 1. The dense block is followed by the transition layer consisting of a BatchNorm-ReLU-Conv with size 1×1 and stride 1, followed by an average-pool of 1×2. In total there were 3 dense blocks followed by transition layers.

The output of the DenseNet blocks is then put through 2 extra convolutional layers for the network to learn representations in the spatial domain. Both layers are put through the BatchNorm-ReLU-Conv sequence, as in the dense blocks, with a kernel size of 3×1 and stride 2. This is followed by an average-pool of 1×3. Finally, a fully-connected layer with 32 neurons is followed by softmax to give the final classification.

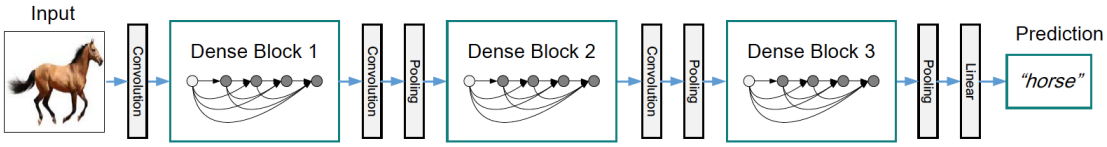

Figure 3.9: DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature map sizes via convolution and pooling. [49]

Layer	Operation		Stride	Output Shape
Input				$E \times T$
	Unsqueeze			$1 \times E \times T$
Convolution	BN-ReLU-Conv(1×5)		(1,3)	$8 \times 14 \times 426$
Pool	MaxPool(1×2)		(1,2)	$8 \times 14 \times 213$
Dense Block 1	[BN-ReLU-Conv(1×1) BN-ReLU-Conv(1×5)]	$\times 4$	(1,3)	$4 \times 14 \times 213$
Transition Layer 1	BN-ReLU-Conv(1×1)		(1,1)	$4 \times 14 \times 213$
	AvgPool(1×2)		(1,2)	$2 \times 14 \times 106$
Dense Block 2	[BN-ReLU-Conv(1×1) BN-ReLU-Conv(1×5)]	$\times 11$	(1,3)	$2 \times 14 \times 106$
Transition Layer 2	BN-ReLU-Conv(1×1)		(1,1)	$2 \times 14 \times 106$
	AvgPool(1×2)		(1,2)	$2 \times 14 \times 53$
Dense Block 3	[BN-ReLU-Conv(1×1) BN-ReLU-Conv(1×5)]	$\times 12$	(1,3)	$2 \times 14 \times 53$
Convolution	BN-ReLU-Conv(3×1)		(2,1)	$2 \times 6 \times 53$
Convolution	BN-ReLU-Conv(3×1)		(2,1)	$2 \times 2 \times 53$
Classification	AvgPool(2×2) 32D fully-connected, softmax		(1,2)	$2 \times 1 \times 26$

Table 3.3: Modified DenseNet implementation. E is number of electrodes, T is number of time points, and K is number of classes. In this table $E = 14$, $T = 1280$ (3 seconds), $K = 6$. BN = Batch Normalisation.

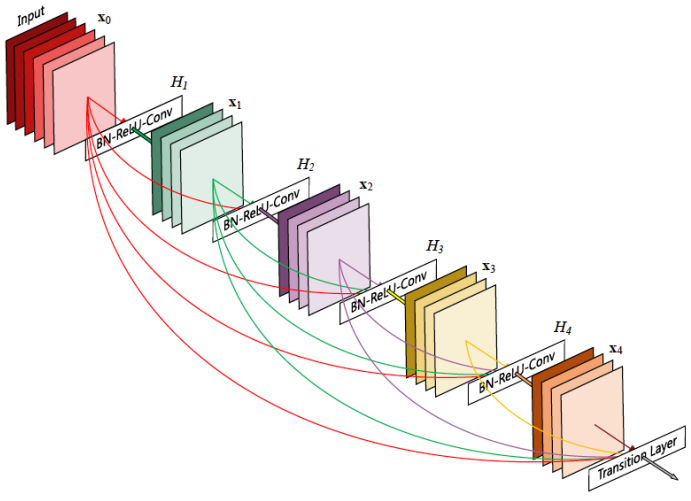

Figure 3.10: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature maps as input. [49]

Training The training parameters are the same as described before, with the exception that stochastic gradient descent (SGD) was used instead of Adam. Following the original paper [49], a weight decay of 10^{-4} and a Nesterov momentum of 0.9 was used. The dropout rate was set to 0.2. Overlapping frame sizes of 1280 and 2560 were tested. DenseNets with 3 or 5 dense blocks were also tested, but no difference was found. The training and validation loss over 100 epochs is seen in Figure 3.11. The loss stayed the same over the epochs and did not decrease.

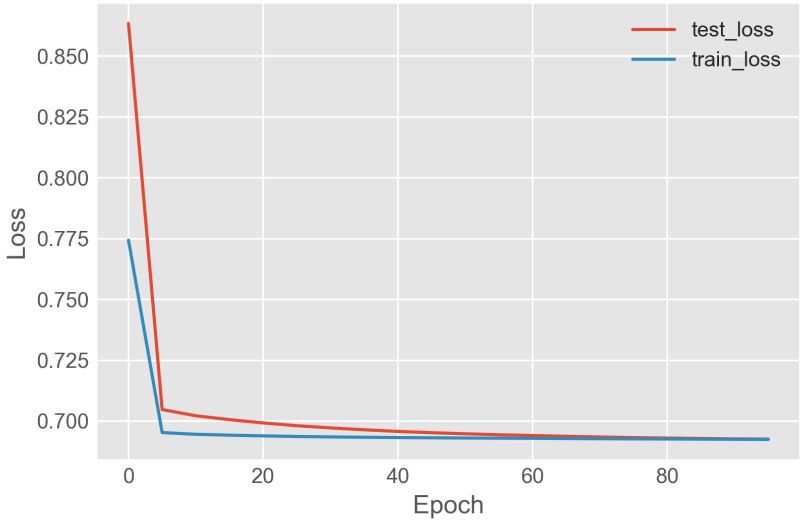

Figure 3.11: Training and test loss for DenseNet over 100 epochs for binary classification. Note the loss axis scale range is very close together in value for the purpose of visualisation.

3.3.3 Spectrograms as Input

Time-frequency representation of EEG data

As CNNs are especially suited for processing imaging data, an alternative approach is to transform EEG signals into two-dimensional image representations. One important constraint when doing so is to retain both time and frequency information in an image — similar to speech signals, the most salient features of multi-channel EEG often reside in the frequency domain.

One way to achieve this is to transform the data into spectrograms, which takes a Fourier transform of sub-segments of the signal to compute its power spectrum, as described in Section 3.3.1. However, concatenating spectral measurements of all the electrodes into a single feature vector clearly ignores the inherent structure of the EEG data in time, frequency, and space.

Instead, the EEG channels can be treated like colour channels of an image. Once transformed into an image, each pixel along the vertical axis of the spectrogram corresponds to spectral frequency, and each pixel along the horizontal axis corresponds to the time bins. The numerical ”brightness” value of a pixel is then equal to the output value of the spectrogram at the particular time and frequency corresponding to that pixel. This transformed time series for each individual channel can be treated as a monochromatic image. By stacking together spectrogram images formed by each EEG channel, the same way colour channels are stacked together in colour images, spectrograms generated from multi-channel EEG data can be processed in the same way standard images are processed in computer vision tasks, and preserving inherent structures in frequency and time.

The spectrograms were generated prior to input into the network, as detailed in Section 3.3.1, using the `scipy.signal.spectrogram` function with a Hann windowing function. For every electrode channel, a spectrogram is generated for each relevant frequency band. As neural networks expect normalised inputs, the logarithm of the spectrograms are taken and used as input into the following CNNs (Figure 3.12).

Similar to [17], a single-frame approach and multi-frame approach were used. In the single-frame approach, a single spectrogram image was generated by applying FFT to the whole trial duration (30 seconds), so that for every trial only a single multi-channel image was produced. The time window length is set to 1280 with overlapping segments of 640 (50% overlap), generating images with 165 frequency bins and 45 time bins. In the multi-frame approach, a spectrogram is computed for a single segment (5 seconds) (256 window size with 50% overlap) and this input goes through data augmentation detailed in 3.2 to generate 11 images of 34 frequency bins and 37 time bins per trial. A particular challenge for spectrograms is the trade-off between time and frequency.

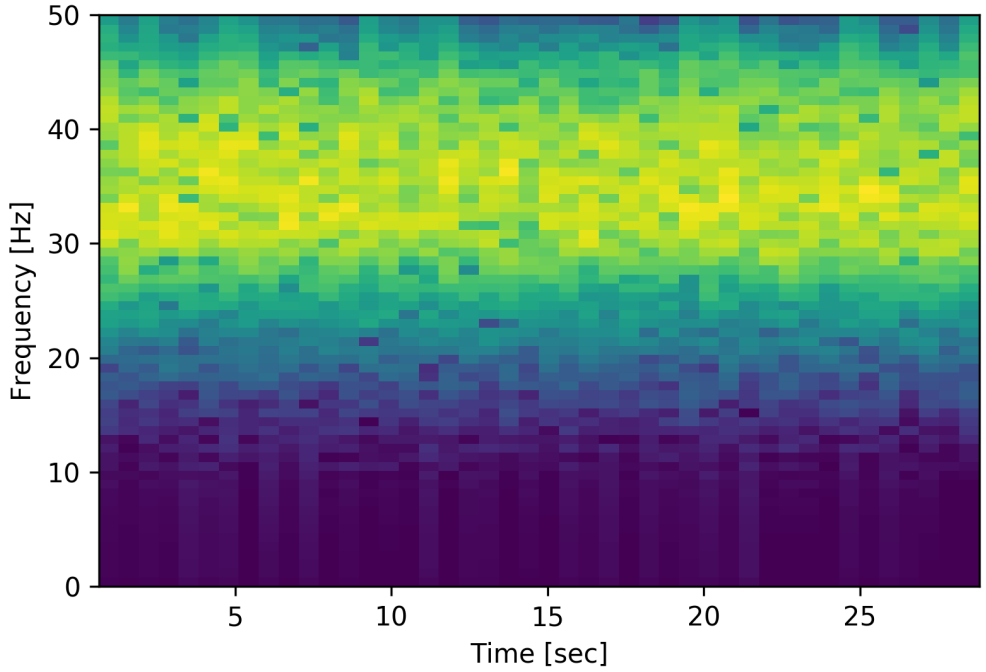

Figure 3.12: Log spectrogram visualisation with a window length of 1280 and overlap of 640 with Hann window function of gamma frequency band from a single channel in a trial where the subject gave the rating ”5”. More brightness in colour (yellow) indicates higher spectral values, and lower brightness in colour (blue) indicates lower spectral values.

SpecNet

Architecture SpecNet is a network inspired by the architecture described in [86], which is a shallow network with small convolution sizes. As the number of distinct spectral features in EEG is much smaller (homogeneous) than it is in a natural image, the model complexity should be reduced to prevent overfitting.

The network takes the spectrogram dimensions frequency $\times$ time from each of the 14 electrode channels as input. It then uses 8 filter kernels to perform convolutions of size 14×3 — 14 for the number of electrode channels to learn global feature correlates. The next layer has 16 kernels with a small kernel size 1×1 followed by a max-pooling layer with size 1×2. A dropout of 0.5 is used on the output of the layer, which then inputs into the fully-connected dense layer with 128 dimensions. A dropout of 0.5 is used again on the output and finally the softmax activation gives the output. ReLU is used as the activation function in all layers.

Layer	Operation	Stride	Output Shape	Parameters
Input			$E \times F \times T$	
1	$8 \times$ Conv2D (14×3)	(1,2)	$8 \times 21 \times 18$	4,712
	ReLU	-	$8 \times 21 \times 18$	-
2	$16 \times$ Conv2D (1×1)	(2,2)	$16 \times 11 \times 9$	144
	ReLU	-	$16 \times 11 \times 9$	-
	MaxPool2D (1×2)	(1,2)	$16 \times 11 \times 4$	-
	Dropout(0.5)	-	$16 \times 11 \times 4$	-
3	Flatten	-	704	-
	FC	-	128	90,240
	ReLU	-	128	-
	Dropout (0.5)	-	128	-
4	Softmax	-	K	774
Total				95,870

Table 3.4: Proposed SpecNet architecture. E is number of electrodes, F is number of frequency bins in specrogram, T is number of time bins in spectrogram, and K is number of classes. In this table $E = 14$, $F = 34$, $T = 37$, $K = 6$. FC = Fully-Connected.

Training The training parameters are as described in 1D CNN, with the exception that RMSProp was used as the optimiser and a weight decay of 0.0005 was used. A classifier was trained for each of the six frequency bands using the filtered dataset. The training and validation loss is shown in Figure 3.13. Over 100 epochs, the loss value barely decreases and no difference was found with longer training epochs.

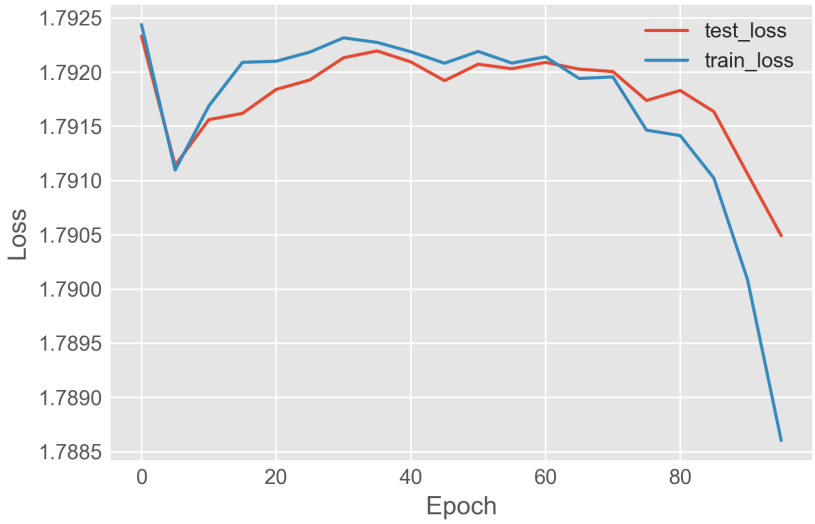

Figure 3.13: Training and test loss for SpecNet over 100 epochs for multiclass classification. Note the loss axis scale range is very close together in value for the purpose of visualisation.

ResNet

Architecture The last architecture tested is an implementation of a Residual Network (ResNet) [44]. ResNet is a highly successful state-of-the-art ImageNet classifier widely used in image recognition tasks. ResNets add the input of a convolutional layer to the output of the same layer, to the effect that the convolutional layer only has to learn to output a residual that changes the previous layer's output. This identity short-cut function is shown in Figure 3.14 — the shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers. This allows ResNets to be successfully trained with a much larger number of layers than traditional convolutional networks. [97] have experimented with a 31-layer ResNet for EEG motor imagery classification, but this was on raw EEG data rather than image representations.

The idea with this implementation is to treat the spectrogram representations as natural images and test whether standard image classifiers could discriminate feature representations in spectrograms as in natural images. A ResNet with 18 layers, modified from the implementation in the `torchvision` library was trained from scratch using 14 channels instead of the original 3 channel architecture designed for coloured-images. To prevent overfitting, ResNet-18, the shallowest of ResNet architectures, is chosen, and the number of feature maps in each layer is reduced. As in the original implementation, a 7×7 convolution is performed on the 14 channel input with stride 2, followed by 3×3 max-pooling. There are then 4 blocks of 3×3 convolutions with with layer size 2. However, instead of using increasing sizes of 64, 128, 256, 512 feature maps in each block in the original network, this implementation reduces the feature maps to 8, 16, 16, 32. This is because the discriminative spectral features in EEG data is far less distinctive and fewer than it would be in a natural image.

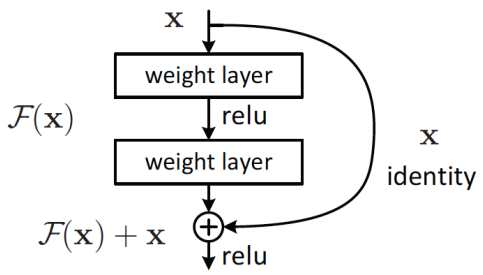

Figure 3.14: Residual block used in ResNet[44]

Layer	Operation		Stride	Output Shape
Input				$E \times F \times T$
Conv1	Conv(7×7)-BN-ReLU		(2,2)	$8 \times 17 \times 19$
Pool	MaxPool(3×3)		(2,2)	$8 \times 9 \times 10$
Conv2-X	Conv(3×3)-BN-ReLU Conv(3×3)BN-ReLU	$\times 2$	(2,2)	$8 \times 9 \times 10$
Conv3-X	Conv(3×3)-BN-ReLU Conv(3×3)-BN-ReLU	$\times 2$	(2,2)	$16 \times 5 \times 5$
Conv4-X	Conv(3×3)-BN-ReLU Conv(3×3)-BN-ReLU	$\times 2$	(2,2)	$16 \times 3 \times 3$
Conv5-X	Conv(3×3)-BN-ReLU Conv(3×3)-BN-ReLU	$\times 2$	(2,2)	$32 \times 2 \times 2$
Classification	AvgPool(1×1) 1000D fully-connected, softmax		(1,1)	$32 \times 1 \times 1$

Table 3.5: Modified ResNet implementation. E is number of electrodes, F is number of frequency bins in spectrogram, T is number of time bins in spectrogram, and K is number of classes. In this table $E = 14$, $F = 34$, $T = 37$, $K = 6$. BN = Batch Normalisation.

Training The training procedure was as described previously, with Adam as the optimiser and a weight decay set to 0.0001. The batch size was set to 64. The training and validation loss is seen in Figure 3.15. Over 100 epochs, the decrease in loss is very little and plateaus after around 50 epochs.

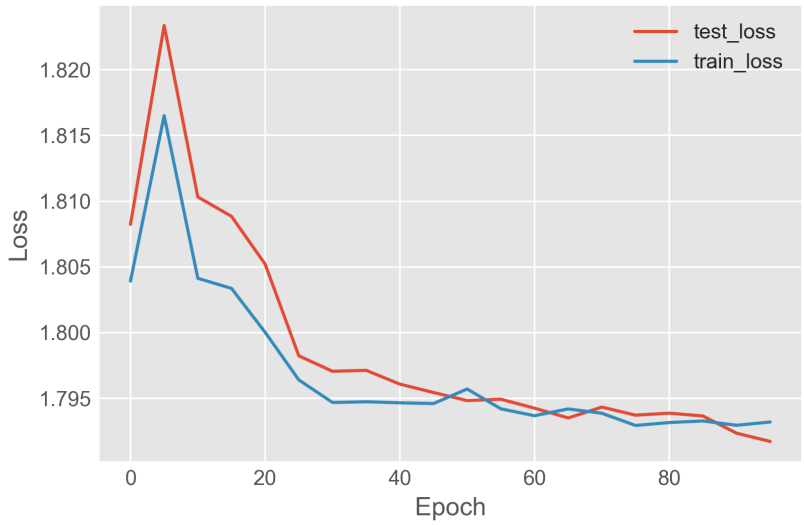

Figure 3.15: Training and test loss for ResNet over 100 epochs for multiclass classification. Note the loss axis scale range is very close together in value for the purpose of visualisation.

Chapter 4

Results

Architecture	6-Class Test Accuracy	Binary Test Accuracy
SVM	45.13% ± 0.80%	77.82% ± 1.65%
*k*NN	44.11% ± 1.36%	76.72% ± 1.28%
2D CNN	25%	61.07%
SpecNet	21.59%	61.08%
ResNet-18	20.42%	50.62%
DenseNet	19.62%	49.26%
1D CNN	19.55%	49.40%
Chance	16.67%	50%

Table 4.1: Classification accuracy results for architectures tested.

Five different CNNs were tested using three different types of data: raw time series data, processed time series data, and spectrogram representations. The models were used to predict a 6 classes of 0 - 5 ratings of liking and binary classes of ”liked” and ”disliked”, which consisted of the labels 0, 1 for ”disliked” and the labels 4, 5 for ”liked”.

The two baseline models surpassed the performance of all deep learning models. Out of the two models, SVM performed best with 45.13% ± 0.80% for 6-class classification and 77.82% for binary classification. *k*NN performed comparable to SVM with 44.11% ± 1.36% and 76.72% ± 1.28%. Both are well above chance level and the confusion matrices are shown in Figure 4.1 and 4.2. Nevertheless, one thing to note is that *k*NN classification required considerably less computation time compared to SVM, a technical parameter which may be significant for the design of a future real-time music appraisal recognition system.

Our models did not reach the performance reported in the original paper [38], who reported a test accuracy of 86% for binary classification in *k*NN. However, [38] used an active paradigm in their experimental design, whereas we used a passive listening paradigm, which is a harder classification task. Taking this into consideration, and the fact that no extensive optimisation has been done for these models,

the results show that traditional feature extraction and machine learning methods are able to classify music liking to a good extent.

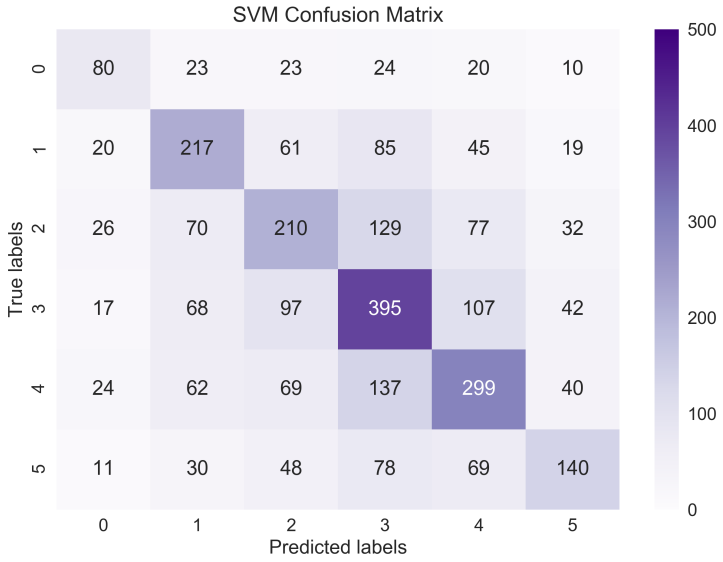

Figure 4.1: Confusion matrix for 6-class classification using SVM

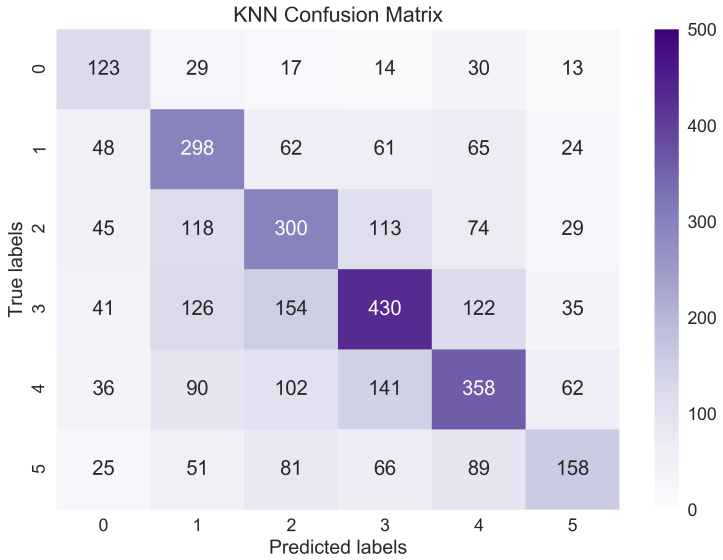

Figure 4.2: Confusion matrix for 6-class classification using kNN

The deep learning models, on the other hand, were not able to effectively classify music preference. 1D CNN performed worse, with test accuracies 19.55% for multiclass and 49.4% for binary classification — basically at chance level. As the least complex model of the five architectures, the model was not able to learn the EEG

features at all. Although DenseNet has been successfully used for EEG classification before [122], it also failed to learn features for music liking on our datasets.

Raw, filtered, and envelope datasets, as detailed in Section 3.1.4, were tested as inputs for the time series models 1D CNN, 2D CNN, and DenseNet, as recent literature has shown promise for the use of raw data as inputs into CNNs [91] [29]. However, no difference was found between the datasets on all of the modelss. Although 2D CNN appeared to perform slightly better with raw data than on envelope or filtered data, it was not a significant difference and likely due to noise, given that there was no change in training and validation loss, as seen in Figure 3.8.

Models using spectrograms as input have also failed to classify music preference on the dataset. At 21.59% and 20.42% for SpecNet and ResNet respectively, the performance is close to chance and this is shown in the plateauing high loss in training and validation. A recent work by [117] also showed that without aggressive data augmentation, ResNet failed to surpass traditional machine learning methods. Although the loss appeared to decrease at a very small scale for SpecNet (Figure 3.13, tuning different learning rates and training with longer epochs did not improve the performance, as the models simply started to overfit. Using single-frame and multi-frame approaches also did not make a difference to the performance of both models.

All models were also experimented with using different configurations of regularisation techniques, such as batch normalisation and dropout, which are used in state-of-the-art classifiers [49] and have shown to significantly improve performance in EEG classification studies [96] [97]. However, this did not improve the performance of the models on our dataset. Different lengths of window overlap frames (Section 3.2), at 768 (3 seconds), 1280 (5 seconds), and 2560 (10 seconds), have also been tested, but also failed to make a difference. All results in Table 4.1 used 1280 data points per window frame.

Inter-subject variability is an important limitation for EEG data. As EEG classification is hard to generalise across subjects, a per-subject classification attempt was also attempted — that is, a model was trained for every individual subject. However, this did not yield any effective results, as the amount of data for a single subject is too limited.

Chapter 5

Discussion and Conclusion

This project aimed to examine whether deep learning architectures can be used for the prediction of music preference and the feasibility of a deep learning-powered bio-personalised music recommendation system. On all accounts, the deep learning models have failed to surpass the baseline models using traditional machine learning. The most likely explanation for this is that there was simply not enough data for the deep learning networks to learn the representations of music liking effectively.

Nonetheless, below lists some points of improvement and further approaches for the implementations of this project:

- More rigorous hyperparameter optimisation. No formal hypermarameter optimisation methods were used for the deep learning methods, as the focus was on experimenting with different architectures. A more systematic search using methods such as grid search [19] or Bayesian optimisation [101] can be used to find a good set of hyperparameters, an important aspect for deep learning algorithms.

- Experimenting with hybrid architectures. Related works have shown that hybrid architectures combining CNNs and LSTMs have achieved good performance using stacked CNN layers to process spatial information and LSTM cells to process temporal information [17][45] [62][29]. In the early stages of this project, basic LSTM architectures were tested, but were not pursued as the performance was no better than CNNs and due to time constraints. On a bigger dataset, CNN+LSTM architectures could be an interesting approach for EEG classification of music liking.

- Separating music ratings by familiarity. Our questionnaire included ”yes” or ”no” ratings of music familiarity that could’ve been utilised to aid classification of music preference for both traditional machine learning and deep learning approaches. [38] separated their classes into liked and familiar music and like and unfamiliar music and achieved higher classification music for liked and familiar music, possibly due to the role of familiarity to music-induced affective responses [94].

- Pre-selecting critical channels. To reduce the complexity of computing and increase classification accuracy, [86] used pre-selected channels relevant to emotion recognition tasks. This strategy could potentially be used on our dataset as a feature reduction strategy, as [5] [57] have found that the EEG channels located over the prefrontal cortex, in particular channel AF3, is the most important for classifying music liking.

As we often say, data is everything. Although recent reviews and studies [91] [29] [126] have reported favourable classification performance for deep learning model, [71] pointed out that for classification tasks with limited data, deep learning approaches are actually the worst classifiers and are far out-performed by traditional EEG and machine learning classifiers. Such may be the case for this project. With 60 trials from 22 subjects, the total number of trials used for training is 1320. Even after rigorous data augmentation, the total sample size for training and testing is only 14,520 across the 6 classes (with window frames of 1280). In contrast, computer vision tasks often have large, publicly available datasets on the scale of millions of samples for training. The fact that the traditional learning methods were able to successful classify music liking adds the argument that the problem for the deep learning approach lies in the limited amounts of data.

The difficulty of training a deep learning model on limited data is compounded by the fact that we used a more challenging passive BCI paradigm. It is conventional wisdom that a good feature of input into a CNN should have visual differences distinctive enough to be apparent to a human observer. This is true of EEG classification tasks such as epileptic onset detection or sleep stage classification, as these tasks traditionally relied on the visual inspection of human experts anyway. Many successful EEG classification models also use ERP-based paradigms, which have more apparent changes in waveform than passive BCI tasks. Although the neural substrates of music preference and EEG biomarkers for music liking is well-established (Section 2.2), it is still a relatively new sub-field, and to the best of this author's knowledge, there are currently no published deep learning models specific to the classification of music liking.

Another challenge for EEG classification is inter-subject variability. EEG data is hard to generalise across different individuals. It might not be enough to have many samples from few subjects — instead, the training data should aim to have as many different subjects as possible for a classifier to be able to classify across different individuals. The data collected for this project had a gender imbalance skewing in favour of males, which may have biased the training data.

Depending on the implementation of a bio-personalised music recommendation system, an opposite approach might be to eliminate the challenge of inter-subject variability altogether — perhaps collecting hours of data from a single subject and training a personalised classifier, similar to the scheme shown in Figure 2.9. In the EEG studies previously reviewed, those that train classifiers on a per subject basis can reach much higher accuracies than those that train classifiers that generalise across subjects [9] [91]. The downside, of course, is that a separate classifier would

have to be trained for every new user, which can be computationally expensive, time-consuming, and impractical to implement in real life.

A promising solution to the limited data and inter-subject variability problem is *transfer learning*. Transfer learning is based on the idea that knowledge used in solving one task could also be useful for another related task. For example, for the mobile EEG monitoring of epileptic seizures, [82] used transfer learning to fine-tune generalised models of epileptic onset using patient-specific data and found that the approach outperformed both the generalised model and purely patient-specific models.

Thinking more broadly, we can also use transfer learning for broader fields that share similarities with EEG classification. A good example is audio signal processing. The extraction of spectral features in the frequency domain is very similar for audio and EEG processing — both use Fourier transformed spectrograms, and audio recognition is a much more common classification task than EEG. The acquisition of EEG data is time-consuming and often differ in experimental protocols depending on the task — for this reason EEG data is often privately acquired and there are very few publicly available datasets [91]. In contrast, large-scale datasets of labelled audio events are publicly available, such as AudioSet by Google Research [90]. These datasets could be used to pre-train deep learning layers which can then be used as feature extractors in the frequency domain for EEG specific models, such as SpecNet as implemented in this project. In addition, audio signals could also potentially be pitch-shifted and re-sampled (audio signals have different sampling rates and frequency ranges) to match wave-forms that are more characteristic of EEG data.

Trying to determine the subjective emotional evaluation of a piece of music somebody listened to based on EEG is a challenging problem. Attempting to do this with a small training set makes the task even harder. Overall, despite the negative results from the deep learning models, this work contributed to the collection of new EEG data and showed that music liking on a scale of 6 classes can be classified using spectrogram-based feature extraction methods and traditional machine learning. Given larger datasets, future works may focus on using transfer learning to improve deep learning for music preference classification.

Appendix A

List of Songs

Artist	**Title**
John Lennon	Imagine
Cutting Crew	(I Just) Died In Your Arms
Tears for Fears	Shout
U2	With or without you
Def Leppard	Love Bites
Morrisey	Every day is like Sunday
Sinead OConnor	Nothing Compares to U
R.E.M.	Losing My Religion
Manic Street Preachers	Motorcycle Emptiness
Everything but the girl	Missing
No Doubt	Dont Speak
Mansun	Wide Open Space
Spice Girls	Wannabe
The Killers	Mr Brightside
Leona Lewis	Bleeding Love
Coldplay	Viva la vida
Alicia Keys	Empire state of mind (feat Jay z)
B.o.B	Nothin On You (feat. Bruno Mars)
Bruno Mars	Grenade
Adele	Someone like you
David Guetta	Titanium (feat Sia)
Emeli Sande	Read All About it Part III
Rihanna	Diamonds
Birdy	Wings
Daft Punk	Get Lucky ft. Pharrell Williams
Ed Sheeran	Sing
Taylor Swift	Shake it Off
Ellie Goulding	Love Me Like You Do
Jason Derulo	Grenade
Ed Sheeran	Perfect

Table A.1: List of 30 songs used in the experiment in random order. This would consist of one experimental session.

Artist	Title
Modjo	Lady Hear Me Tonight
Raining Pleasure	Fake
Snow Patrol	Chasing Cars
King of Leon	Use somebody
The Black Eyed Peas	I Gotta Feeling
Bruno Mars	Just the way you are
Eminem	Love The Way You Lie (ft. Rihanna)
Jessie J	Price Tag ft. B.o.B
Emeli Sande	My Kind of Love
P!nk	Try
Florence + The Machine	Spectrum
Ed Sheeran	Thinking out Loud
Pharrell Williams	Happy
The Weeknd	Cant Feel My Face
Adele	When we were young
Justin Timberlake	Cant Stop the Feeling!
Ed Sheeran	Shape of You
Chainsmokers & Coldplay	Something Just like this
Led Zeppelin	Stairway To Heaven
Aerosmith	Dream On
George Michael	Careless Whisper
Pet Shop Boys	Its a sin
Whitesnake	Is this love
Alannah Myles	Black Velvet
Depeche Mode	Enjoy the Silence
Queen	Bohemian Rapsody
Nirvana	Smells Like Teen Spirit
Tasmin Archer	Sleeping Satellite
Oasis	Wonderwall
Elton John	Something About the Way You Look Tonight

Table A.2: List of 30 songs used in the experiment in random order. This would consist of one experimental session.

Appendix B

Ethics Consideration Checklist

The table below outlines key ethical questions which were considered during this project. Below expands on any questions which were answered positively.

- **Does your project involve human participants?**
  22 participants were recruited to participate in the EEG experiment, approved by ethics declaration in the department. All participants signed an informed consent form and all data were anonymised.
- **Does your project involve personal data collection and/or processing?**
  All participant's EEG data were used as training data and stored on Imperial College London Department of Computing servers. They were also asked to complete a questionnaire which asked information on their ethnicity, age, disabilities, and education level.
- **Does it involve the collection and/or processing of sensitive personal data?**
  See previous answer.
- **Does it involve tracking or observation of participants?**
  Participant's EEG data were tracked during the task of music listening.
- **Does your project involve further processing of previously collected personal data (secondary use)?**
  A dataset of EEG data collected previously from associates in Greece was used in the early stages of the project to test architectures and experiment on.
- **Does your project have the potential for military applications?**
  BCIs are already an area of interest to the Defense Advanced Research Projects Agency Program (DARPA) in the USA. Potentially military applications might include mental state monitoring for soldiers or controlling drones with the mind.
- **Does your project have an exclusive civilian application focus?** The project is intended to assess the feasibility of a bio-personalised music recommendation system using deep learning.

- **Are there any other ethics issues that should be taken into consideration?** All ethical questions associated with potential future ”mind reading” technology.

	Yes	**No**
Does your project involve Human Embryonic Stem Cells?		✓
Does your project involve the use of human embryos?		✓
Does your project involve the use of human foetal tissues / cells?		✓
Does your project involve human participants?	✓	
Does your project involve human cells or tissues??		✓
Does your project involve personal data collection and/or processing?	✓	
Does it involve the collection and/or processing of sensitive personal data?	✓	
Does it involve processing of genetic information?		✓
Does it involve tracking or observation of participants?	✓	
Does your project involve further processing of previously collected personal data (secondary use)?	✓	
Does your project involve animals?		✓
Does your project involve developing countries?		✓
If your project involves low income countries, are any benefit-sharing actions planned?		✓
Could the situation in the country put the individuals taking part in the project at risk?		✓
Does your project involve the use of elements that may cause harm to the environment, animals or plants?		✓
Does your project deal with endangered fauna and/or flora/protected areas?		✓
Does your project involve the use of elements that may cause harm to humans, including project staff?		✓
Does your project involve other harmful materials or equipment?		✓
Does your project have the potential for military applications?	✓	
Does your project have an exclusive civilian application focus?	✓	
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		✓
Does your project affect current standards in military ethicss?		✓
Does your project have the potential for malevolent/criminal/terrorist abuse?		✓
Does your project involve information on the use of biological, chemical, nuclear security sensitive materials and explosives, and means of their delivery?		✓
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards?		✓
Does your project have the potential for terrorist or criminal abuse		✓
Will your project use or produce software for which there are copyright licensing implications?		✓
Will your project use or produce goods or information for which there are data protection, or other legal implications?		✓
Are there any other ethics issues that should be taken into consideration?	✓	

Table B.1: Ethics Checklist

Bibliography

[1] Khald Aboalayon, Miad Faezipour, Wafaa Almuhammadi, and Saeid Moslehpour. Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016. pages 7

[2] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, 100:270–278, 2018. pages 16

[3] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D. P Subha. Automated EEG-based screening of depression using deep convolutional neural network. *Computer Methods and Programs in Biomedicine*, 161:103–113, 2018. pages 16

[4] U. Rajendra Acharya, S. Vinitha Sree, G. Swapna, Roshan Joy Martis, and Jasjit S. Suri. Automated EEG analysis of epilepsy: A review. *Knowledge-Based Systems*, 45:147–165, 2013. pages 7

[5] Dimitrios A. Adamos, Stavros I. Dimitriadis, and Nikolaos A. Laskaris. Towards the bio-personalization of music recommendation systems: A single-sensor EEG biomarker of subjective music preference. *Information Sciences*, 343-344:94–108, 2016. pages 2, 9, 15, 27, 47

[6] Aaron Alexander-Bloch, Jay N. Giedd, and Ed Bullmore. Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience*, 14(5):322–336, 2013. pages 9

[7] Salma Alhagry, Aly Aly, and Reda A. Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017. pages 17

[8] Elena A. Allen, Erik B. Erhardt, Yonghua Wei, Tom Eichele, and Vince D. Calhoun. Capturing inter-subject variability with group independent component analysis of fMRI data: A simulation study. *NeuroImage*, 59(4):4141–4159, 2012. pages 9, 10

[9] Mohammad A. Almogbel, Anh H. Dang, and Wataru Kameyama. EEG-signals based cognitive workload detection of vehicle driver using deep learning. In

2018 20th International Conference on Advanced Communication Technology (ICACT), pages 256–259. IEEE, 2018. pages 16, 47

[10] E Altenmüller. Hits to the left, flops to the right: Different emotions during listening to music are reflected in cortical lateralisation patterns. *Neuropsychologia*, 40(13):2242–2256, 2002. pages 9

[11] Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019. pages 3, 16

[12] Jo Aoe, Ryohei Fukuma, Takufumi Yanagisawa, Tatsuya Harada, Masataka Tanaka, Maki Kobayashi, You Inoue, Shota Yamamoto, Yuichiro Ohnishi, and Haruhiko Kishima. Automatic diagnosis of neurological diseases using MEG signals with a deep neural network. *Scientific Reports*, 9(1):5057, 2019. pages 16, 18, 32

[13] P Aricò, G Borghini, G Di Flumeri, N Sciaraffa, and F Babiloni. Passive BCI beyond the lab: Current trends and future directions. *Physiological Measurement*, 39(8):08TR02, 2018. pages 7

[14] Martijn Arns, C. Keith Conners, and Helena C. Kraemer. A decade of EEG theta/beta ratio research in ADHD: A meta-analysis. *Journal of attention disorders*, 17(5):374–383, 2013. pages 7

[15] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5. IEEE, 2017. pages 18

[16] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271 [cs]*, 2018. pages 16

[17] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. *arXiv:1511.06448 [cs]*, 2015. pages 16, 19, 38, 46

[18] Thomas Baumgartner, Michaela Esslen, and Lutz Jäncke. From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International journal of psychophysiology*, 60(1):34–43, 2006. pages 9

[19] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012. pages 46

[20] Joydeep Bhattacharya and Hellmuth Petsche. Musicians and the gamma band: A secret affair? *NeuroReport*, 12(2):371–374, 2001. pages 9

[21] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in neuroinformatics*, 9:16, 2015. pages 8

[22] A. J. Blood and R. J. Zatorre. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences*, 98(20):11818–11823, 2001. pages 9

[23] Steven Brown, Michael J. Martinez, and Lawrence M. Parsons. Passive music listening spontaneously engages limbic and paralimbic systems. *Neuroreport*, 15(13):2033–2037, 2004. pages 9

[24] Camilo J. Cela-Conde, Luigi Agnati, Joseph P. Huston, Francisco Mora, and Marcos Nadal. The neural foundations of aesthetic appreciation. *Progress in Neurobiology*, 94(1):39–48, 2011. pages 9

[25] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. pages 30

[26] Wikimedia Commons. International_10-20_system_for_EEG-MCN.svg. https://commons.wikimedia.org/wiki/File:International_10-20_system_for_EEG-MCN.svg, 17. pages 6

[27] Wikimedia Commons. 21 electrodes of International 10-20 system for EEG.svg. https://commons.wikimedia.org/wiki/File:21_electrodes_of_International_10-20_system_for_EEG.svg, 2010. pages 6

[28] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017. pages 14

[29] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16(3):031001, 2019. pages 3, 16, 17, 19, 20, 45, 46, 47

[30] W. Dimpfel, F. Schober, and M. Spler. The influence of caffeine on human EEG under resting condition and during mental loads. *The Clinical Investigator*, 71(3), 1993. pages 21

[31] Hauke Dose, Jakob S. Møller, Helle K. Iversen, and Sadasivan Puthusserypady. An end-to-end deep learning approach to MI-EEG signal classification for BCIs. *Expert Systems with Applications*, 114:532–542, 2018. pages 16, 17, 32

[32] Emotive. Emotiv Epoc+ 14 Channel Mobile EEG. https://www.emotiv.com/product/emotiv-epoc-14-channel-mobile-eeg/, 2019. pages 2, 22

[33] Reza Fazel-Rezai, Brendan Z. Allison, Christoph Guger, Eric W. Sellers, Sonja C. Kleih, and Andrea Kübler. P300 brain computer interface: Current challenges and emerging trends. *Frontiers in Neuroengineering*, 5, 2012. pages 7

[34] Joseph T. Giacino, Joseph J. Fins, Steven Laureys, and Nicholas D. Schiff. Disorders of consciousness after acquired brain injury: The state of the science. *Nature Reviews Neurology*, 10(2):99, 2014. pages 7

[35] Alexandre Gramfort, Daniel Strohmeier, Jens Haueisen, Matti S. Hämäläinen, and Matthieu Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70:410–422, 2013. pages 8

[36] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE, 2013. pages 18

[37] Vipin Gupta, Mayur Dahyabhai Chopda, and Ram Bilas Pachori. Cross-Subject Emotion Recognition Using Flexible Analytic Wavelet Transform From EEG Signals. *IEEE Sensors Journal*, 19(6):2266–2274, 2019. pages 14

[38] S. K. Hadjidimitriou and L. J. Hadjileontiadis. EEG-Based Classification of Music Appraisal Responses Using Time-Frequency Analysis and Familiarity Ratings. *IEEE Transactions on Affective Computing*, 4(2):161–172, 2013. pages 9, 14, 25, 26, 27, 30, 43, 46

[39] Stelios K. Hadjidimitriou and Leontios J. Hadjileontiadis. Toward an EEG-Based Recognition of Music Liking Using Time-Frequency Analysis. *IEEE Transactions on Biomedical Engineering*, 59(12):3498–3510, 2012. pages 2, 14, 25, 26, 27

[40] S. Hagihira. Changes in the electroencephalogram during anaesthesia and their physiological basis. *British journal of anaesthesia*, 115(suppl_1):i27–i31, 2015. pages 7

[41] Mehdi Hajinoroozi, Zijing Mao, Tzyy-Ping Jung, Chin-Teng Lin, and Yufei Huang. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, 47:549–555, 2016. pages 16, 20

[42] Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding. In *2018 6th International Conference on Brain-Computer Interface (BCI)*, pages 1–6. IEEE, 2018. pages 16

[43] Md Musaddaqul Hasib, Tapsya Nayak, and Yufei Huang. A hierarchical LSTM model with attention for modeling EEG non-stationarity for human decision

prediction. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 104–107. IEEE, 2018. pages 17

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, 2015. pages 35, 41

[45] Ryan Hefron, Brett Borghetti, Christine Schubert Kabban, James Christensen, and Justin Estepp. Cross-Participant EEG-Based Assessment of Cognitive Workload Using Multi-Path Convolutional Recurrent Neural Networks. *Sensors*, 18(5):1339, 2018. pages 16, 17, 46

[46] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580 [cs]*, 2012. pages 32

[47] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. pages 11, 12

[48] Tomoyasu Horikawa and Yukiyasu Kamitani. Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features. *Frontiers in Computational Neuroscience*, 11, 2017. pages 16

[49] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*, 2016. pages 35, 36, 37, 45

[50] Patrick G. Hunter and E. Glenn Schellenberg. Music and Emotion. In Mari Riess Jones, Richard R. Fay, and Arthur N. Popper, editors, *Music Perception*, volume 36, pages 129–164. Springer New York, 2010. pages 9, 10

[51] Cosimo Ieracitano, Nadia Mammone, Alessia Bramanti, Amir Hussain, and Francesco C. Morabito. A Convolutional Neural Network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings. *Neurocomputing*, 323:96–107, 2019. pages 16, 18

[52] iMotions. *EEG Pocket Guide*. iMotions Biometric Research Platform, 2016. pages 5, 6

[53] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, 2015. pages 28

[54] Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, and Haojun Xu. Deep Convolutional Neural Networks for mental load classification based on EEG data. *Pattern Recognition*, 76:582–595, 2018. pages 16, 19, 20

[55] Viktor Jirsa and Viktor Müller. Cross-frequency coupling in real and virtual brain networks. *Frontiers in computational neuroscience*, 7:78, 2013. pages 9

[56] Kai Keng Ang, Zhang Yang Chin, Haihong Zhang, and Cuntai Guan. Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 2390–2397. IEEE, 2008. pages 14

[57] F.P. Kalaganis, D.A. Adamos, and N.A. Laskaris. Musical NeuroPicks: A consumer-grade BCI for on-demand music streaming services. *Neurocomputing*, 280:65–75, 2018. pages 15, 47

[58] Min-Ki Kim, Miyoung Kim, Eunmi Oh, and Sung-Phil Kim. A Review on the Computational Methods for Emotional State Estimation from the Human EEG. *Computational and Mathematical Methods in Medicine*, 2013:1–13, 2013. pages 8

[59] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. pages 30

[60] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. pages 32

[61] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and K.R. Rao. Cognitive Analysis of Working Memory Load from Eeg, by a Deep Recurrent Neural Network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2576–2580. IEEE, 2018. pages 17, 19

[62] Prathamesh M Kulkarni, Zhengdong Xiao, Eric J Robinson, Apoorva Sagarwal Jami, Jianping Zhang, Haocheng Zhou, Simon E Henin, Anli A Liu, Ricardo S Osorio, Jing Wang, and Zhe Chen. A deep learning approach for real-time detection of sleep spindles. *Journal of Neural Engineering*, 16(3):036004, 2019. pages 17, 46

[63] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. pages 16, 17, 32

[64] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov./1998. pages 11

[65] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. pages 28

[66] Jongpil Lee, Jiyoung Park, Keunhyoung Kim, and Juhan Nam. SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification. *Applied Sciences*, 8(1):150, 2018. pages 18

[67] Pierre-Majorique Leger and Rene Riedl. *Fundamentals of Neuro Information Systems Information Systems and the Brain*. Springer Berlin, 2015. OCLC: 890650508. pages 4, 8

[68] Jinpeng Li, Zhaoxiang Zhang, and Huiguang He. Implementation of EEG Emotion Recognition System Based on Hierarchical Convolutional Neural Networks. In Cheng-Lin Liu, Amir Hussain, Bin Luo, Kay Chen Tan, Yi Zeng, and Zhaoxiang Zhang, editors, *Advances in Brain Inspired Cognitive Systems*, volume 10023, pages 22–33. Springer International Publishing, 2016. pages 16

[69] Zhenqi Li, Xiang Tian, Lin Shu, Xiangmin Xu, and Bin Hu. Emotion Recognition from EEG Using RASM and LSTM. In Benoit Huet, Liqiang Nie, and Richang Hong, editors, *Internet Multimedia Computing and Service*, volume 819, pages 310–318. Springer Singapore, 2018. pages 17

[70] Yuan-Pin Lin, Chi-Hong Wang, Tien-Lin Wu, Shyh-Kang Jeng, and Jyh-Horng Chen. EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 489–492. IEEE, 2009. pages 14, 27

[71] F Lotte, L Bougrain, A Cichocki, M Clerc, M Congedo, A Rakotomamonjy, and F Yger. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. pages 8, 14, 47

[72] Fabien Lotte. A Tutorial on EEG Signal-processing Techniques for Mental-state Recognition in Brain–Computer Interfaces. In Eduardo Reck Miranda and Julien Castet, editors, *Guide to Brain-Computer Music Interfacing*, pages 133–161. Springer London, 2014. pages 8

[73] Plonsey Malmivuo, Jaakko Malmivuo, and Robert Plonsey. *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, USA, 1995. pages 4

[74] Hector P. Martinez, Yoshua Bengio, and Georgios N. Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33, 2013. pages 10

[75] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2):314–324, 2012. pages 22

[76] Jinyoung Moon. Extraction of User Preference for Video Stimuli Using EEG-Based User Responses. *ETRI Journal*, 35(6):1105–1114, 2013. pages 14

[77] Murugappan Murugappan, Nagarajan Ramachandran, and Yaacob Sazali. Classification of human emotion from EEG using discrete wavelet transform.

Journal of Biomedical Science and Engineering, 03(04):390–396, 2010. pages 14

[78] Sebastian Nagel and Martin Spüler. World's Fastest Brain-Computer Interface: Combining EEG2Code with Deep Learning. Preprint, Neuroscience, 2019. pages 16, 17, 32

[79] Satoshi Nakamura, Norihiro Sadato, Tsutomu Oohashi, Emi Nishina, Yoshitaka Fuwamoto, and Yoshiharu Yonekura. Analysis of music–brain interaction with simultaneous measurement of regional cerebral blood flow and electroencephalogram beta rhythm in human subjects. *Neuroscience Letters*, 275(3):222–226, 1999. pages 9

[80] Neuralink. Neuralink. https://www.neuralink.com/, 2019. pages 3

[81] Ozan Ozdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdogmus. Adversarial Deep Learning in EEG Biometrics. *IEEE Signal Processing Letters*, 26(5):710–714, 2019. pages 16

[82] Adam Page, Colin Shea, and Tinoosh Mohsenin. Wearable seizure detection using convolutional neural networks with transfer learning. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1086–1089. IEEE, 2016. pages 16, 48

[83] Yaozhang Pan, Cuntai Guan, Juanhong Yu, Kai Keng Ang, and Ti Eu Chan. Common frequency pattern for music preference identification using frontal EEG. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 505–508. IEEE, 2013. pages 9, 14, 27, 30

[84] G. Pfurtscheller and C. Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001. pages 7

[85] Sergey M. Plis, Devon R. Hjelm, Ruslan Salakhutdinov, Elena A. Allen, Henry J. Bockholt, Jeffrey D. Long, Hans J. Johnson, Jane S. Paulsen, Jessica A. Turner, and Vince D. Calhoun. Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience*, 8, 2014. pages 20, 32

[86] Rui Qiao, Chunmei Qing, Tong Zhang, Xiaofen Xing, and Xiangmin Xu. A novel deep-learning based framework for multi-subject emotion recognition. In *2017 4th International Conference on Information, Cybernetics and Computational Social Systems (ICCSS)*, pages 181–185. IEEE, 2017. pages 16, 18, 39, 47

[87] L R Quitadamo, F Cavrini, L Sbernini, F Riillo, L Bianchi, S Seri, and G Saggio. Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: A review. *Journal of Neural Engineering*, 14(1):011001, 2017. pages 14

[88] Rajesh P N Rao. *Brain-Computer Interfacing: An Introduction*. pages 7, 9

[89] Yuanfang Ren and Yan Wu. Convolutional deep belief networks for feature extraction of EEG signal. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2850–2853. IEEE, 2014. pages 20

[90] Google Research. AudioSet. https://research.google.com/audioset/. pages 48

[91] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: A systematic review. *arXiv:1901.05498 [cs, eess, stat]*, 2019. pages 3, 16, 17, 25, 28, 30, 45, 47, 48

[92] Giulio Ruffini, David Ibañez, Marta Castellano, Laura Dubreuil, Jean-François Gagnon, Jacques Montplaisir, and Aureli Soria-Frisch. Deep learning with EEG spectrograms in rapid eye movement behavior disorder. Preprint, Neuroscience, 2018. pages 16, 18

[93] Elham S. Salama, Reda A.El-Khoribi, Mahmoud E.Shoman, and Mohamed A.Wahby. EEG-Based Emotion Recognition using 3D Convolutional Neural Networks. *International Journal of Advanced Computer Science and Applications*, 9(8), 2018. pages 16, 18

[94] V. N. Salimpoor, I. van den Bosch, N. Kovacevic, A. R. McIntosh, A. Dagher, and R. J. Zatorre. Interactions Between the Nucleus Accumbens and Auditory Cortices Predict Music Reward Value. *Science*, 340(6129):216–219, 2013. pages 9, 46

[95] Daniela Sammler, Maren Grigutsch, Thomas Fritz, and Stefan Koelsch. Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2):293–304, 2007. pages 9

[96] R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–7. IEEE, 2017. pages 16, 45

[97] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization: Convolutional Neural Networks in EEG Analysis. *Human Brain Mapping*, 38(11):5391–5420, 2017. pages 16, 17, 24, 28, 32, 41, 45

[98] Barbara Schmidt and Simon Hanslmayr. Resting frontal EEG alpha-asymmetry predicts the evaluation of affective musical stimuli. *Neuroscience Letters*, 460(3):237–240, 2009. pages 9

[99] Jonathon Shlens. A Tutorial on Independent Component Analysis. *arXiv:1404.2986 [cs, stat]*, 2014. pages 8, 23

[100] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. pages 32

[101] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012. pages 46

[102] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114, 2018. pages 16

[103] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. pages 28

[104] Avital Sternin, Sebastian Stober, Adrian M Owen, and Jessica A Grahn. Tempo Estimation from the EEG signal during perception and imagination of music. page 8. pages 17

[105] Sebastian Stober. Toward Studying Music Cognition with Information Retrieval Techniques: Lessons Learned from the OpenMIIR Initiative. *Frontiers in Psychology*, 8:1255, 2017. pages 17

[106] Sebastian Stober, Daniel J. Cameron, and Jessica A. Grahn. Using Neural Convolutional Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings. *NIPS*, 2014. pages 17

[107] Sebastian Stober, Avital Sternin, Adrian M. Owen, and Jessica A. Grahn. Deep Feature Learning for EEG Recordings. *arXiv:1511.04306 [cs]*, 2015. pages 17

[108] Yingnan Sun, Frank P.-W. Lo, and Benny Lo. EEG-based user identification system using 1D-convolutional long short-term memory neural networks. *Expert Systems with Applications*, 125:259–267, 2019. pages 17, 18

[109] Zhichuan Tang, Chao Li, and Shouqian Sun. Single-trial EEG classification of motor imagery using deep convolutional neural networks. *Optik*, 130:11–18, 2017. pages 17

[110] Kaggle Team. Grasp-and-Lift EEG Detection Winners' Interview: 3rd place, Team HEDJ. http://blog.kaggle.com/2015/10/05/grasp-and-lift-eeg-detection-winners-interview-3rd-place-team-hedj/, 2015. pages 3, 7

[111] Jason Teo, Chew Lin Hou, and James Mountstephens. Deep learning for EEG-Based preference classification. *AIP Conference Proceedings*, 1891(020141):8, 2017. pages 16

[112] K. C. Tseng, Bor-Shyh Lin, Chang-Mu Han, and Psi-Shi Wang. Emotion recognition of EEG underlying favourite music by support vector machine. In *2013 1st International Conference on Orange Technologies (ICOT)*, pages 155–158. IEEE, 2013. pages 14

[113] Orestis Tsinalis, Paul M. Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. *arXiv:1610.01683 [cs, stat]*, 2016. pages 16, 17

[114] Jan van Erp, Fabien Lotte, and Michael Tangermann. Brain-Computer Interfaces: Beyond Medical Applications. *Computer*, 45(4):26–34, 2012. pages 7

[115] Michel J. A. M. van Putten, Sebastian Olbrich, and Martijn Arns. Predicting sex from brain rhythms with deep learning. *Scientific Reports*, 8(1):3069, 2018. pages 16

[116] Gyanendra K. Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102:162–172, 2014. pages 14

[117] Fang Wang, Sheng-hua Zhong, Jianfeng Peng, Jianmin Jiang, and Yan Liu. Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks. In Klaus Schoeffmann, Thanarat H. Chalidabhongse, Chong Wah Ngo, Supavadee Aramvith, Noel E. O'Connor, Yo-Sung Ho, Moncef Gabbouj, and Ahmed Elgammal, editors, *MultiMedia Modeling*, volume 10705, pages 82–93. Springer International Publishing, 2018. pages 45

[118] Xiaoyan Wei, Lin Zhou, Ziyi Chen, Liangjun Zhang, and Yi Zhou. Automatic seizure detection using three-dimensional CNN based on multi-channel EEG. *BMC Medical Informatics and Decision Making*, 18(S5):111, 2018. pages 18

[119] R. W. Wilkins, D. A. Hodges, P. J. Laurienti, M. Steen, and J. H. Burdette. Network Science and the Effects of Music Preference on Functional Brain Connectivity: From Beethoven to Eminem. *Scientific Reports*, 4(1):6130, 2015. pages 9

[120] Haiyan Xu and Konstantinos N. Plataniotis. Affective states classification using EEG and semi-supervised deep learning approaches. In *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2016. pages 19

[121] Miku Yanagimoto and Chika Sugimoto. Recognition of persisting emotional valence from EEG using convolutional neural networks. In *2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 27–32. IEEE, 2016. pages 16

[122] Yi Yu, Samuel Beuret, Donghuo Zeng, and Keizo Oyama. Deep Learning of Human Perception in Audio Event Classification. In *2018 IEEE International*

Symposium on Multimedia (ISM), pages 188–189. IEEE, 2018. pages 17, 35, 45

[123] Thorsten O Zander and Christian Kothe. Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general. *Journal of Neural Engineering*, 8(2):025005, 2011. pages 7

[124] R. J. Zatorre and V. N. Salimpoor. From perception to pleasure: Music and its neural substrates. *Proceedings of the National Academy of Sciences*, 110(Supplement_2):10430–10437, 2013. pages 9

[125] Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82:128–150, 2017. pages 16

[126] Xiang Zhang, Lina Yao, Xianzhi Wang, Jessica Monaghan, David Mcalpine, and Yu Zhang. A Survey on Deep Learning based Brain Computer Interface: Recent Advances and New Frontiers. *arXiv:1905.04149 [cs, eess, q-bio]*, 2019. pages 3, 12, 13, 47

[127] Ziping Zhao, Zhongtian Bao, Yiqin Zhao, Zixing Zhang, Nicholas Cummins, Zhao Ren, and Bjorn Schuller. Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. *IEEE Access*, 7:97515–97525, 2019. pages 18

[128] Wei-Long Zheng, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. EEG-based emotion classification using deep belief networks. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014. pages 19