National Technical University of Athens
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

Affective Analysis and Interpretation of Brain Responses to Music Stimuli

Diploma Thesis

Avramidis Kleanthis

Supervisor: Prof. Petros Maragos

Athens, July 2021

National Technical University of Athens
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

Affective Analysis and Interpretation of Brain Responses to Music Stimuli

Diploma Thesis

Avramidis Kleanthis

Supervisor: Prof. Petros Maragos

Approved by the examining committee:

.....	.....	.....
Petros Maragos	Alexandros Potamianos	Constantinos Tzafestas
Professor	Associate Professor	Associate Professor
NTUA	NTUA	NTUA

Athens, July 2021

.....
Kleanthis Avramidis
Electrical and Computer Engineer, NTUA

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας κλείνει ένα σημαντικό προσωπικό κεφάλαιο, αυτό της φοίτησης στη Σχολή Ηλεκτρολόγων Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου. Ήταν τα πιο σύντομα 6 χρόνια της ζωής μου, γεμάτα με νέες γνώσεις και εμπειρίες. Με αυτή την αφορμή, νιώθω την υποχρέωση να ευχαριστήσω:

Τον επιβλέποντα καθηγητή κ. Πέτρο Μαραγκό και όλα τα μέλη-φίλους του Εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σήματος για την ευκαιρία και τα εφόδια που μου έδωσαν για να εκκινήσω την ερευνητική μου πορεία στο πλαίσιο της διπλωματικής αλλά και της γενικότερης συνεργασίας μας, εδώ και περίπου 2 χρόνια.

Ιδιαιτέρως τη μεταδιδακτορική ερευνήτρια Δρ. Νάνσυ Ζλαντίντση και το διδακτορικό φοιτητή Χρήστο Γαρούφη, στα πρόσωπα των οποίων βρήκα δύο πολύ καλούς φίλους. Παιδιά, σας ευχαριστώ για την καίρια συμβολή σας σε όλη τη διαδικασία εκπόνησης της εργασίας αλλά και για την αμέριστη στήριξη τα 2 χρόνια της συνεργασίας μας, είστε οι καλύτεροι.

Τον Tim Greer, το Νίκο Αθανασίου και όλους όσους βοήθησαν με τις ιδέες τους και το υλικό τους στο να πάρει η παρούσα εργασία την τελική της μορφή, καθώς και όλους όσους, ανεβάζοντας τη δουλειά τους στο διαδίκτυο, προωθούν την ελεύθερη διακίνηση της γνώσης.

Την οικογένειά μου, τους φίλους που έκανα στη σχολή και με στήριξαν όλο αυτό τον καιρό, ιδίως όμως τους Παναγιώτη, Μανώλη, Χρήστο, Νικόλα και Χάρη, που ανέχτηκαν την εξαετή μου τρέλα που ονομάζεται ΣΗΜΜΥ και παραμένουν δίπλα μου, πάντα.

Κλεάνθης Αβραμίδης
12/07/2021

Dedicated to those 32 people who offered the EEG data I utilized.
Guys, you couldn't have made it any more difficult.

Abstract

The analysis of human emotions is a widely researched topic in the scientific fields of Psychology and Neuroscience, trying to investigate the nature and elicitation mechanisms of our feelings. From a computational perspective, however, it remains rather underexplored. While Artificial Intelligence has made overwhelming progress in modeling rational intelligence, there are yet no highly reliable systems to analyze affect, as considerable barriers exist in this process: Emotion expression can be highly subjective and its interpretation varies depending on the context, whereas it poses an inter-subject variability. Yet, most Signal Processing and Machine Learning studies concentrate on behavioral processing of emotions, through modalities like speech, text and facial expressions. To address the challenges of Affective Analysis, in this thesis we choose to process brain signals, and specifically the Electroencephalogram (EEG), as a means to derive emotional information. Recorded physiological and neural signals are capable of being more objective and reliable affective indicators, whereas they can also contribute to develop human-aid systems for applications like the treatment or rehabilitation from brain diseases. Importantly, we consider music as the means to induce emotions for the EEG recordings, since music is known to have a deep emotional impact on humans.

Our approach can be divided into two main parts: In the first one, we analyze the complex structure of the EEG and examine novel feature extraction schemes that are based on two multifractal algorithms, namely Multiscale Fractal Dimension and Multifractal Detrended Fluctuation Analysis. In this way we attempt to quantify the variability of the observed signals' complexity across multiple timescales. Our proposed EEG features surpass widely used baselines on Emotion Recognition, whereas they show competitive results in challenging subject-independent experiments and recognition of arousal, indicating that it is highly correlated with the EEG's fragmented structure. In the second part, we utilize a two-branch neural network as a bimodal EEG-music framework, which learns common latent representations between the EEG signals and their music stimuli in order to examine their correspondence. Through this model, we perform supervised emotion recognition experiments and retrieval of music rankings to EEG input queries. By applying this system to independent subject data, we also extract interesting patterns regarding the latent similarity of brain and music signals, the temporal variation of the music-induced emotions and the activated brain regions in each case. As a whole, this study deals with core problems regarding the interpretation of complex EEG signals and illustrates multiple ways that music stimulates the brain activity.

— ***Keywords:*** Music Perception, Music Cognition, Emotion Recognition, Electroencephalography, Multifractal Analysis, Cross-Modal Learning, Metric Learning

Εκτεταμένη Περίληψη

Η ανάλυση των ανθρώπινων συναισθημάτων είναι ένα δημοφιλές ερευνητικό πεδίο της Ψυχολογίας και των Νευροεπιστημών, που προσπαθούν να διερενήσουν τη φύση τους και τους μηχανισμούς παραγωγής τους. Από υπολογιστική άποψη, ωστόσο, παραμένει ένα περισσότερο ανεξερεύνητο πεδίο. Ενώ η Τεχνητή Νοημοσύνη έχει κάνει μεγάλη πρόοδο στην ανάλυση της ανθρώπινης λογικής, ακόμα δεν έχουμε στη διάθεσή μας έμπιστα ευφυή συστήματα για την αναγνώριση συναισθημάτων, λόγω μιας σειράς σημαντικών εμποδίων: Η συναισθηματική έκφραση έχει ενίοτε υποκειμενική ερμηνεία, ενώ συνήθως διαφορετικά άτομα αποδίδουν διαφορετικά συναισθήματα ή και διαφορετικές εκφράσεις του ίδιου συναισθήματος. Πάντως, η σχετική έρευνα στους τομείς της Επεξεργασίας Σήματος και της Μηχανικής Μάθησης επικεντρώνεται στην συμπεριφορική ανάλυση του συναισθήματος, χρησιμοποιώντας τον προφορικό ή γραπτό λόγο και εκφράσεις του προσώπου. Για να αντιμετωπίσουμε τις προκλήσεις που θέτει η Ανάλυση Συναισθήματος, στην εργασία αυτή επιλέγουμε να αναλύσουμε νευρολογικά σήματα, συγκεκριμένα το Ηλεκτροεγκεφαλογράφημα (ΗΕΓ) ως μέσο εξαγωγής συναισθηματικής πληροφορίας. Τα φυσιολογικά και νευρολογικά σήματα μπορούν να είναι πιο αντικειμενικοί δείκτες του συναισθήματος ενώ η ανάλυσή τους μπορεί να συνεισφέρει στην ανάπτυξη ευφυών συστημάτων για την υποβοήθηση του ανθρώπου και στην κατανόηση της διαδικασίας λήψης αποφάσεων. Για την πρόκληση των συναισθημάτων θεωρούμε μουσικά σήματα, καθώς η μουσική είναι γνωστή για την ισχυρή της συναισθηματική επίδραση στους ανθρώπους και στη λειτουργία του ανθρώπινου εγκεφάλου.

Η προσέγγισή μας μπορεί να διαιρεθεί σε δύο βασικά μέρη: Στο πρώτο, μελετούμε την πολύπλοκη δομή των ΗΕΓ σημάτων και εξετάζουμε καινοτόμες μεθόδους εξαγωγής χαρακτηριστικών, βασισμένες σε δύο multifractal αλγορίθμους, τους Multiscale Fractal Dimension και Multifractal Detrended Fluctuation Analysis. Επιχειρούμε έτσι να ποσοτικοποιήσουμε πόσο μεταβάλλεται η πολυπλοκότητα αυτών των σημάτων σε διαφορετικές κλίμακες παρατήρησης. Τα προτεινόμενα χαρακτηριστικά πετυχαίνουν μεγαλύτερη ακρίβεια από ευρέως χρησιμοποιούμενες μεθόδους στην Αναγνώριση Συναισθήματος, ενώ αποδίδουν σημαντικά αποτελέσματα σε πειράματα πολλών συμμετεχόντων και στην αναγνώριση της μετρικής arousal, υποδεικνύοντας έτσι πως σχετίζεται σε μεγάλο βαθμό με την περίπλοκη δομή του ΗΕΓ. Στο δεύτερο μέρος κατασκευάζουμε ένα πολυτροπικό δίκτυο για ΗΕΓ και μουσικά σήματα προκειμένου να αναλύσουμε την συσχέτισή τους όσον αφορά το συναίσθημα και να εξάγουμε κοινές αναπαραστάσεις. Μέσω αυτού του μοντέλου αξιολογούμε πειράματα κατηγοριοποίησης για επισημειώσεις συναισθήματος, αλλά και πειράματα εξαγωγής μουσικών κομματιών, σχετικών με ΗΕΓ εισόδους. Εφαρμόζοντας αυτό το σύστημα ξεχωριστά σε δεδομένα διαφορετικών ανθρώπων, εξάγουμε σημαντικά μοτίβα σχετικά με την ομοιότητα εγκεφαλικών και μουσικών σημάτων, τις χρονικές μεταβολές των συναισθηματικών εκφράσεων και τις εγκεφαλικές περιοχές που ενεργοποιούνται ανά περίπτωση. Συνολικά, η εν λόγω εργασία καταπιάνεται με θεμελιώδη ζητήματα σχετικά με την κατανόηση των πολύπλοκων σημάτων ΗΕΓ και αποτυπώνει πολλαπλούς τρόπους με τους οποίους η μουσική επηρεάζει την ανθρώπινη εγκεφαλική λειτουργία.

Εισαγωγή

Η Επιστήμη των Συναισθημάτων

Στην καθημερινή μας ζωή βιώνουμε μια ποικιλομορφία από διαφορετικά συναισθήματα ή συνδυασμούς συναισθημάτων (χαρά, λύπη, απογοήτευση, έκπληξη κλπ). Ένα συναίσθημα είναι μια υποκειμενική κατάσταση στην οποία βρισκόμαστε και την αντιλαμβανόμαστε από το πώς *νιώθουμε*, από φυσιολογική και ψυχολογική άποψη. Στην πάροδο του χρόνου, πολλές θεωρίες έχουν προταθεί για να εξηγήσουν αυτό το φαινόμενο. Επί παραδείγματι, η θεωρία των *James-Lange* [70] ορίζει πως τα συναισθήματα προκύπτουν από βιολογικές "εξάρσεις" του ανθρώπινου οργανισμού. Άλλες θεωρίες [20] προτάσσουν την ανεξαρτησία της φυσιολογικής με την ψυχολογική αντίδραση. Σε κάθε περίπτωση, φαίνεται πως η φυσιολογία παίζει σημαντικό ρόλο είτε στη δημιουργία είτε στην ενδυνάμωση του συναισθήματος.

Οι Νευροεπιστήμες και η Ψυχολογία έχουν αναπτύξει 2 βασικές προσεγγίσεις για την κατηγοριοποίηση των συναισθημάτων: την διανυσματική και την κατηγορική. Χαρακτηριστικό παράδειγμα της πρώτης κατηγορίας, που θα χρησιμοποιήσουμε σε αυτή την εργασία, είναι το πρωτόκολλο *Valence-Arousal* του Russel [128], σύμφωνα με το οποίο κάθε συναίσθημα ορίζεται ως ένα σημείο στο διδιάστατο χώρο με διευθύνσεις αρνητικό-θετικό (valence) και άτονο-έντονο (arousal). Η κατηγορική προσέγγιση, από την άλλη, χρησιμοποιεί διακριτές κλάσεις συναισθημάτων. Μια από τις πρώτες και πιο ευρέως χρησιμοποιούμενες κατηγοριοποιήσεις μέχρι και σήμερα είναι αυτή του Paul Ekman (1970), που εισήγαγε έξι καθολικά αποδεκτά συναισθήματα [40]: χαρά, απέχθεια, φόβο, θυμό, έκπληξη και λύπη.

Στο πλαίσιο αυτό, στόχος μας είναι να χτίσουμε υπολογιστικά μοντέλα που να είναι σε θέση να διερευνήσουν και να αναγνωρίσουν το συναίσθημα μέσα από μια δεδομένη κατάσταση. Με τον όρο *Affective Computing* αναφερόμαστε στην μελέτη αυτή των συστημάτων και συσκευών που επεξεργάζονται, αναγνωρίζουν και κατανοούν τα ανθρώπινα συναισθήματα, με σκοπό τη βελτίωση των παρεχόμενων υπηρεσιών αλλά και την ανάλυση της ανθρώπινης ψυχολογίας. Για να το επιτύχουμε αυτό, χρειαζόμαστε θεωρητικά εργαλεία, που θα αναλύσουμε στη συνέχεια, και μεγάλες συλλογές δεδομένων, στις οποίες βρίσκουμε κατάλληλους περιγραφητές. Οι πιο γνωστές μορφές δεδομένων για αυτό το σκοπό είναι:

- **Εκφράσεις Προσώπου**: Η πιο άμεση οδός έκφρασης των συναισθημάτων, αφού αποτελεί το κύριο μέσο επικοινωνίας και έκφρασης. Συστήματα ανάλυσης ανθρώπινων εκφράσεων υπάρχουν αμέτρητα και διακρίνονται σε στατικά (εικόνες) και δυναμικά (βίντεο), ενώ δρουν σε περιγραφητές όπως η θέση των στοιχείων του προσώπου και η κίνησή τους. Η πρόοδος στον όγκο των δεδομένων και τους αλγορίθμους Τεχνητής Νοημοσύνης (Artificial Intelligence - AI) που έχουμε στη διάθεσή μας έχουν ωθήσει την έρευνα στο συγκεκριμένο πεδίο, αλλά και τον σκεπτικισμό σχετικά με το ζήτημα της καθολικότητας και της συνέπειας [11] των συναισθηματικών εκφράσεων.

- **Φυσική Γλώσσα**: Ο πυρήνας της ανθρώπινης επικοινωνίας βασίζεται στην φυσική γλώσσα, γραπτή ή προφορική. Η αναγνώριση του συναισθήματος που εκφράζεται μέσω του λόγου χρησιμοποιείται σε συστήματα επικοινωνίας ανθρώπου-υπολογιστή, σε συστήματα συστάσεων ενώ υποβοηθούν και επιστήμες όπως η Ψυχολογία. Η σχετική έρευνα έχει ιστορία τριών δεκαετιών αλλά ακόμα δεν έχει αποδώσει απόλυτα έμπιστα συστήματα, μιας και ένα σήμα ομιλίας περιέχει ποικίλες μεταβλητές (ποιος μιλάει, σε ποια γλώσσα, με τι λεξιλόγιο κλπ). Από την άλλη, η γραπτή γλώσσα ακολουθεί άλλο δρόμο επεξεργασίας που συμπεριλαμβάνει γραμματικούς και συντακτικούς κανόνες στην ανάλυσή της. Προσφάτως, η Μηχανική Μάθηση, δρώντας πάνω

σε τεράστιους όγκους λεξιλογικών δεδομένων και ειδικά διαμορφωμένα αναδρομικά μοντέλα, έχει καταφέρει να ξεπεράσει προηγούμενες προσεγγίσεις.

- **Ήχος και Μουσική**: Ενώ υιοθετεί αρκετά στοιχεία από την ανάλυση ομιλίας, η αναγνώριση συναισθήματος από μουσικά σήματα είναι αυτόνομος κλάδος με μεγάλη ανάπτυξη τα τελευταία χρόνια. Συνιστά μια σημαντική ενότητα της Επεξεργασίας Μουσικής Πληροφορίας αφού, όπως θα δούμε, η μουσική μπορεί να προκαλέσει πολύ έντονα συναισθήματα σε σύγκριση με άλλα φαινόμενα. Είναι κοινώς αποδεκτό πως πολλά μουσικά στοιχεία ενέχουν συναισθηματικές συνιστώσες, για παράδειγμα οι ελάσσονες κλίμακες δημιουργούν συναισθήματα λύπης ή και φόβου. Ο συγκεκριμένος κλάδος βρίσκει εφαρμογή κυρίως σε συστήματα μουσικών συστάσεων, αλλά μπορεί να βοηθήσει και στην ανάλυση των ψυχολογικών αντιδράσεων του ανθρώπου.

- **Βιομετρικά Σήματα**: Όλες οι προαναφερθείσες περιπτώσεις χρησιμοποιούν σήματα συμπεριφοράς για να προσδιορίσουν το συναίσθημα. Αυτό ωστόσο δεν είναι πάντα ακριβές, μιας και η έκφραση και η αντίληψη της συναισθηματικής κατάστασης είναι υποκειμενική, ενώ δεν είναι ιδιαίτερα δύσκολο για έναν άνθρωπο να προσποιηθεί ή να κρύψει ένα πραγματικό συναίσθημα. Για το λόγο αυτό, τα τελευταία χρόνια ενθαρρύνεται η διερεύνηση αντίστοιχων περιγραφητών σε βιοσήματα, που εξελίσσονται χωρίς τον ενσυνείδητο έλεγχό μας και άρα μπορούν να είναι πιο αξιόπιστα (θερμοκρασία, εγκεφαλικά σήματα κλπ) [139]. Η έρευνα εδώ χρησιμοποιεί κυρίως στατιστικά εργαλεία για να μοντελοποιήσει τις ιδιότητες των βιοσημάτων, τα οποία πάσχουν συνήθως από υπερβολική παρουσία θορύβου και απαιτούν δαπανηρή διαδικασία καταγραφής.

Ο Ανθρώπινος Εγκέφαλος

Η έκφραση και αντίληψη συναισθημάτων είναι μια προηγμένη λειτουργία του ανθρώπινου εγκεφάλου, του πιο πολύπλοκου οργάνου στο ανθρώπινο σώμα, υπεύθυνο για τις αισθήσεις, την κίνηση και τη συμπεριφορά μας, που παρά τις μακραίωνες προσπάθειες του ανθρώπου, παραμένει ακόμα και σήμερα ένα μυστήριο. Ερευνητικά ακουμπά κυρίως την περιοχή των Νευροεπιστημών. Σήμερα ωστόσο, με τα πανίσχυρα υπολογιστικά εργαλεία και αλγορίθμους που έχουμε στη διάθεσή μας, είμαστε σε θέση να αναλύουμε πολλά περισσότερα δεδομένα και συνεπώς δικαιούμαστε να αναμένουμε σημαντικές εξελίξεις στο εγγύς μέλλον από τον χώρο της Τεχνητής Νοημοσύνης και της Επιστήμης Υπολογιστών.

Όπως όλα τα ανθρώπινα όργανα, ο εγκέφαλος αποτελείται από κύτταρα που ρυθμίζουν τη δομή και τη λειτουργία του. Κάποια εξ' αυτών ωστόσο, τα νευρικά κύτταρα, δουλεύουν ώστε να μπορούμε να αισθανόμαστε, να σκεφτόμαστε και να δρούμε, μέσω των πληροφοριών που διακινούν. Ένας ανθρώπινος εγκέφαλος αποτελείται από τουλάχιστον 90 δισ. νευρώνες, καθένας εκ των οποίων συνδέεται με ηλεκτροχημικές διεργασίες με χιλιάδες άλλους, φτιάχοντας ένα υπέρμετρα πολύπλοκο πλέγμα. Το νευρικό κύτταρο είναι η στοιχειώδης δομή του νευρικού συστήματος. Αποτελείται από το σώμα του νευρώνα, έναν άξονα και ένα σύνολο από δενδρίτες, μέσω των οποίων συνδέεται με δενδρίτες άλλων νευρώνων (συνάψεις), σχηματίζοντας δίκτυα μεταφοράς πληροφορίας, μέσω ηλεκτροχημικών σημάτων.

Την πρώτη συστηματική καταγραφή των ηλεκτρικών αυτών σημάτων έκανε ο Richard Caton [21] το 1875, όταν τοποθέτησε 2 ηλεκτρόδια στο εξωτερικό του κρανίου ενός ανθρώπου και μέτρησε την ηλεκτρική του δραστηριότητα. Από τότε, ο όρος "Ηλεκτροεγκεφαλογράφημα" (ΗΕΓ) χρησιμοποιείται ευρέως για να δηλώσει την ηλεκτρική δραστηριότητα του εγκεφάλου, η γνώση της οποίας έχει αποδειχθεί κρίσιμη σε πολλούς ιατρικούς και άλλους τομείς. Σήμερα, το ΗΕΓ καταγράφεται ψηφιακά, με πολλά εξειδικευμένα ηλεκτρόδια

και ακολουθώντας διεθνή standards τοποθέτησης και εξαγωγής. Το ηλεκτρικό σήμα που καταγράφεται είναι αρκετά ασθενές, ενώ μέχρι να γίνουν αντιληπτά τα αντίστοιχα πεδία από τα ηλεκτρόδια, διαπερνώνται από ισχυρό θόρυβο μέσα στο κρανίο. Συνεπώς, χρησιμοποιούνται τεχνικές αποθορυβοποίησης και ενίσχυσής τους πριν την οποιαδήποτε επεξεργασία. Ένα από τα πλέον σημαντικά χαρακτηριστικά της εγκεφαλικής ηλεκτρικής δραστηριότητας είναι οι συχνότητες ή ρυθμοί που εμφανίζουν και είναι δείκτες μιας πληθώρας ανθρώπινων λειτουργιών. Έχουν προσδιοριστεί 5 βασικοί εγκεφαλικοί ρυθμοί:

- *Δέλτα* (δ): Βρίσκεται στα 0.5-4Hz, σχετίζεται κυρίως με το βαθύ ύπνο.
- *Θήτα* (ϑ): Βρίσκεται στα 4-7.5Hz και σχετίζεται με λειτουργίες του ασυνείδητου και την έμπνευση, ενώ εξετάζεται ιδιαίτερα κατά την πρώιμη παιδική ηλικία.
- *Άλφα* (α): Βρίσκεται στα 8-13Hz και αποτελεί ίσως τον πιο σημαντικό και πιο συχνά παρατηρούμενο ρυθμό, ενώ προσδιορίστηκε πρώτος, το 1929. Υποδεικνύει τη χαλάρωση, την αποφόρτιση ή και την ένταση, την προσοχή κλπ.
- *Βήτα* (β): Βρίσκεται στα 13-30Hz. Είναι ο κύριος ρυθμός που σχετίζεται με την ενεργή δραστηριότητα, τη σκέψη και την προσοχή, ενώ απαντάται κυρίως σε ενήλικες.
- *Γάμμα* (γ): Πάνω από 30Hz, έχει αξιοποιηθεί για τη διάγνωση εγκεφαλοπαθειών.

Η Αντίληψη της Μουσικής

Η ικανότητά μας να ακούμε και να επεξεργαζόμαστε ήχους οφείλεται στο ακουστικό μας σύστημα, αποτελούμενο κυρίως από τις λειτουργίες του αυτιού και του εγκεφάλου. Τα ηχητικά κύματα εισέρχονται στο αυτί και δονούν τη λεγόμενη τυμπανική μεμβράνη, η οποία με τη σειρά της μεταφέρει και ενισχύει αυτές τις δονήσεις στο υγρό του εσωτερικού αυτιού όπου το σήμα αναλύεται και μετατρέπεται σε ηλεκτρικές ώσεις προς το ακουστικό τμήμα του εγκεφάλου. Ειδικά στην περίπτωση της μουσικής, πολλές διαφορετικές εγκεφαλικές περιοχές φαίνεται να υπεισέρχονται στην ανάλυση των συνιστωσών του μουσικού σήματος. Η κατηγοριοποίηση αυτή γίνεται με βάση θεμελιώδη χαρακτηριστικά του ήχου, όπως η ένταση, η συχνότητα, η διάρκεια, ο ρυθμός, το ηχόχρωμα κλπ. Ο ανθρώπινος εγκέφαλος επεξεργάζεται και οργανώνει αυτές τις πληροφορίες σε πιο πολύπλοκες έννοιες, όπως το μέτρο, η μελωδία και η αρμονία. Πρόκειται συνεπώς για σημαντικά στοιχεία και αμέτρητες έρευνες έχουν αναλωθεί στην εύρεση περιγραφητών που να τα προσδιορίζουν σε ηχητικά δεδομένα.

Όπως είπαμε, διαφορετικές εγκεφαλικές περιοχές αναλύουν διαφορετικά μουσικά στοιχεία, π.χ. ο ακουστικός λοβός τους τόνους, η παρεγκεφαλίδα το ρυθμό κ.ο.κ. Το πιο μυστήριο χαρακτηριστικό από αυτά είναι σαφώς το συναίσθημα που ενέχει ή προκαλεί ένα μουσικό κομμάτι. Από την εμπειρία μας αντιλαμβανόμαστε πως η μουσική μπορεί να προκαλέσει πολύ ισχυρά συναισθήματα, συνήθως ισχυρότερα από ότι άλλες μορφές ερεθισμάτων. Ως αποτέλεσμα, η σχετική έρευνα πάνω στην ανάλυση των εγκεφαλικών αποκρίσεων σε μουσικά ερεθίσματα, που είναι και το θέμα της παρούσας εργασίας, έχει μεγάλο ενδιαφέρον. Μια από τις πιο σημαντικές παραμέτρους στη μουσική αντίληψη φαίνεται πως είναι η ικανότητά μας να αναγνωρίζουμε και να αναμένουμε χρονικά μοτίβα. Πρόκειται για μια εγγενή ικανότητα του ανθρώπου αλλά και βασικό χαρακτηριστικό των μουσικών σημάτων, εκφραζόμενο μέσω του ρυθμού και της επανάληψης. Εμείς από την πλευρά μας θα διερευνήσουμε την ύπαρξη συσχετίσεων ανάμεσα στα συναισθηματικά χαρακτηριστικά των ΗΕΓ σημάτων και χαρακτηριστικών των μουσικών ηχητικών κυματομορφών, μέσω τεχνικών Μηχανικής Μάθησης.

Θεωρητικό Υπόβαθρο

Αρχές Επεξεργασίας Σήματος

Για να προσεγγίσουμε συστηματικά το εν λόγω θέμα, θα χρησιμοποιήσουμε τεχνικές από τον τομέα της Επεξεργασίας Σημάτων. Η Ψηφιακή Επξεργασία Σήματος (ΨΕΣ) είναι ένας ταχύτατα αναπτυσσόμενος κλάδος που θεμελιώνεται στα Μαθηματικά, τη Φυσική και την Επιστήμη των Υπολογιστών. Το βασικό στοιχείο της ανάλυσής μας θα είναι το σήμα (signal), δηλαδή μια φυσική ποσότητα που κωδικοποιεί ένα είδος πληροφορίας. Αν και ορισμένα σήματα μπορούν να μοντελοποιηθούν ντετερμινιστικά μέσω εξισώσεων, τα περισσότερα φυσικά σήματα συνήθως περιγράφονται μόνο από στατιστική άποψη. Σε κάθε περίπτωση, ένα σήμα αναλύεται μέσω ενός συστήματος (system), που υλοποιεί μια αντιστοίχιση (συνάρτηση) μιας εισόδου σε μια μοναδική έξοδο. Μια τέτοια συνάρτηση είναι αυτή που προσδιορίζει το φασματικό περιεχόμενο ενός σήματος και καλείται *Μετασχηματισμός Fourier*. Ο συγκεκριμένος μας λέει πως τα απολύτως ολοκληρώσιμα σήματα μπορούν να παρασταθούν ως γραμμικός συνδυασμός (άπειρων) ημιτονοειδών κυμάτων ή μιγαδικών εκθετικών:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\omega)e^{j\omega t}d\omega, \quad \text{όπου} \quad X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t}dt$$

Ο Μ/Σ Fourier χρησιμοποιείται ευρέως στην ανάλυση σημάτων με σημαντικό συχνοτικό περιεχόμενο, όπως είναι για παράδειγμα η μουσική (νότες) και τα εγκεφαλικά σήματα (ρυθμοί). Ανάμεσα στις διαφρετικές εκδοχές του μετασχηματισμού, στην παρούσα εργασία θα επικεντρωθούμε στον Διακριτό Μ/Σ Fourier (DFT) που εφαρμόζεται τυπικά σε ψηφιακά, πεπερασμένα σήματα μέσω του αλγορίθμου *Fast Fourier Transform* (FFT):

$$X[n] = \sum_{k=0}^{N-1} x[k]e^{-j\frac{2\pi}{N}nk} \quad (n = 0, 1, \ldots, N-1)$$

Το φασματικό περιεχόμενο ενός σήματος δε μπορεί, ωστόσο, να μας προσδώσει πληροφορία σχετικά με την χρονική τοποθέτηση των εμφανιζόμενων συχνοτήτων. Ένας άμεσος τρόπος για να προσεγγίσουμε το πρόβλημα είναι να σπάσουμε το εκάστοτε σήμα σε, πιθανώς επικαλυπτόμενα, κομμάτια και να εφαρμόσουμε τον μετασχηματισμό σε καθένα από αυτά. Η τεχνική αυτή λέγεται Μ/Σ Fourier βραχέος χρόνου (STFT). Συνενώνοντας στη συνέχεια τα φάσματα που προκύπτουν στον άξονα του χρόνου, καταλήγουμε σε μια διδιάστατη χρονοσυχνοτική αναπαράσταση, το φασματογράφημα (spectrogram). Συγκεκριμένα, δεδομένου ενός ψηφιακού σήματος $x[n]$ και παραθύρου $w[n]$, η μαθηματική έκφρασή του είναι η εξής:

$$\text{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega\pi}$$

Χρησιμοποιώντας τις παραπάνω τεχνικές μπορούμε να αναλύσουμε επαρκώς μια μεγάλη γκάμα ντετερμινιστικών σημάτων. Παρόλ' αυτά, τα περισσότερα σήματα σε φυσικές συνθήκες είναι πολύπλοκα και σπάνια μπορούν να προσδιοριστούν με ακρίβεια μέσω μαθηματικών εξισώσεων (π.χ. η κίνηση των ηλεκτρονίων). Για τέτοια σήματα προτιμούμε να αναλύουμε τα στατιστικά τους χαρακτηριστικά, χρησιμοποιώντας στοιχεία από τη θεωρία πιθανοτήτων. Συγκεκριμένα, θεωρούμε πως κάθε εμφάνιση ενός τυχαίου σήματος είναι ένα ενδεχόμενο σε ένα δειγματικό χώρο. Τέτοιοι δειγματικοί χώροι σημάτων ονομάζονται *τυχαίες διαδικασίες* (random processes) και το εκάστοτε σήμα είναι στην ουσία μια τυχαία

μεταβλητή του χώρου αυτού. Οι στατιστικές μετρικές που ορίζουμε για τυχαίες μεταβλητές (μέση τιμή, διασπορά κλπ) μπορούν να επεκταθούν άμεσα, ενώ μας ενδιαφέρει ιδιαίτερα η συνάρτηση αυτοσυσχέτισης δύο τυχαίων διαδικασιών:

$$R_X(t_1, t_2) = \mathbb{E}[X(t_1), X(t_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X(t_1),X(t_2)}(x_1, x_2)\, dx_1 dx_2$$

Η συνάρτηση αυτοσυσχέτισης αποδεικνύεται χρήσιμη σε πληθώρα υπολογισμών, ιδίως σε ό,τι έχει να κάνει με την εκτίμηση του φάσματος τυχαίων σημάτων. Από τη στιγμή που δεν μπορούμε να προσδιορίσουμε αναλυτικά ένα τυχαίο σήμα, επιλέγουμε να εφαρμόσουμε τον Μ/Σ Fourier στην συνάρτηση αυτοσυσχέτισης προκειμένου να έχουμε εικόνα του φάσματός του. Το μέγεθος που προκύπτει καλείται Πυκνότητα Φάσματος Ισχύος (PSD) και χρησιμοποιείται ευρέως για την εξαγωγή φασματικών χαρακτηριστικών σε φυσικά σήματα.

Στην παρούσα εργασία ασχολούμαστε εκτενώς και με την έννοια των φράκταλ σχημάτων. Ο όρος φράκταλ (fractal) προέρχεται από τη λατινική λέξη fractum (σπασμένο) και προτάθηκε από τον Mandelbrot [45] προκειμένου να περιγράψει ασυνήθιστα σχήματα που δε μπορούν να μοντελοποιηθούν από έννοιες της συμβατικής γεωμετρίας. Αυτό που παρατήρησε μελετώντας τη μορφή της βρετανικής ακτογραμμής είναι ότι αυξάνοντας την ακρίβεια των μετρήσεων, το προκύπτον μήκος της αυξάνεται επίσης. Αυτό υποδεικνύει την ομοιότητα που έχει η πολυπλοκότητα του σχήματος της ακτογραμμής σε διαφορετικές κλίμακες. Ως εκ τούτου, φράκταλ αλγόριθμοι και τεχνικές επινοήθηκαν για να αναλύσουν τέτοιου είδους δομές αυτο-ομοιότητας (self-similarity) και επαναληπτικής πολυπλοκότητας, με κυριότερη την έννοια της φράκταλ διάστασης. Πράγματι, τα φράκταλ σχήματα έχουν ιδιότητες που προσομοιάζουν σε σχήματα διάστασης μεγαλύτερης από την τοπολογική τους. Πολλοί αλγόριθμοι έχουν προταθεί για τον προσδιορισμό της, ενώ κάποιοι που θα αξιοποιήσουμε στην παρούσα εργασία είναι η Minkowski-Bouligand Dimension [41] και η Higuchi Dimension [50]. Πολλά φυσικά σχήματα δύνανται να εμφανίσουν φράκταλ ιδιότητες, που μπορούν μάλιστα να δώσουν χρήσιμες σημασιολογικές πληροφορίες. Τέτοια παραδείγματα είναι η μουσική, το γήινο ανάγλυφο αλλά και τα περισσότερα βιοσήματα [150, 39].

Αρχές Μηχανικής Μάθησης

Ο χώρος της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI) περιγράφεται γενικά ως ο χώρος μελέτης των ευφυών πρακτόρων (agents), δηλαδή κάθε συσκευής που μπορεί να αντιληφθεί το περιβάλλον της και να εκτελέσει ενέργειες με στόχο τη μεγιστοποίηση ενός οφέλους, κατ' αντιστοιχία με τον τρόπο που αντιλαμβανόμαστε τη φυσική και ανθρώπινη νοημοσύνη. Η Μηχανική Μάθηση (Machine Learning - ML) είναι ένας υπόχωρος του AI που ασχολείται με την εκπαίδευση ευφυών υπολογιστικών μοντέλων πάνω σε δεδομένα. Οι κυριότεροι αλγόριθμοι μάθησης κατηγοριοποιούνται σε μάθηση υπό επίβλεψη (supervised learning), όπου τα δεδομένα συνοδεύονται από επισημειώσεις που θέλουμε να προβλέψουμε, και μάθηση χωρίς επίβλεψη (unsupervised learning), όπου αυτές απουσιάζουν και καλούμαστε να εξάγουμε σημασιολογικά και στατιστικά στοιχεία από τα ίδια τα δεδομένα.

Τα περισσότερα πειράματα στη Μηχανική Μάθηση χρησιμοποιούν αλγορίθμους μάθησης με επίβλεψη. Συγκεκριμένα, δίνεται συνήθως ένα σύνολο δεδομένων $\boldsymbol{x}$ και ένα σύνολο επισημειώσεων y για κάθε στοχείο του συνόλου δεδομένων. Ο στόχος είναι να προσδιοριστεί με όσο μεγαλύτερη ακρίβεια, μια συνάρτηση $y = f(\boldsymbol{x})$ που να μπορεί να γενικευτεί σε αυθαίρετα δεδομένα του ίδιου τύπου. Αναφέρουμε περιληπτικά 2 βασικούς αλγόριθμους αυτής της κατηγορίας: τη γραμμική παλινδρόμηση (Linear Regression) και τις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM).

Η γραμμική παλινδρόμηση είναι ίσως ο πιο απλός και παλιός αλγόριθμος μάθησης, και περιλαμβάνει την εκτίμηση της πιο αντιπροσωπευτικής ευθείας που να περιγράφει την τάση ενός συνόλου δεδομένων σε έναν διανυσματικό χώρο. Στην απλούστερη περίπτωση που εξετάζουμε, τα δεδομένα υπακούν σε μια γραμμική μορφή $\boldsymbol{y} = \boldsymbol{bX}$, όπου:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ 1 & x_{31} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{bmatrix}$$

και οι συντελεστές b εκτιμώνται μέσω της ελαχιστοποίησης του σφάλματος:

$$\hat{b} = \arg\min_{b} \|\boldsymbol{y} - \boldsymbol{bX}\|_2^2$$

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) [32] είναι ένας από τους πιο αποδοτικούς αλγόριθμους κατηγοριοποίησης, καθώς χρησιμοποιεί τετραγωνικό προγραμματισμό προκειμένου να προσδιορίσει βέλτιστα υπερεπίπεδα που θα διαχωρίζουν τις κλάσεις των δεδομένων στο χώρο. Θεωρούμε χάριν απλότητας το δυαδικό πρόβλημα: Έστω ένα σύνολο από N διανύσματα εισόδου $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ και αντίστοιχες επισημειώσεις $y_1, ..., y_N$ όπου $x_i \in \mathbb{R}^d$ και $y_i \in \{-1, 1\}$. Όλα τα υπερεπίπεδα στο $\mathbb{R}^d$ μπορούν να εκφραστούν μέσω ενός διανύσματος $\boldsymbol{w}$ και μιας σταθεράς, όπως φαίνεται στην εξίσωση:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$$

Ένα τέτοιο υπερεπίπεδο $(\boldsymbol{w}, b)$ θα διαχωρίζει επιτυχώς τα δεδομένα όταν

$$y_i \left(\mathbf{x_i} \cdot \mathbf{w} + b\right) \geq 1 \quad \forall i$$

και επίσης όταν μπορεί να γενικεύει αποδοτικά σε νέα δεδομένα. Το τετραγωνικό πρόβλημα που προκύπτει επιλύεται με τη βοήθεια πολλαπλασιαστών Lagrange. Υπάρχουν όμως περιπτώσεις κατανομών δεδομένων που δεν είναι "γραμμικά" διαχωρίσιμες από υπερεπίπεδα. Σε αυτή την περίπτωση, για να γενικεύσουμε την SVM προσέγγιση, μπορούμε να ορίσουμε μια αντιστοίχιση $\mathbf{z} = \phi(\mathbf{x})$ (πυρήνας - kernel) μέσω της οποίας να μετασχηματίσουμε τα διανύσματα των δεδομένων σε έναν άλλο χώρο, στον οποίο ο διαχωρισμός τους είναι ευκολότερος.

Νευρωνικά Δίκτυα

Η πιο ευρέως χρησιμοποιούμενη μέθοδος μάθησης σήμερα στηρίζεται στην έννοια των νευρωνικών δικτύων (Artificial Neural Networks - ANN), ενός μοντέλου που επιχειρεί να προσομοιώσει τις λειτουργικότητες του ανθρώπινου εγκεφάλου. Η βασική υπολογιστική μονάδα των ANN είναι ο νευρώνας (perceptron). Κατά αντιστοιχία με τον βιολογικό νευρώνα, το perceptron δέχεται ένα σύνολο εισόδων τις οποίες αθροίζει, εφαρμόζοντας ανάλογα βάρη, και χρησιμοποιεί μια συνάρτηση ενεργοποίησης προκειμένου να διαχωρίσει τις εξόδους σε συγκεκριμένες κλάσεις. Ένας νευρώνας εκπαιδεύεται στην κατηγοριοποίηση δεδομένων μέσω του αλγορίθμου Perceptron [127]. Προκειμένου ωστόσο ένα μοντέλο να μάθει πολύπλοκες, μη γραμμικές συναρτήσεις, χρησιμοποιούνται αρχιτεκτονικές συνδυασμένων νευρώνων σε διακριτά επίπεδα, όπου κάθε νευρώνας συνδέεται με όλους τους νευρώνες του προηγούμενου επιπέδου και κανέναν του επιπέδου του, όπως φαίνεται στο Figure 1. Τέτοια

δίκτυα εκπαιδεύονται μέσω ενός αλγορίθμου ακολουθιακού υπολογισμού των μεταβολών των βαρών τους (Back Propagation), ενώ η εκπαίδευσή τους ελέγχεται από τις επισημειώσεις των δεδομένων μέσω μιας συνάρτησης κόστους.

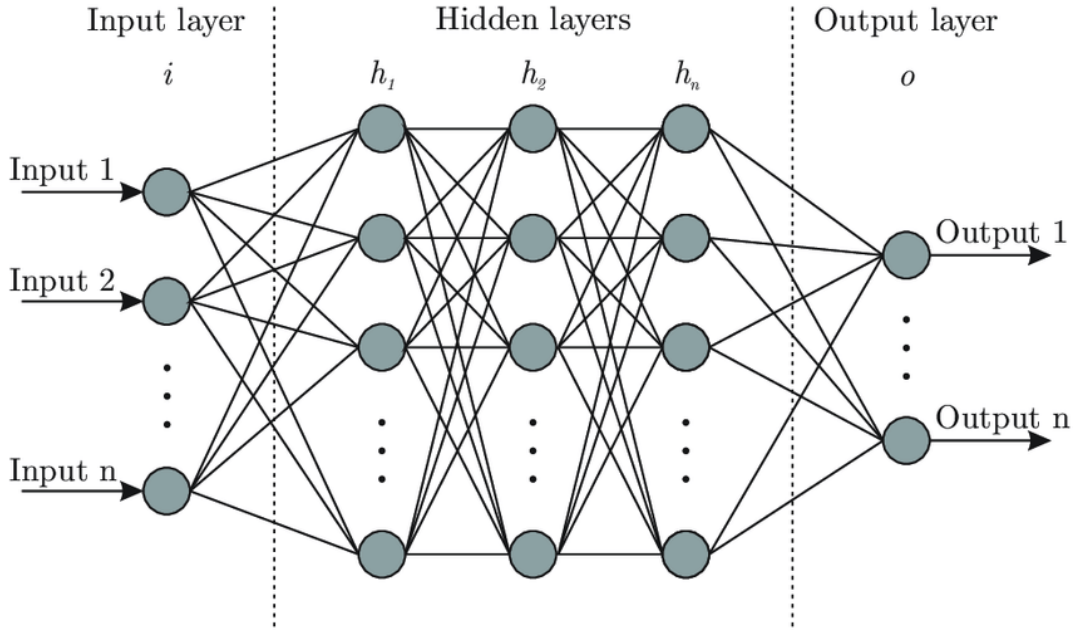

Figure 1: Νευρωνικό Δίκτυο 5 επιπέδων, που περιέχει n εισόδους, 3 κρυφά επίπεδα και ένα επίπεδο n εξόδων. Είναι παράδειγμα ενός βαθιού νευρωνικού δικτύου (DNN). Πηγή: [18].

Προηγένα Θέματα Νευρωνικών Δικτύων

Το νευρωνικό δίκτυο του Figure 1 αποτελείται από περισσότερα του ενός κρυφά επίπεδα. Δίκτυα με αυτή την ιδιότητα ονομάζονται *Βαθιά Νευρωνικά Δίκτυα* (DNNs) και ο τομέας που ασχολείται με τη μελέτη τους ονομάζεται Βαθιά Μάθηση. Αλγόριθμοι και μοντέλα βαθιάς μάθησης έχουν μόλις πρόσφατα γενικευθεί στην σχετική βιβλιογραφία, λόγω των ισχυρών τους δυνατοτήτων και της ικανότητας ταχείας εκπαίδευσης που προσφέρουν τα σύγχρονα υπολογιστικά συστήματα. Η κυριότερη διαφορά τους είναι η δυνατότητα να εξάγουν χαρακτηριστικά από τα δεδομένα και να τα συνδυάζουν αποτελεσματικά στα κρυφά τους επίπεδα, κάτι που στις ως τώρα μεθόδους γινόταν χειρωνακτικά.

Σύντομα ωστόσο έγινε αντιληπτό πως διαφορετικοί τύποι δεδομένων απαιτούν διαφορετικό τρόπο αντιμετώπισης. Στην επεξεργασία και αναγνώριση εικόνων, για παράδειγμα, προτάθηκαν τα Συνελικτικά Νευρωνικά Δίκτυα (CNN) για την εξαγωγή και κατηγοριοποίηση των χαρακτηριστικών τους. Η ιδέα βασίζεται στους κλασικούς αλγορίθμους Υπολογιστικής Όρασης που χρησιμοποιούσαν πυρήνες (kernels) προκειμένου να εξάγουν ακμές, γωνίες ή άλλα χαρακτηριστικά. Μέσω των αλγορίθμων βαθιάς μάθησης, τα CNNs εκπαιδεύουν ειδικά προσαρμοσμένους πυρήνες στα εκάστοτε δεδομένα. Έκτοτε, η χρησιμότητα των δικτύων αυτών έχει εξαπλωθεί και πολλές διαφορετικές αρχιτεκτονικές έχουν προταθεί στη βιβλιογραφία [140], κάποιες εκ των οποίων θα χρησιμοποιήσουμε στην εργασία αυτή ως πρότυπα. Μια άλλη, επίσης δημοφιλής, κατηγορία νευρωνικών δικτύων είναι τα αναδρομικά δίκτυα (RNNs) τα οποία ορίστηκαν πρωταρχικά για την ανάλυση δεδομένων με χρονική συσχέτιση (χρονοσειρές, σειριακά δεδομένα κλπ). Ονομάζονται έτσι επειδή εφαρμόζουν την ίδια επεξεργασία σε κάθε στοιχείο μιας χρονοσειράς, λαμβάνοντας υπόψη προηγούμενα στοιχεία.

Τέλος, μια κατηγορία αλγορίθμων που θα παίξουν σημαντικό ρόλο στο ερευνητικό κομμάτι της εργασίας είναι η πολυτροπική μάθηση (Multimodal Learning), δηλαδή η μάθηση

που περιλαμβάνει διαφορετικά είδη δεδομένων. Πρόκειται για μια αναπτυσσόμενη ερευνητική περιοχή, μιας και όλο και περισσότερα δεδομένα γίνονται διαθέσιμα σε όλο και περισσότερες μορφές. Επιχειρείται έτσι να προσεγγιστεί πιο πιστά ο τρόπος που ο άνθρωπος αντιλαμβάνεται και εξάγει στοιχεία για μια κατάσταση. Οι αλγόριθμοι βαθιάς μάθησης βρίσκουν ιδανική εφαρμογή σε αυτή την περιοχή μιας και έχουν τη δυνατότητα να εξάγουν συγκεκριμένα χαρακτηριστικά από κάθε είδος πληροφορίας, προκειμένου αυτά να δρουν συμπληρωματικά στα εκάστοτε προβλήματα κατηγοριοποίησης. Η συπληρωματικότητα αυτή επιτυγχάνεται είτε με την από κοινού συγχώνευση και επεξεργασία τους ως ένα διανύσμα χαρακτηριστικών είτε μετρώντας και αξιολογώντας την ομοιότητά τους. Η τεχνική αυτή ονομάζεται Μετρική Μάθηση (Metric Learning) και στοχεύει, αντί να προβλέψει μια συγκεκριμένη κατηγορία, να προβλέψει την ομοιότητα ζευγών δεδομένων μέσω της απόστασης των διανυσμάτων των χαρακτηριστικών τους στον εκάστοτε διανυσματικό χώρο.

Multifractal Ανάλυση Σημάτων ΗΕΓ

Στην πρώτη ενότητα πειραμάτων ασχολούμαστε με την ανάπτυξη καινοτόμων αλγορίθμων για την επεξεργασία των fractal και multifractal ιδιοτήτων των σημάτων ΗΕΓ καθώς και με το κατά πόσο αυτές οι ιδιότητες υποδεικνύουν συναισθηματική πληροφορία. Το ΗΕΓ έχει ευρέως καθιερωθεί ως ένα φράκταλ σήμα με πολύπλοκη δομή και σημαντικές αλλοιώσεις λόγω θορύβου, κάτι που έχει σταθεί εμπόδιο στις προσπάθειες σημασιολογικής ανάλυσής του. Παρόλ' αυτά χρησιμοποείται ευρέως στη βιβλιογραφία για Αναγνώριση Συναισθήματος, καθώς αποδίδει μια υψηλής ακρίβειας χρονική ανάλυση και τα χρονοσυχνοτικά χαρακτηριστικά των ρυθμών του έχουν αποδειχθεί σημαντικοί συναισθηματικοί δείκτες.

Φράκταλ Ανάλυση σε Πολλαπλές Κλίμακες

Παρότι έχουν προταθεί πολλές εκδοχές υπολογισμού της φράκταλ διάστασης ενός ΗΕΓ, η χαώδης και μη γραμμική μορφή του συγκεκριμένου σήματος επιβάλλει την αναζήτηση πιο πολύπλοκων τεχνικών. Εστιάζουμε συγκεκριμένα στην υπόθεση πως τα ΗΕΓ εμφανίζουν διαφορετικές μορφές πολυπλοκότητας σε διαφορετικές κλίμακες, κάτι που επιφέρει μια μεταβλητότητα στον υπολογισμό της διάστασης. Για την αντιμετώπιση αυτού του ζητήματος προτείνουμε την εφαρμογή της φράκταλ διάστασης πολλαπλών κλιμάκων (Multiscale Fractal Dimension - MFD) [90], ενός αλγορίθμου που βασίζεται στη μέτρηση του εμβαδού της αναπαράστασης του σήματος σε πολλαπλές κλίμακες, όταν αυτό καλύπτεται από ένα κάλυμμα δίσκων ανάλογης ακτίνας, με κέντρα στα σημεία του σήματος (κάλυμμα Minkowski). Ο αλγόριθμος είναι γνωστός ως "Μέθοδος μορφολογικής κάλυψης" (Figure 2):

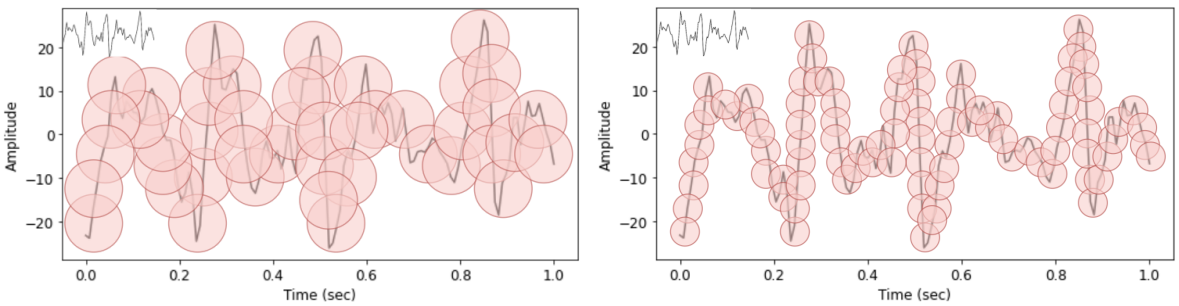

Figure 2: Αναπαράσταση της Μορφολογικής Κάλυψης Minkowski για ένα δείγμα σήματος ΗΕΓ.

Όπως είναι λογικό, το εμβαδό του καλύμματος σε κάθε κλίμακα μας δίνει μια διαφορετική οπτική για την πολυπλοκότητα του υπό εξέταση σήματος. Συγκεκριμένα, αποδεικνύεται πως το εμβαδό αυτό $A_B(s)$ συνδέεται με τη μετρούμενη κλίμακα s μέσω της σχέσης

$$\log[A_B(s)] = (2 - D)\log(s) + \text{constant}$$

όπου D η ζητούμενη φράκταλ διάσταση. Για να εξάγουμε τις περισσότερες σημαντικές μεταβολές στην πολυπλοκότητα, εφαρμόζουμε τον αλγόριθμο σε ένα μικρό παράθυρο από κλίμακες (ακτίνες δίσκων) και αναπαριστούμε τη φράκταλ διάσταση που προκύπτει συναρτήσει του χρόνου (φρακτόγραμμα - fractogram).

Ο δεύτερος αλγόριθμος που θα χρησιμοποιήσουμε είναι αυτός της Multifractal Detrended Fluctuation Analysis (MFDFA) [62]. Η μέθοδος βασίζεται στην εκτίμηση του Hurst εκθέτη H που είναι συμπληρωματικό μέγεθος της φράκταλ διάστασης, σύμφωνα με την σχέση $D = 2 - H$. Δεδομένης εισόδου $x[n]$ μήκους N, ο αλγόριθμος υπολογίζει πρώτα το συσσωρευτικό άθροισμα $y[n] = \sum_{m=1}^{N} (x[m] - \mu_x)$ και το χωρίζει σε μη-επικαλυπτόμενα παράθυρα. Στη συνέχεια, αφαιρείται από κάθε υπο-σήμα η βασική του τάση μέσω Γραμμικής Παλινδρόμησης, ούτως ώστε η τελική μορφή να περιέχει αποκλειστικά τις μη-γραμμικές πολυπλοκότητες του σήματος. Τέλος, υπολογίζεται η RMS τιμή κάθε τέτοιου τμήματος και λαμβάνεται ως έξοδος η μέση RMS τιμή για όλα τα διαθέσιμα παράθυρα.

Το αποτέλεσμα αυτής της διαδικασίας είναι ένα διάνυσμα τιμών, μία για κάθε διαθέσιμη κλίμακα ανάλυσης (αντίστοιχα με το μήκος των παραθύρων). Για φράκταλ σήματα, η γραφική παράσταση στον άξονα των κλιμάκων είναι ευθεία σε log-log αναπαράσταση, υποδεικνύοντας την ισχύ του εκθετικού νόμου. Σύμφωνα με αυτόν, η κλίση της ευθείας είναι η εκτίμηση του εκθέτη H. Για να μεταβούμε από την απλή μέθοδο DFA στην Multifractal DFA, αρκεί απλά να υπολογίσουμε την τελική τιμή όχι κατά RMS αλλά κατά μια σειρά από ροπές q. Ως αποτέλεσμα, προκύπτει μια διαφορετική γραμμή για κάθε ροπή που, στην περίπτωση των φράκταλ σημάτων, θα είναι παράλληλες ευθείες (κοινό H) ενώ στα multifractals θα είναι συγκλίνουσες προς μεγαλύτερες κλίμακες, όπως φαίνεται στο Figure 3.

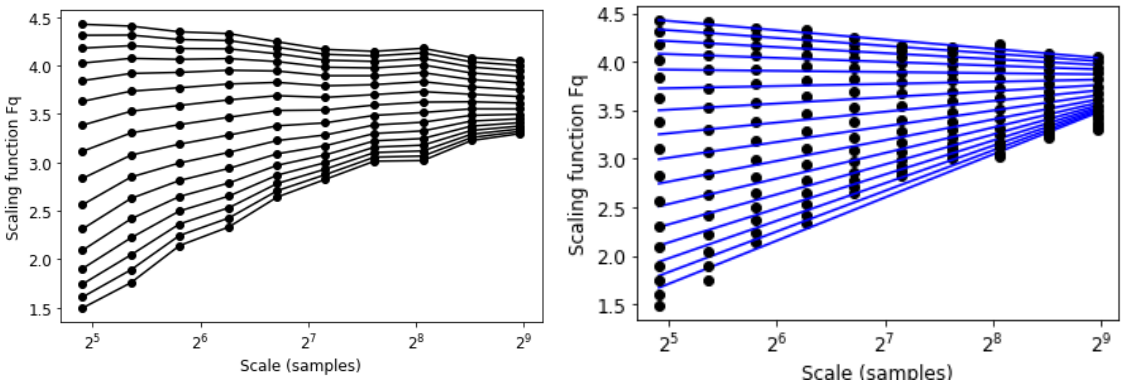

Figure 3: MFDFA σε δείγμα ΗΕΓ: Απεικονίζουμε 16 DFA γραμμικές αναπαραστάσεις της $F_q(s)$ μαζί με τη γραμμική πρόβλεψη της της κλίσης που προσδιορίζει τον γενικευμένο Hurst εκθέτη $H(q)$.

Το ΗΕΓ ως Multifractal Σήμα

Τόσο η παρούσα, όσο και η ανάλυση της επόμενης ενότητας, βασίζονται στο σύνολο δεδομένων DEAP [65], το οποίο παρουσιάζουμε αναλυτικά στο Appendix της εργασίας. Συνοπτικά, το DEAP είναι ένα ευρύτατα χρησιμοποιούμενο σύνολο δεδομένων για πολυτροπική αναγνώριση συναισθήματος, με έμφαση στα βιολογικά σήματα και ιδιαίτερα το

ΗΕΓ. Τα δεδομένα προέρχονται από 32 ανθρώπους που παρακολούθησαν 40 επιλεγμένα μουσικά κλιπ του ενός λεπτού ενώ γινόταν καταγραφή του ΗΕΓ, διάφορων άλλων βιολογικών σημάτων καθώς και (σε μερικούς) η έκφραση του προσώπου. Μετά την παρακολούθηση κάθε κλιπ, ο εκάστοτε συμμετέχων βαθμολογούσε το συναίσθημα που ένιωσε στη διανυσματική αναπαράσταση (valence, arousal, dominance, liking, familiarity). Το ΗΕΓ μετρήθηκε σύμφωνα με το καθιερωμένο σύστημα 10-20 [58], με 32 ηλεκτρόδια. Εμείς χρησιμοποιούμε τα προεπεξεργασμένα δεδομένα ΗΕΓ του DEAP, τα οποία και έχουν υποστεί υποδειγματοληψία στα 128Hz και ζωνοπερατό φιλτράρισμα στις μπάντες ϑ μέχρι και γ.

Για τη διερεύνηση των φράκταλ χαρακτηριστικών των ΗΕΓ σημάτων πρώτα εξετάζουμε την στατικότητά τους. Χρησιμοποιούμε το Augmented Dickey-Fuller (ADF) Test [37] και, προς έκπληξή μας, καταλήγουμε στο ότι τα δεδομένα μας είναι αυστηρά στατικά, δηλαδή οι βασικές στατιστικές τους μετρικές δεν αλλάζουν σε μεγάλη κλίμακα. Μετά από αναπαραγωγή της διαδικασίας προεπεξεργασίας των δεδομένων, αποδίδουμε την παρατηρούμενη στατικότητα στο ζωνοπερατό φιλτράρισμα του ΗΕΓ, που αποκλείει τις χαμηλότερες συχνότητες. Βασισμένοι σε αυτό, μπορούμε να μοντελοποιήσουμε τα δεδομένα μας ως fractional γκαουσιανό θόρυβο (fractional Gaussian noise - fGn) ελάχιστου H. Πράγματι, τρέχοντας τον DFA αλγόριθμο για τον υπολογισμό του εκθέτη, παρατηρούμε πως στη συντριπτική πλειοψηφία των δεδομένων, ο εκθέτης προσεγγίζει το μηδέν. Αυτό υποδεικνύει μια εξαιρετικά μεγάλη φράκταλ διάσταση και μια πολύ fragmented δομή. Η εικόνα αυτή συμφωνεί με την παρατηρούμενη στατικότητα των fGn σημάτων και φανερώνει αρνητικές χρονικές συσχετίσεις στη δομή τους.

Εξαγωγή Φράκταλ Χαρακτηριστικών

Στην ανάλυσή μας αξιοποιούμε τόσο τα δοθέντα ΗΕΓ όσο και ξεχωριστά τις συχνοτικές τους μπάντες, εκτός της ϑ που δεν αναφέρεται στη βιβλιογραφία ως κρίσιμη σε συναισθηματικούς δείκτες. Οι μπάντες λαμβάνονται με ζωνοπερατό φιλτράρισμα στο αρχικό σήμα. Χωρίζουμε επίσης τα ηλεκτρόδια που θα εξετάσουμε, διαλέγοντας 12 μπροστά-αριστερά και 12 μπροστά-δεξιά ηλεκτρόδια, σε συμφωνία με την σχετική βιβλιογραφία [178]. Στα πειράματά μας θα συγκρίνουμε τα χαρακτηριστικά των προτεινόμενων αλγορίθμων με βασικές μεθόδους που χρησιμοποιούνται ευρέως, την πυκνότητα φάσματος ισχύος (PSD) του δεύτερου μισού του σήματος και την Higuchi φράκταλ διάσταση του σήματος, μετά από παραθυροποίηση.

Σχετικά με τα MFD χαρακτηριστικά, χωρίζουμε και εδώ το σήμα σε 7 παράθυρα των 15 sec, επικαλυπτόμενα κατά 50%. Για κάθε κομμάτι υπολογίζουμε το φρακτόγραμμα για κλίμακες από 1 ως 274 σημεία και λαμβάνουμε σαν διάνυσμα χαρακτηριστικών 30 γραμμικά δειγματοληπτημένες τιμές από κάθε φρακτόγραμμα. Ενώ πειραματιστήκαμε και με αυτά τα δεδομένα, η προτεινόμενη οδός είναι η εξαγωγή μέσης και διάμεσης τιμής και της τυπικής απόκλισης των 7 παραθύρων για κάθε σημείο, καταλήγοντας σε ένα 90D διάνυσμα. Σχετικά με τα MFDFA χαρακτηριστικά, χρησιμοποιούμε ως διάνυσμα τις τιμές και τετμημένες του multifractal φάσματος $D(q)$, που προκύπτει μέσω των σχέσεων:

$$D(q) = q'h(q) - t(q'), \quad h(q_n) = \frac{t(q_n) - t(q_{n-1})}{q_n - q_{n-1}}, \quad t(q) = qH(q) - 1.$$

Πειραματική Αξιολόγηση

Πραγματοποιούμε πειράματα τόσο για κάθε έναν συμμετέχοντα ξεχωριστά (subject dependent - SD) όσο και από κοινού για όλους τους συμμετέχοντες (subject independent - SI). Μετά την εξαγωγή των χαρακτηριστικών, ενοποιούμε το σύνολο των καναλιών-ηλεκτροδίων

και το τελικό διάνυσμα δίνεται ως είσοδος σε μια Μηχανή Διανυσμάτων Υποστήριξης (SVM). Ως επισημειώσεις χρησιμοποιούμε τις μετρικές valence και arousal σε ξεχωριστά πειράματα δυαδικής κατηγοριοποίησης (high - low) με όριο την ενδιάμεση βαθμολογία (5).

Features	Channels	Raw Signal	Alpha Band	Beta Band	Gamma Band	Combined
PSD		**0.642** — 0.652	0.598 — 0.645	0.629 — 0.639	**0.635** — 0.620	0.631 — 0.648
HFD	Front	0.615 — 0.638	0.605 — 0.655	0.591 — 0.643	0.601 — 0.634	**0.638** — 0.645
MFD	Left	0.620 — 0.661	**0.626** — **0.669**	0.591 — **0.653**	0.594 — 0.636	0.612 — **0.661**
MFDFA		0.577 — **0.662**	0.571 — 0.643	0.577 — 0.649	0.592 — **0.651**	0.586 — 0.658
PSD		0.627 — 0.644	0.616 — 0.645	**0.637** — 0.641	0.623 — 0.627	0.623 — 0.646
HFD	Front	0.606 — 0.644	0.604 — 0.655	0.595 — 0.633	0.572 — 0.627	0.623 — 0.644
MFD	Right	0.607 — 0.655	0.605 — 0.652	0.566 — 0.652	0.602 — 0.641	0.597 — 0.657
MFDFA		0.587 — 0.655	0.573 — 0.641	0.603 — 0.650	0.573 — 0.620	0.586 — 0.652

Table 1: Ακρίβεια για τα Subject Dependent πειράματα στη μορφή: Valence — Arousal

Features	Channels	Raw Signal	Alpha Band	Beta Band	Gamma Band	Combined
PSD		0.554 — 0.569	0.547 — 0.564	0.549 — 0.562	0.553 — 0.570	0.546 — 0.564
HFD	Front	0.541 — 0.601	0.552 — 0.588	0.541 — 0.616	0.545 — 0.584	**0.585** — **0.621**
MFD	Left	0.553 — 0.606	**0.566** — **0.631**	0.545 — **0.618**	**0.554** — 0.580	0.559 — 0.615
MFDFA		**0.569** — **0.630**	0.546 — 0.600	0.545 — 0.598	0.532 — 0.545	0.553 — 0.608
PSD		0.553 — 0.580	0.557 — 0.560	**0.558** — 0.573	0.552 — 0.579	0.555 — 0.575
HFD	Front	0.525 — 0.573	0.566 — 0.582	0.544 — 0.595	0.549 — 0.567	0.571 — 0.605
MFD	Right	0.552 — 0.601	0.556 — 0.605	0.547 — 0.587	0.545 — **0.588**	0.560 — 0.607
MFDFA		0.555 — 0.619	0.552 — 0.580	0.549 — 0.591	0.539 — 0.584	0.544 — 0.599

Table 2: Ακρίβεια για τα Subject Independent πειράματα στη μορφή: Valence — Arousal

Τα αποτελέσματα των κύριων πειραμάτων παρατίθενται στα Tables 1 και 2. Σημειώνουμε καταρχάς την αναμενόμενη διάκριση ανάμεσα στα 2 είδη πειραμάτων, SD και SI, που αναδεικνύει την ιδιότητα των ΗΕΓ σημάτων να προσδιορίζονται εν πολλοίς από ατομικούς παράγοντες. Τα PSD χαρακτηριστικά φαίνεται να αποδίδουν περισσότερο στα SD πειράματα, όπου για το raw signal λαμβάνουμε ακρίβεια 64% στην πρόβλεψη του valence και 65.2% του arousal. Αντίθετα, στο δεύτερο πίνακα παρατηρούμε σημαντική πτώση, συγκριτικά με άλλα χαρακτηριστικά, κάτι που υποδεικνύει πως τα συχνοτικά χαρακτηριστικά του ΗΕΓ επηρεάζονται από το συναίσθημα, αλλά διαφέρουν από άτομο σε άτομο. Αντίθετα, οι multifractal μέθοδοι αποδίδουν καλά και στα 2 είδη πειραμάτων, ιδίως στην αναγνώριση του arousal, όπου πετυχαίνουν 5% αύξηση. Στο SD πείραμα, τα MFD άλφα ρυθμού και τα MFDFA των ΗΕΓ πετυχαίνουν πάνω από 66%, ενώ τα υψηλότερα σκορ στο SI αγγίζουν το 63%.

Επικεντρώνοντας περισσότερο στις φράκταλ μεθόδους αναγνώρισης, τα multifractal χαρακτηριστικά εμφανίζουν παρόμοια απόδοση με την Higuchi διάσταση στο valence και καλύτερη απόδοση στο arousal, ενισχύοντας περαιτέρω την εικόνα πως η πολυπλοκότητα των ΗΕΓ σημάτων σε πολλαπλές κλίμακες μπορεί να είναι δείκτης συναισθηματικής έκφρασης. Σε ένα πείραμα δεύτερου χρόνου φαίνεται επιπλέον πως ο συνδυασμός των φράκταλ διαστάσεων βελτιώνει τις επιδόσεις του μοντέλου στην αναγνώριση arousal, ιδίως όσον αφορά τον άλφα ρυθμό. Μπορούμε πλέον να προβλέψουμε το arousal με ακρίβεια 67% και 64% στα SD και SI πειράματα αντίστοιχα. Τέλος, σημαντική παρατήρηση συνιστά και η βελτιωμένη επίδοση των σημάτων από το αριστερό μέρος του εγκεφάλου σε σχέση με το δεξί, ενώ ο συνδυασμός τους δε φαίνεται να αυξάνει τις δυνατότητες του μοντέλου. Καταληκτικά, είναι εμφανές πως τα multifractal χαρακτηριστικά των ΗΕΓ σημάτων και οι αρνητικές τους συσχετίσεις μπορούν να ληφθούν υπόψη στο σχεδιασμό συστημάτων αναγνώρισης συναισθήματος.

Διατροπική Μάθηση μεταξύ ΗΕΓ και Μουσικής

Η Διατροπική Μάθηση (Cross-Modal Learning) είναι μια κατηγορία αλγορίθμων πολυτροπικής μάθησης που αποσκοπεί στην εξεύρεση συνδετικών στοιχείων μεταξύ 2 ή περισσότερων ειδών δεδομένων. Τα δεδομένα επεξεργάζονται ώστε να προβληθούν σε έναν κοινό διανυσματικό χώρο, από όπου θα μπορούσαμε, δίνοντας εισόδους ενός είδους, να εξάγουμε σχετική πληροφορία από άλλο είδος δεδομένων. Εδώ θα επικεντρωθούμε στη διατροπική μάθηση μεταξύ ΗΕΓ του συνόλου DEAP και των μουσικών σημάτων που χρησιμοποιήθηκαν ως συναισθηματικά ερεθίσματα. Συγκεκριμένα, προτείνουμε ένα μοντέλο που θα μπορεί, αναλύοντας τα 2 είδη δεδομένων, να κάνει ακριβέστερες προβλέψεις συναισθήματος, αλλά και να επιστρέφει εκτιμήσεις μουσικών κομματιών σε ΗΕΓ "ερωτήματα", δηλαδή εισόδους για τις οποίες αναζητούμε τις πιο όμοιες αναπαραστάσεις μουσικών κομματιών στον κοινό διανυσματικό χώρο. Με την αξιοποίηση της συνδυαστικής αυτής πληροφορίας για καθέναν από τους συμμετέχοντες του πειράματος, εξάγουμε χρήσιμα στοιχεία σχετικά με τη φύση της μουσικής αντίληψης και τις χρονικές της μεταβολές.

Γεφυρώνοντας το Χάσμα

Θα κατασκευάσουμε ένα μοντέλο με 2 βασικούς κλάδους - νευρωνικά δίκτυα, έναν για τα εγκεφαλικά και έναν για τα μουσικά σήματα, ενώ θα αξιοποιήσουμε από κοινού τα κοινά χαρακτηριστικά τους και τις επισημειώσεις τους, όπως προτάθηκε στο [177]. Συγκεκριμένα, έχουμε στη διάθεσή μας μια συλλογή από n ζεύγη ΗΕΓ-μουσικής, που δηλώνουμε ως $T = \{(x_i^a, x_i^b)\}_{i=1}^n$, όπου x_i^a είναι το ΗΕΓ στοιχείο του i-οστού ζεύγους και x_i^b το αντίστοιχο μουσικό ερέθισμα. Κάθε ζεύγος συνοδεύεται από μια συναισθηματική ετικέτα $y_i \in \mathbb{R}^2$ για τις κατηγορίες valence και arousal. Για κάθε ζεύγος i, στόχος μας είναι να μάθουμε μια διανυσματική μορφή $u(i) = f(x_i^a, Y^a) \in \mathbb{R}^d$ για το ΗΕΓ και $v(i) = g(x_i^b, Y^b) \in \mathbb{R}^d$ για το μουσικό κομμάτι, όπου d είναι η διάσταση του κοινού διανυσματικού χώρου και Y^a, Y^b οι παράμετροι προς εκπαίδευση των 2 συναρτήσεων, ούτως ώστε να ικανοποιούνται όσο το δυνατόν οι εξής ιδιότητες: α) η ομοιότητα μεταξύ στοιχείων της ίδιας κατηγορίας να είναι μεγαλύτερη από αυτή μεταξύ στοιχείων που ανήκουν σε διαφορετικές κατηγορίες, και β) η ομοιότητα του εκάστοτε ζεύγους δεδομένων να είναι επίσης μεγαλύτερη από την ομοιότητα τυχαίων ζευγών μεταξύ των 2 ειδών δεδομένων.

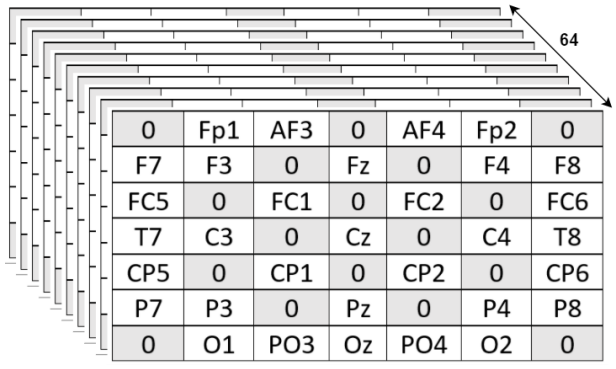

Figure 4: Η αρχιτεκτονική του ΗΕΓ σήματος εισόδου, που αναπαριστά τη φυσική τοπολογία των ηλεκτροδίων στον ανθρώπινο εγκέφαλο, σε ένα 9x9 πλέγμα.

Τα χαρακτηριστικά των σημάτων που θα αξιοποιήσουμε είναι σημαντικά, ούτως ώστε να μπορούμε να ανιχνεύσουμε συσχετίσεις ανάμεσα στα 2 είδη δεδομένων. Όσον αφορά το ΗΕΓ, αξιοποιούμε τη διαστατικότητα των 32 καναλιών ώστε να τα οργανώσουμε σε ένα

πλέγμα που να προσομοιάζει στην τοπολογία του τον ανθρώπινο εγκέφαλο, όπως φαίνεται στο Figure 4. Κατ' αυτό τον τρόπο μπορούμε να εξάγουμε τόσο χωρική όσο και χρονική πληροφορία για τα ΗΕΓ σήματα, χρησιμοποιώντας ένα 3D Συνελικτικό Δίκτυο 3 επιπέδων. Σχετικά με τα μουσικά σήματα, κι επειδή είναι περιορισμένα σε αριθμό, επιλέγουμε να εξάγουμε ενδιάμεσα χαρακτηριστικά μέσω μεταφοράς μάθησης (transfer learning) από το MusiCNN [118], ένα συνελικτικό δίκτυο, εκπαιδευμένο πάνω σε χιλιάδες μουσικά κομμάτια από μεγαλύτερα σύνολα. Τα 2 νευρωνικά δίκτυα εκπαιδεύονται αρχικά ξεχωριστά πάνω στις επισημειώσεις των δεδομένων τους και στη συνέχεια προσαρμόζονται ως σύνολο μέσω της προβολής των διανυσμάτων εξόδου σε ένα κοινό επίπεδο δικτύου, που αναπαριστά τον κοινό διανυσματικό χώρο (Figure 5). Κάθε δίκτυο εκπαιδεύεται ξεχωριστά για την πρόβλεψη του valence και του arousal, ως προβλήματα δυαδικής κατηγοριοποίησης. Όπως και στο προηγούμενο πείραμα, έτσι κι εδώ χρησιμοποιούμε την ενδιάμεση τιμή του εύρους των ετικετών ως το διαχωριστικό όριο σε high και low κλάσεις.

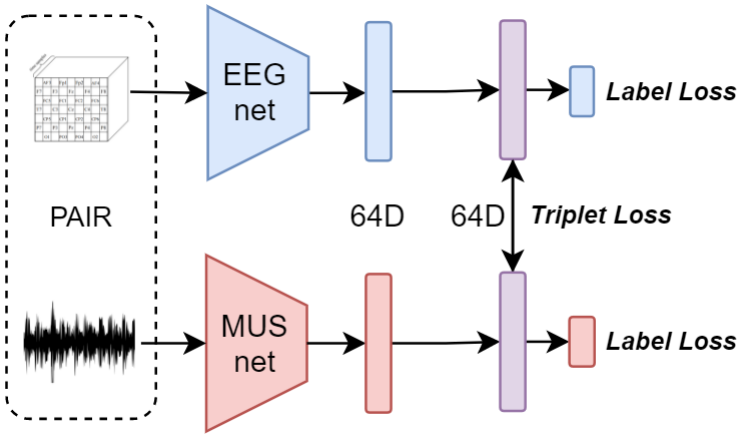

Figure 5: Το προτεινόμενο δίκτυο 2 κλάδων για τα ΗΕΓ και τα αντίστοιχα μουσικά σήματα.

Η Πολυτροπική Διαδικασία Μάθησης

Στόχος μας είναι να προσδιορίσουμε ένα κοινό χώρο στον οποίο τα δείγματα των 2 διαφορετικών ειδών δεδομένων να μπορούν να εμφανίζουν ομοιότητες σχετικά με την συναισθηματική τους πληροφορία. Χρησιμοποιούμε λοιπόν έναν συνδυασμό 4 συναρτήσεων σφάλματος για τον προσανατολισμό της διαδικασίας εκπαίδευσης. Οι πρώτοι 2 παράγοντες πρακτικά υπολογίζουν το σφάλμα των προβλέψεων του δικτύου για τα δύο είδη δεδομένων:

$$\mathcal{J}_1 = \lambda_{11}\mathrm{CE}_a + \lambda_{12}\mathrm{CE}_b$$

Για τη μείωση της απόστασης των όμοιων δεδομένων στον κοινό χώρο που δημιουργούμε, χρησιμοποιούμε 2 μετρικές συναρτήσεις τριών εισόδων. Πρακτικά, η πρώτη είσοδος καλείται οδηγός (anchor) και οι επόμενες είναι ένα δείγμα ίδιας και διαφορετικής κατηγορίας, αντίστοιχα. Οι συναρτήσεις αυτές μετρούν την απόσταση μεταξύ των δοθέντων σημείων (εδώ χρησιμοποιούμε απόσταση συνημιτόνου), επιχειρώντας να οδηγήσουν στη μείωση της απόστασης μεταξύ του οδηγού και του "θετικού" του δείγματος, ταυτόχρονα μεγαλώνοντας την απόσταση με το "αρνητικό" του δείγμα. Οι 2 συναρτήσεις δηλώνονται ως εξής:

$$\mathcal{J}_2 = \max\left(\cos(u_a - v_p) - \cos(u_a - u_n), 0\right)$$

$$\mathcal{J}_3 = \max\left(\cos(u_a - v_p) - \cos(u_a - v_n), 0\right)$$

ενώ η τελική συνάρτηση σφάλματος είναι ένας γραμμικός συνδυασμός των προαναφερθέντων:

$$\mathcal{J} = \mathcal{J}_1 + \lambda_2 \mathcal{J}_2 + \lambda_3 \mathcal{J}_3$$

Η κύρια πρόκληση που καλούμαστε να λύσουμε στο συγκεκριμένο πρόβλημα είναι η σημασιολογική απόσταση που επιβάλλει η διαφορετικότητα στο είδος των δεδομένων, και συγκεκριμένα όπως αυτή εκφράζεται στις επισημειώσεις τους. Τα 40 μουσικά κομμάτια του DEAP έχουν ανεξάρτητες επισημειώσεις, οι οποίες σε μερικές περιπτώσεις αντικρούουν τις μέσες βαθμολογίες που τους έχουν ανατεθεί στο πείραμα. Επιλέξαμε να αποσύρουμε από την ανάλυσή μας 6/40 κομμάτια με τις περισσότερες αναντιστοιχίες. Όσον αφορά τα ΗΕΓ, από το 1 λεπτό του πειράματος αποσύρουμε τα πρώτα 7 και τα τελευταία 3 δευτερόλεπτα για να αποφύγουμε το θόρυβο της αρχικής προσαρμογής αλλά και της κόπωσης, ενώ χωρίζουμε τα υπόλοιπα 50 δευτερόλεπτα σε 50 ανεξάρτητα μέρη του ενός δευτερολέπτου. Ακόμη, δεκαπλασιάζουμε την ποσότητα των δεδομένων κάνοντας επαύξηση μέσω θορύβου [162] ούτως ώστε να υποβοηθήσουμε την εκπαίδευση των βαθιών μοντέλων. Ως χαρακτηριστικά εισόδου επιλέγουμε το μονόπλευρο μέτρο του Διακριτού Μ/Σ Fourier του σήματος. Σε κάθε κομμάτι ΗΕΓ αντιστοιχίζεται το μουσικό κομμάτι που χρησιμοποιήθηκε στη συγκεκριμένη καταγραφή, του ίδιου δευτερολέπτου καθώς και των 2 αμέσως προηγούμενων.

Τα αποτελέσματα των πειραμάτων αξιολογούνται με την πρόβλεψη των συναισθηματικών επισημειώσεων, όπου χρησιμοποιούμε τη μετρική απλής ακρίβειας, καθώς και την εξαγωγή σχετικών μουσικών απαντήσεων σε ΗΕΓ εισόδους, όπου χρησιμοποιούμε 2 μετρικές αποστάσεων: Precision@10 (P@10) και mean Average Precision (mAP). Οι συγκεκριμένες μετρικές αξιολογούν την συσχέτιση μεταξύ ενός δείγματος - ερωτήματος εισόδου και των σχετικών του απαντήσεων, κατατάσσοντας την απόστασή του στον κοινό διανυσματικό χώρο με όλα τα διαθέσιμα (μουσικά) δείγματα. Η μεν πρώτη αξιολογεί τα 10 κοντινότερα δείγματα ενώ η δεύτερη το σύνολο της κατάταξης. Επίσης, τα αποτελέσματα παρουσιάζονται και συγκεντρωτικά ανά καταγραφή, παίρνοντας την πλειοψηφική πρόβλεψη των 50 τμημάτων κάθε κομματιού για την μετρική ακρίβειας και την ενδιάμεση τιμή για τις μετρικές κατάταξης.

Πειραματική Αξιολόγηση

Dimension	**Non-Aggregated**	**Aggregated**
Valence EEG	0.610 → 0.604	**0.633** → 0.632
Arousal EEG	0.645 → 0.641	0.645 → **0.662**
Valence MUS	0.680 → 0.646	**0.743** → 0.689
Arousal MUS	0.838 → 0.837	0.833 → **0.838**

Table 3: Αναγνώριση Συναισθήματος από τα προ-εκπαιδευμένα στα τελικά μοντέλα - μεσοσταθμικές τιμές από 32 διαφορετικά μοντέλα, ένα για κάθε συμμετέχοντα.

Για τα πειράματα ακολουθείται η ροή 1) μεμονωμένη εκπαίδευση του κλάδου ΗΕΓ 2) μεμονωμένη εκπαίδευση του κλάδου μουσικής 3) από κοινού εκπαίδευση προσαρμογής των 2 κλάδων. Παραπάνω φαίνονται τα αποτελέσματα της ακρίβειας στην πρόβλεψη συναισθήματος. Βλέπουμε πως για τη μουσική τα αποτελέσματα είναι αρκετά υψηλά, παρά το ότι δουλέψαμε μόνο με 34 κομμάτια. Αυτό υποδεικνύει την επιτυχή μεταφορά μάθησης που θα βοηθήσει στη συνέχεια και στην προσέγγιση των συναισθηματικών χαρακτηριστικών των δεδομένων. Από την άλλη, τα ΗΕΓ εμφανίζουν μεγάλη μεταβλητότητα ανά ξεχωριστό συμμετέχοντα, δίνοντας μεσοσταθμικά 63.3% πρόβλεψη valence και 64.5% arousal. Γενικώς, παρατηρούμε

πως τα συγκεντρωτικά νούμερα δίνουν μια πιο καθαρή αίσθηση του εκάστοτε συναισθήματος. Συνεπώς, προχωρώντας στην αξιολόγηση της από κοινού εκπαίδευσης, βλέπουμε μια ελαφρά μείωση στην ακρίβεια του valence, ενώ η πρόβλεψη του arousal ενισχύεται από την από κοινού μάθηση, με 2% βελτίωση στα ΗΕΓ και κρατώντας τα υψηλά ποσοστά στη μουσική.

Valence	**Accuracy**	**P@10**	**mAP**
Non-Aggregated	0.610 → 0.604	**0.617**	0.577
Aggregated	0.633 → 0.632	**0.659**	0.576
Arousal	**Accuracy**	**P@10**	**mAP**
Non-Aggregated	0.645 → 0.641	0.653	**0.674**
Aggregated	0.645 → 0.662	0.677	**0.679**

Table 4: Ακρίβεια ανάκτησης μουσικών κομματιών από ΗΕΓ εισόδους - μεσοσταθμικές τιμές.

Στο Table 4 φαίνονται επίσης τα αποτελέσματα για τις μετρικές κατάταξης. Παρά τη μικρή μείωση στην πρόβλεψη του valence, ο κοινός διανυσματικός χώρος αποδίδει μια πιο πιστή αναπαράσταση των συσχετίσεων και υποδεικνύει πως η πλειοψηφία των ΗΕΓ σημάτων θα μπορούσε να εξάγει συναισθηματικά συνεκτικές κατατάξεις μουσικών κομματιών. Για το valence παρατηρούμε επίσης μια σημαντική απόσταση μεταξύ των 2 μετρικών. Σε συνδυασμό με τις προηγούμενες παρατηρήσεις, φαίνεται πως ο κοινός χώρος για το valence κυριαρχείται από τοπικά clusters μεγάλης ομοιότητας, τα οποία πιάνει μόνο η P@10 μετρική. Για να επιβεβαιώσουμε αυτόν τον ισχυρισμό θα παραθέσουμε στη συνέχεια οπτικοποιήσεις των χώρων που προκύπτουν. Από την άλλη, το arousal δείχνει να αποδίδει πιο συνεκτικές αναπαραστάσεις, εξ ου και οι αυξημένες επιδόσεις σε όλες τις κατηγορίες, αλλά και στις γενικές κατατάξεις (mAP). Συγκεκριμένα, μέχρι και το 68%, μεσοσταθμικά, των ΗΕΓ εξάγουν συναισθηματικά κοντινά μουσικά κομμάτια από το δίκτυο.

Σχεδιάσαμε πολλά ακόμη πειράματα προκειμένου να δοκιμάσουμε και να επιβεβαιώσουμε την αποδοτικότητα των επιλογών μας στη μοντελοποίηση και εκμάθηση του συγκεκριμένου προβλήματος. Όπως φαίνεται παρακάτω, οι 4 συναρτήσεις σφάλματος αποδίδουν βέλτιστα όταν συνδυάζονται, με τις συναρτήσεις επίβλεψης να έχουν τη μεγαλύτερη επίδραση στα προκύπτοντα αποτελέσματα. Θετική επίδραση τουλάχιστον 4% φαίνεται να έχει η επαύξηση των δεδομένων μέσω θορύβου, καθώς τα δεδομένα είναι αριθμητικά περιορισμένα ώστε να υποστηρίξουν την εκπαίδευση ενός βαθιού δικτύου. Ακόμη, τα αποτελέσματα φαίνεται να χειροτερεύουν με την απουσία της αρχικής μεμονωμένης εκπαίδευσης των 2 δικτύων καθώς και με την απουσία της μεταφοράς μάθησης για τα μουσικά κομμάτια.

Ποιοτική Ανάλυση Αποτελεσμάτων

Από τη στιγμή που εκπαιδεύουμε ένα μοντέλο για κάθε διαφορετικό συμμετέχοντα, είναι λογικό τα αποτελέσματα να έχουν σημαντική μεταβλητότητα μεταξύ τους. Για να εξάγουμε χρήσιμα στοιχεία σχετικά με τη μουσική αντίληψη προσεγγίζουμε επιλεγμένες περιπτώσεις συμμετεχόντων και μουσικών κομματιών. Αρχικά, δίνουμε οπτικοποιημένα παραδείγματα του κοινού διανυσματικού χώρου των 2 ειδών δεδομένων, μέσω του αλγορίθμου t-SNE [155]. Όσον αφορά το valence (Figure 6a) είναι εμφανές, σε σύγκριση και με τις arousal αναπαραστάσεις, πως τείνουν να σχηματίζονται τοπικά συνεκτικοί υπόχωροι ανάμεσα στα δεδομένα και υπάρχει δυσκολία στην ομογενοποίησή τους. Από την άλλη, οι αναπραστάσεις για το arousal (Figure 6b) είναι αρκετά πιο συνεπείς σημασιολογικά, επιτυγχάνοντας σε μεγάλο βαθμό τη γεφύρωση του χάσματος μεταξύ ΗΕΓ και μουσικής πληροφορίας.

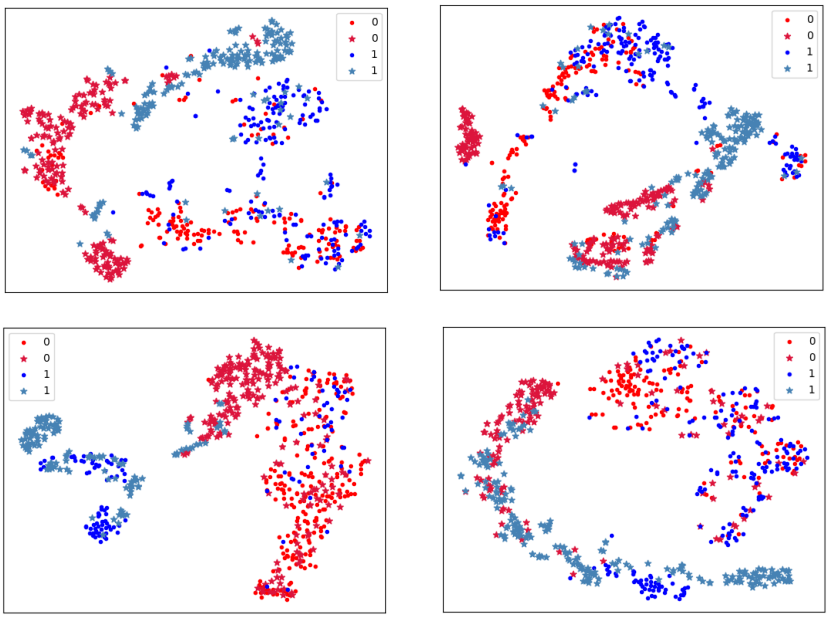

Figure 6: t-SNE αναπαράσταση του κοινού χώρου για τα test δείγματα 2 επιλεγμένων trials για a) Valence (πάνω) και b) Arousal (κάτω). Με τελείες δηλώνονται τα ΗΕΓ (έντονα χρώματα) και με αστερίσκους τα μουσικά δείγματα (άτονα χρώματα).

Παρακάτω (Figure 7) φαίνονται οι διακυμάνσεις της μετρικής mAP καθώς εξελίσσεται χρονικά ένα μουσικό κομμάτι, σε μεσοσταθμική προβολή όλων των συμμετεχόντων και μετά από φιλτράρισμα ομαλοποίησης. Είναι σαφές πως και στις 2 περιπτώσεις πειραμάτων υπάρχουν κοινές τάσεις όσον αφορά τη χρονική εξέλιξη του παραγόμενου συναισθήματος. Τα παρατηρούμενα μοτίβα εστιάζουν σε υψηλά επίπεδα αναγνώρισης κυρίως στην αρχή των κομματιών, ενώ φαίνεται να χτίζουν σταδιακά το παραγόμενο συναίσθημα. Επιπρόσθετα πειράματα απαιτούνται ωστόσο για την επιβεβαίωση της επικράτησης αυτών των μοτίβων.

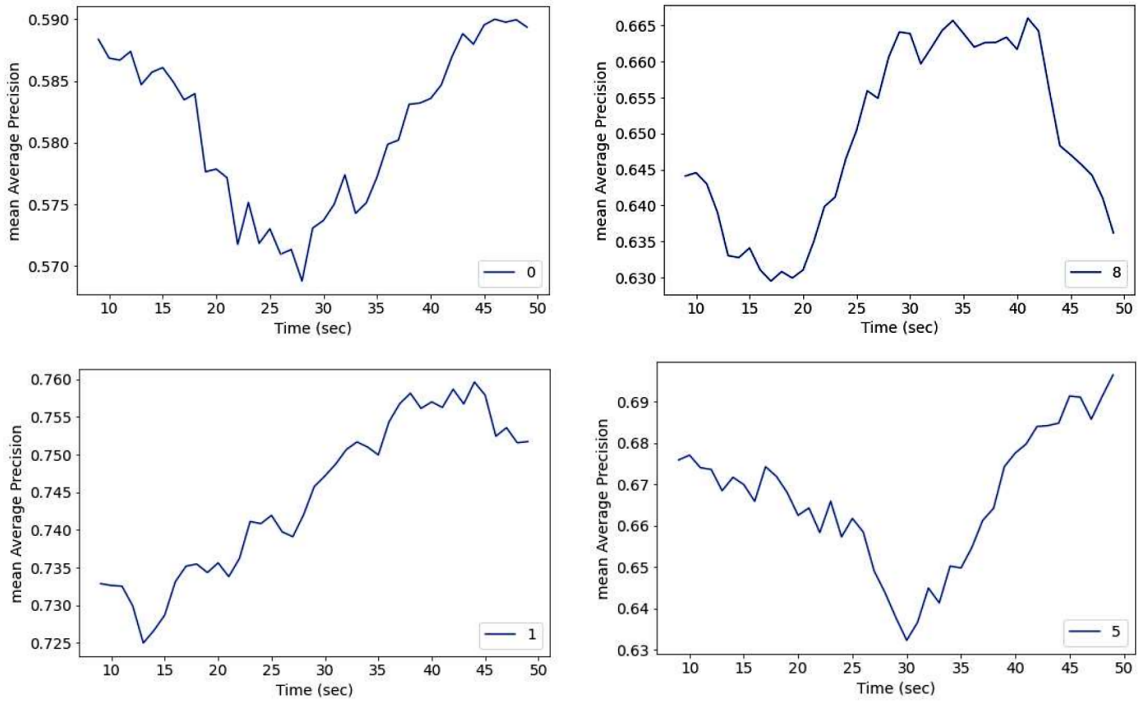

Figure 7: mAP τιμές για καθένα από τα 50 sec. επιλεγμένων κομματιών. Οι τιμές είναι μεσοσταθμικές των 32 συμμετεχόντων σε κάθε sec. ξεχωριστά και έχουν υποστεί median φιλτράρισμα. Κομμάτια 0,8: Απεικόνιση Valence, Κομμάτια 1,5: Απεικόνιση Arousal

Τέλος, στο Figure 8 παρουσιάζουμε και απεικονίσεις ενδιάμεσων επιπέδων του 3D συνελικτικού δικτύου ανάλυσης των ΗΕΓ σημάτων. Πέρα από την αποδοτικότητα στην εξαγωγή χρήσιμων χαρακτηριστικών, η δομή αυτή μας επιτρέπει να εξετάσουμε τις τοπολογικές περιοχές του εγκεφάλου που ενεργοποιούνται περισσότερο στη μουσική και συναισθηματική αντίληψη. Χρησιμοποιούμε την έξοδο του πρώτου επιπέδου του CNN που διατηρεί μια 5X5 τοπολογική δομή. Παρατηρούμε καθαρά πως τα περιφερειακά κανάλια ενεργοποιούνται περισσότερο σε όλες τις περιπτώσεις, ενώ υπάρχει μια μικρή τάση ενίσχυσης των πίσω-αριστερά περιοχών στο valence και των μπροστά-δεξιών περιοχών στο arousal. Οι παρατηρήσεις αυτές επιβεβαιώνονται για ένα σημαντικό αριθμό συμμετεχόντων.

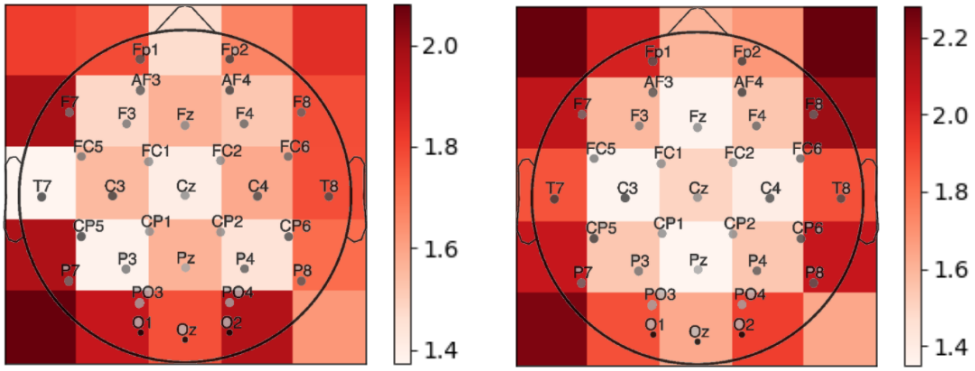

Figure 8: Οπτικοποίηση των βαρών του πρώτου επιπέδου του ΗΕΓ δικτύου για τον συμμετέχοντα 9, για Valence (αριστερά) και Arousal (δεξιά). Πρόκειται για μεσοσταθμικές τιμές των 50 sec.

Συμπεράσματα

Συμπερασματικά, η παρούσα εργασία επιχειρεί να αναλύσει και να παρουσιάσει στοιχεία για τη μουσική αντίληψη του ανθρώπου, από μια υπολογιστική σκοπιά και μπορεί να χωριστεί σε 2 μεγάλες υποενότητες: Στην πρώτη ασχολούμαστε με τη πολυπλοκότητα της μορφής των ΗΕΓ σημάτων και προτείνουμε αλγόριθμους πολυ-φράκταλ ανάλυσης προκειμένου να εξάγουμε χαρακτηριστικά που να συνδέονται με το συναίσθημα που προκαλεί το μουσικό άκουσμα. Οι προτεινόμενοι αλγόριθμοι αποδεικνύονται αποδοτικοί και τα αποτελέσματά τους είναι καλύτερα συγκριτικά με άλλους αλγορίθμους εξαγωγής χαρακτηριστικών που χρησιμοποιούνται συχνά στη βιβλιογραφία. Μέσα από τα πειράματα υποδεικνύουμε πως τα φράκταλ χαρακτηριστικά των ΗΕΓ σημάτων μπορούν να λαμβάνονται υπόψη στην Αναγνώριση Συναισθήματος, ιδίως όσον αφορά την μετρική arousal και τον άλφα εγκεφαλικό ρυθμό.

Στη δεύτερη ενότητα επεκτείνουμε τα εργαλεία μας και επιχειρούμε να εκπαιδεύσουμε βαθιά μοντέλα μάθησης με στόχο την εξεύρεση παραγόντων που να συσχετίζουν τα εγκεφαλικά σήματα με τα μουσικά ακουστικά ερεθίσματα. Κατασκευάζουμε ένα μοντέλο που δέχεται από κοινού δεδομένα ΗΕΓ και αντίστοιχα μουσικά ερεθίσματα με στόχο την εξεύρεση κοινών στοιχείων και την αναγνώριση του συναισθήματος, τόσο απευθείας, μέσω της πρόβλεψης των επισημειώσεων, όσο και εμμέσως, με την εξαγωγή σχετικών μουσικών κομματιών από ΗΕΓ εισόδους. Εφαρμόζοντας το προτεινόμενο μοντέλο ανεξάρτητα σε δεδομένα 32 διαφορετικών ανθρώπων, εξάγουμε ενδιαφέροντα μοτίβα σχετικά με την συσχέτιση μουσικής και εγκεφαλικής απόκρισης, τις εγκεφαλικές περιοχές που υπεισέρχονται σε αυτή την ανάλυση καθώς και τη χρονική μεταβλητότητα της συναισθηματικής έκφρασης.

Contents

Abstract - Περίληψη **6**

Contents **25**

List of Figures **27**

List of Tables **30**

List of Acronyms **31**

1 Introduction **33**
1.1 The Science of Emotion . . . 33
1.1.1 Theories of Emotion . . . 33
1.1.2 Defining Emotions . . . 34
1.1.3 Expressing and Perceiving Emotions . . . 35
1.2 The Human Brain . . . 37
1.2.1 Physiology of Neurons . . . 38
1.2.2 The Electroencephalogram (EEG) . . . 39
1.3 Music Perception . . . 41
1.3.1 The Auditory System . . . 41
1.3.2 Principles of Music Structure . . . 42
1.3.3 Haunted by Music . . . 43
1.4 Thesis Structure & Contributions . . . 44

2 Theoretical Background **46**
2.1 Signal Processing Fundamentals . . . 46
2.1.1 What is a Signal? . . . 46
2.1.2 Time - Frequency Representations . . . 46
2.1.3 Probability Theory and Statistics . . . 49
2.1.4 Fractal Signal Analysis . . . 52
2.2 Machine Learning Fundamentals . . . 54
2.2.1 Supervised Learning Algorithms . . . 54
2.2.2 Neural Networks and Optimization . . . 57
2.2.3 Training and Evaluation Issues . . . 60
2.3 Deep into Neural Networks . . . 62
2.3.1 From Feature Engineering to Deep Learning . . . 63
2.3.2 Convolutional Neural Networks . . . 63
2.3.3 Recurrent Neural Networks . . . 64
2.3.4 Multimodal and Metric Learning . . . 65

3 Multifractal Analysis on EEG **67**
3.1 Literature Review . . . 67
3.2 Multifractal Algorithms . . . 68
3.2.1 Multiscale Fractal Dimension . . . 68
3.2.2 Multifractal Detrended Fluctuation Analysis . . . 70
3.3 EEG as a Multifractal Signal . . . 71
3.3.1 Stationarity of EEG Signals . . . 71
3.3.2 Hurst Exponent Estimation . . . 72
3.4 Extraction of Fractal Features . . . 73
3.4.1 Baseline Features . . . 73
3.4.2 MFD Features . . . 74
3.4.3 MFDFA Features . . . 74
3.5 Experimental Evaluation . . . 75
3.5.1 Model Formulation . . . 75
3.5.2 Results & Discussion . . . 75

4 EEG & Music Cross-Modal Learning **79**
4.1 Literature Review . . . 79
4.1.1 Music Cognition . . . 79
4.1.2 Cross-Modal Learning . . . 80
4.2 Bridging the Semantic Gap . . . 81
4.2.1 Problem Formulation . . . 81
4.2.2 The Issue of Input Representations . . . 81
4.2.3 The Proposed Framework . . . 82
4.3 The Multimodal Training Procedure . . . 83
4.3.1 Optimization Methods . . . 83
4.3.2 Feature Extraction . . . 84
4.3.3 Evaluation Metrics . . . 86
4.4 Experimental Evaluation . . . 87
4.4.1 Predicting Emotion Tags . . . 87
4.4.2 Retrieving Tracks from EEG Queries . . . 87
4.4.3 Ablation Studies . . . 88
4.5 Qualitative Analysis . . . 89
4.5.1 The Common Latent Space . . . 89
4.5.2 Temporal Variation of Recognition . . . 91
4.5.3 Scalp Network Activations . . . 92

5 Conclusions **94**
5.1 EEG Affective Features . . . 94
5.2 Cross-Modal Learning . . . 95
5.3 Suggestions for Future Work . . . 95

Appendix **96**
DEAP Dataset . . . 96

List of Publications **100**

Bibliography **102**

List of Figures

1 Νευρωνικό Δίκτυο 5 επιπέδων, που περιέχει n εισόδους, 3 κρυφά επίπεδα και ένα επίπεδο n εξόδων. Είναι παράδειγμα ενός βαθιού νευρωνικού δικτύου (DNN). Πηγή: [18]. 14
2 Αναπαράσταση της Μορφολογικής Κάλυψης Minkowski για ένα δείγμα σήματος ΗΕΓ. 15
3 MFDFA σε δείγμα ΗΕΓ: Απεικονίζουμε 16 DFA γραμμικές αναπαραστάσεις της $F_q(s)$ μαζί με τη γραμμική πρόβλεψη της της κλίσης που προσδιορίζει τον γενικευμένο Hurst εκθέτη $H(q)$. 16
4 Η αρχιτεκτονική του ΗΕΓ σήματος εισόδου, που αναπαριστά τη φυσική τοπολογία των ηλεκτροδίων στον ανθρώπινο εγκέφαλο, σε ένα 9x9 πλέγμα. 19
5 Το προτεινόμενο δίκτυο 2 κλάδων για τα ΗΕΓ και τα αντίστοιχα μουσικά σήματα. 20
6 t-SNE αναπαράσταση του κοινού χώρου για τα test δείγματα 2 επιλεγμένων trials για a) Valence (πάνω) και b) Arousal (κάτω). Με τελείες δηλώνονται τα ΗΕΓ (έντονα χρώματα) και με αστερίσκους τα μουσικά δείγματα (άτονα χρώματα). 23
7 mAP τιμές για καθένα από τα 50 sec. επιλεγμένων κομματιών. Οι τιμές είναι μεσοσταθμικές των 32 συμμετεχόντων σε κάθε sec. ξεχωριστά και έχουν υποστεί median φιλτράρισμα. Κομμάτια 0,8: Απεικόνιση Valence, Κομμάτια 1,5: Απεικόνιση Arousal 23
8 Οπτικοποίηση των βαρών του πρώτου επιπέδου του ΗΕΓ δικτύου για τον συμμετέχοντα 9, για Valence (αριστερά) και Arousal (δεξιά). Πρόκειται για μεσοσταθμικές τιμές των 50 sec. 24

1.1 [22] Plutchik's Emotion Wheel. 34
1.2 [173] The Valence-Arousal Space. 34
1.3 [89] The 6 (+neutral) basic emotion categories, according to Paul Eckman. 34
1.4 Heatmap Visualization of the affect in text. Source: Baziotis et al. [14] . . 36
1.5 Illustration of the white matter fiber architecture of the brain, measured from diffusion spectrum imaging (DSI). Shown are the corpus callosum, cerebellum, and others. Source: https://humanconnectomeproject.org. . . 38
1.6 The structure of a neuron (adopted from Attwood and MacKay [5]). 39
1.7 Schematic diagram of an EEG recording experiment. Source: [101] 40
1.8 Four typical dominant brain normal rhythms. Source: [131] 41
1.9 Subdivisions of notes' duration. Source: learnpianoforfree.weebly.com . . . 42

2.1 An example STFT spectrogram, extracted from a music signal (violin). . . 48
2.2 Example types of wavelets. Source: reference.wolfram.com/language 49

2.3 Visualization of 3 noisy signals with fundamental frequencies at 100, 300 and 800 Hz. Depicted are their STFT spectrogram (left) and Power Spectral Density (right). Source: https://ccrma.stanford.edu/ jcaceres/yamaha/ 51
2.4 Left to right: the Mandelbrot Set, the Koch's Snowflake and the Herpinski Triangle. 52
2.5 Visual Inspection of $N = r^D$ for $r = 1, 2$. Source: [137] 53
2.6 Estimating the box-counting dimension of the coast of Great Britain [YouTube]. 53
2.7 Visual examples of a linear regression implementation that will be used in Detrended Fluctuation Analysis (DFA) algorithm (Chapter 3), to determine measures regarding the EEG complexity. Here we depict the DFA result for two sample 30-sec. EEGs. 55
2.8 Illustration of a Linear SVM functionality. Source: [149] 56
2.9 The Rosenblatt's Perceptron [127]. 58
2.10 A 5-layer Neural Network containing n inputs, 3 hidden layers and an output layer with n outputs. It is an example of a Deep Neural Network (DNN). Source: [18]. 59
2.11 The Cross-Validation concept diagram for $k = 5$. Source: towardsdatascience.com 60
2.12 Confusion Matrix of a binary classification experiment, along with the definitions of most common metrics (Precision, Recall / sensitivity). From manisha-sirsat.blogspot.com 62
2.13 Kernels for Edge Detection: Roberts (R_1, R_2), Prewitt (P_1, P_2), Sobel (S_1, S_2) 63
2.14 The architecture of an example VGG-16 [140] network for Image Analysis. 64
2.15 Diagram of an RNN cell's structure. 64
2.16 Structure of LSTM (left) and GRU cells (right). Red: Sigmoid, Blue: Tanh. From towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation 65
2.17 Concept of a contrastive objective on multimodal output embeddings. Source: laboratoirehubertcurien.univ-st-etienne.fr/en/teams/data-intelligence . . . 66

3.1 Minkowski cover simulation of a sample EEG signal from DEAP. At the up-left of both diagrams we depict the raw EEG sample waveform. 68
3.2 Fractogram of a sample EEG trial (Subject 8, Track 6, Channel CP1, α band, 0-15, 15-30 and 30-45 sec. windows). The utilized configuration is the one described in Section 3.4.2. 69
3.3 MFDFA on an EEG: depicting 16 DFA graphs for $F_q(s)$ along with the linear regression lines, the slopes of which determine the generalized Hurst Exponent $H(q)$. 70
3.4 Graphical examples of fractal time series. The upper graphs represent fractional Brownian motions (fBm) and the lower graphs, the corresponding fractional Gaussian noises (fGn), for three typical values of the Hurst Exponent H: $0.25, 0.5, 0.75$. Source: [35] 72
3.5 The configuration of the selected left (blue) and right (red) DEAP channels. 73
3.6 Sample MFD profiles of 2 EEG signals (Subjects 5, 20) along with the mean and standard deviation features extracted from their 7 subsignals. . . 74
3.7 $t(q)$ and $D(q)$ MFDFA components for a sample EEG signal. 75

4.1 Concept of the proposed model: By using EEG data of music listening we attempt to derive embeddings that could resemble the stimulus and the music-induced affect. 81
4.2 EEG input shape that resembles the channel topology on the scalp. 81
4.3 Framework Architecture: a) EEG net Architecture b) Music net Architecture. 82
4.4 The proposed bi-stream network. The 2D dense layers shown in Figure 4.3 are substituted by 64D dense layers and are then connected to the 64D common space. 83
4.5 The function of triplet losses. Here the arrows correspond to cosine distances. A: anchor, P: positive sample, N: negative sample. 83
4.6 Feature vectors of 2 sample EEG trials at different seconds, for channel Fp1. Up: Subject 8 at the 20th, 30th and 40th seconds (left to right). Down: Subject 8 (same time). 85
4.7 Two ranking examples to clarify the usage of the information retrieval metrics P@10 and mAP. Adapted from slides of Rada Mihalcea: web.eecs.umich.edu/mihalcea 86
4.8 Per-Subject Accuracy after the fine-tuning Sessions. 89
4.9 Latent Space t-SNE visualisation for the test data of subjects for Valence. Dots denote EEG samples (bright colors) while asterisks denote music samples (dim colors). 90
4.10 Latent Space t-SNE visualisation for the test data of subjects for Arousal. Dots denote EEG samples (bright colors) while asterisks denote music samples (dim colors). 90
4.11 Valence mAP scores over the 50 time samples for specific numbered tracks. The scores have been averaged across all 32 subjects, for each time sample separately. 91
4.12 Arousal mAP scores over the 50 time samples for specific numbered tracks. The scores have been averaged across all 32 subjects, for each time sample separately. 91
4.13 Post-ReLU Activation of the first CNN block for subjects 5, 9 in Valence. We present the activation averaged on time and feature axes. 92
4.14 Post-ReLU Activation of the first CNN block for subjects 5, 9 in Arousal. We present the activation averaged on time and feature axes. 92

5.1 [65] Online Assessment Ratings. Selected 40 Videos are highlighted in green. 97
5.2 [65] The mean locations of the stimuli on the arousal/valence plane. Liking is encoded by color: dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: small symbols stand for low dominance and big for high dominance. 99

List of Tables

1 Ακρίβεια για τα Subject Dependent πειράματα στη μορφή: Valence — Arousal 18
2 Ακρίβεια για τα Subject Independent πειράματα στη μορφή: Valence — Arousal 18
3 Αναγνώριση Συναισθήματος από τα προ-εκπαιδευμένα στα τελικά μοντέλα - μεσοσταθμικές τιμές από 32 διαφορετικά μοντέλα, ένα για κάθε συμμετέχοντα. 21
4 Ακρίβεια ανάκτησης μουσικών κομματιών από ΗΕΓ εισόδους - μεσοσταθμικές τιμές. 22

3.1 Hurst Exponent of 27 randomly sampled DEAP EEG trials, computed through monofractal DFA and averaged, with respect to EEG channels and the number of trials. 73
3.2 Subject Dependent Task Accuracy in the form: Valence — Arousal 76
3.3 Subject Independent Task Accuracy in the form: Valence — Arousal . . . 76
3.4 Performance Accuracy at the Subject Dependent Setting for various feature types, presented in [80]. Importantly, features that aggregate the inter-channel correlations seem to capture the most discriminating emotional information. 77
3.5 MFDFA on signals' cumulative sum Valence–Arousal Accuracy 77
3.6 MFD-HFD Combined Arousal Accuracy 78

4.1 Inconsistent Stimuli of the DEAP Dataset and how we handle them. 85
4.2 Emotion Prediction from pre-trained to fine-tuned models - means over 32 subjects. 87
4.3 Retrieval Scores from fine-tuned models - mean values over 32 subjects. . . 88
4.4 Ablation on the utilized Objective Function on Valence (aggregated scores). 88
4.5 Ablation on the utilized Objective Function on Arousal (aggregated scores). 88
4.6 Ablation on critical choices in building the EEG model. We depict accuracy scores in their aggregated format and after the fine-tuning Sessions. Form: Valence – Arousal 89

List of Acronyms

ΗΕΓ	Ηλεκτροεγκεφαλογράφημα
Μ/Σ	Μετασχηματισμός
ΨΕΣ	Ψηφιακή Επεξεργασία Σήματος
AI	Artificial Intelligence
ADF	Augmented Dickey-Fuller (Test)
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
AP	Action Potential
CCA	Canonical Correlation Analysis
CDF	Cumulative Distribution Function
CE	Cross Entropy
CNS	Central Nervous System
CWT	Continuous Wavelet Transform
DCCA	Deep Canonical Correlation Analysis
DFA	Detrended Fluctuation Analysis
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DSI	Diffusion Spectrum Imaging
DSP	Digital Signal Processing
DWT	Discrete Wavelet Transform
DXA	Detrended Cross-Correlation Analysis
ECG	Electrocardiogram
EEG	Electroencephalogram
EMD	Empirical Mode Decomposition
EMG	Electromyogram
EOG	Electrooculogram
ERP	Event-Related Potential
ERS	Event-Related Synchronization
ESA	Energy Separation Algorithm
fBm	fractional Brownian motion
FER	Facial Expression Recognition
FFNN	Feed-Forward Neural Network
FFT	Fast Fourier Transform
fGn	fractional Gaussian noise
fMRI	functional Magnetic Resonance Imaging
GPU	Graphical Processing Unit
GRU	Gated Recurrent Unit
GSR	Galvanic Skin Response

HDF	Higuchi Fractal Dimension
HMM	Hidden Markov Model
IMF	Intrinsic Mode Function
LRAP	Label Ranking Average Precision
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MER	Music Emotion Recognition
MFD	Multiscale Fractal Dimension
MFDFA	Multifractal Detrended Fluctuation Analysis
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NLP	Natural Language Processing
PDF	Probability Density Function
PSD	Power Spectral Density
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
RMS	Root Mean Square
RNN	Recurrent Neural Network
RSP	Respiration
SAM	Self-Assessment Manikin
SER	Speech Emotion Recognition
SIFT	Scale-Invariant Feature Transform
STFT	Short-Time Fourier Transform
SURF	Speeded-Up Robust Features
STG	Superior Temporal Gyrus
SVM	Support Vector Machine
VCCA	Variational Canonical Correlation Analysis

Chapter 1

Introduction

1.1 The Science of Emotion

As we move through our daily lives, we experience a variety of emotions (happiness, surprise, disappointment, anger etc.). An emotion is a subjective state of being that we often describe as our feelings. The words *emotion* and *mood* are sometimes used interchangeably, but psychologists use these words to refer to two distinct things. Typically, the word emotion indicates a subjective, affective state that is relatively intense and occurs in response to an external stimulus. Mood, on the other hand, refers to the prolonged, less intense, affective state that does not occur in response to something we experience, and may as well not be consciously recognized [15].

1.1.1 Theories of Emotion

Our emotional states are comprised of combinations of 3 components: physiological arousal, psychological appraisal and subjective experiences. Therefore, different people may have variable emotional experiences even when faced with similar circumstances. Over time, several theories of emotion have been proposed to explain how the various components of emotion interact. The *James-Lange Theory of Emotion* [70] asserts that emotions arise from physiological arousal. For instance, if you were to encounter some threat in your environment, such as a robber in your home, your sympathetic nervous system would initiate significant physiological arousal, which would make your heart race and increase your respiration rate. According to the James-Lange theory, you would only experience a feeling of fear after this physiological arousal had taken place. Furthermore, different arousal patterns would be associated with different feelings.

Other theorists, however, doubted that physiological arousal is distinct enough to result in the wide variety of emotions that we know. According to the *Cannon-Bard Theory of Emotion*, physiological arousal and emotional experience occur simultaneously, yet independently [20, 69]. These and other theories have each garnered empirical support [27]; however, more recent studies [34] suggest that physiological arousal does not seem to be necessary for the emotional experience, but this arousal does appear to be involved in enhancing the intensity of the emotional experience. The *Schachter-Singer* [133] theory is another variation that takes into account both physiological arousal and the emotional experience. According to this theory, emotions are composed of physiological and cognitive factors that, in context, produce the emotional experience. In any case, the takeout is that studying physiological signals and responses to external stimuli is one of the most prominent methods to properly study the nature of induced emotions.

1.1.2 Defining Emotions

Cognitive Science, Neuroscience and Psychology have developed two main approaches for describing how humans perceive and classify emotion: *continuous* and *categorical*. The continuous approach tends to use dimensions such as negative/positive, calm/aroused etc. A representative example here is the Plutchik's emotion wheel [114], shown in Figure 1.1. However, the most widely used framework in this category is the Valence-Arousal Protocol, proposed by James Russel (1980) [128]. In this scale, each emotional state can be placed on a two-dimensional plane with arousal and valence as the horizontal and vertical axes. While arousal and valence explain most of the variation in emotional states, sometimes a third dimension of dominance is also included. Arousal can range from inactive (e.g. uninterested, bored) to active (e.g. alert, excited), whereas valence ranges from unpleasant (e.g. sad, stressed) to pleasant (e.g. happy, elated). Dominance ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything).

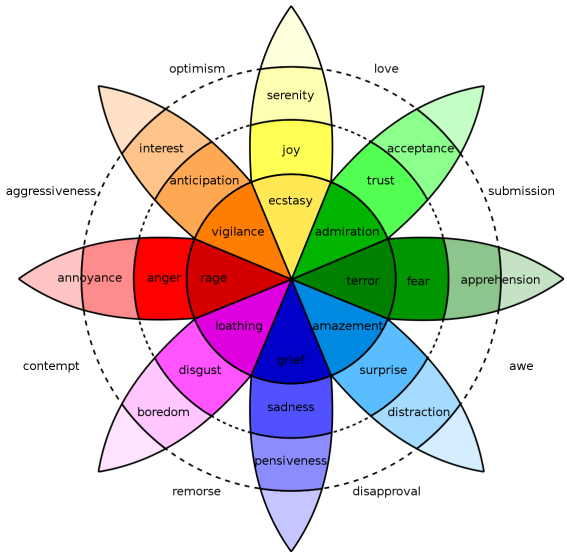

Figure 1.1: [22] Plutchik's Emotion Wheel.

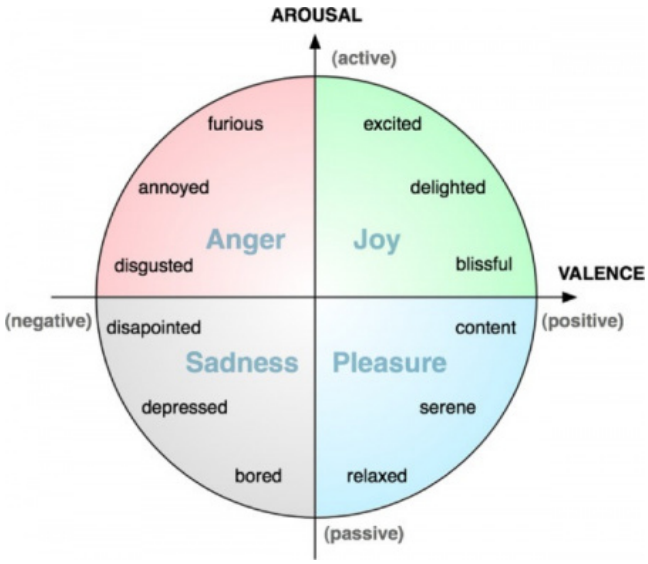

Figure 1.2: [173] The Valence-Arousal Space.

The categorical approach tends to use discrete classes to define emotion. During the 1970s, psychologist Paul Ekman identified six basic emotions that he suggested were universally experienced in all human cultures [40]. The emotions he identified were happiness, sadness, disgust, fear, surprise, and anger. He later expanded his list of basic emotions to include such things as pride, shame, embarrassment, and others, however those first 6 emotions, sometimes along with a 7th neutral class, have been widely utilized in psychological and computational experiments, either through explicit annotation, or combined labeling across categories, i.e. a happy-surprised face or a fearful-surprised one.

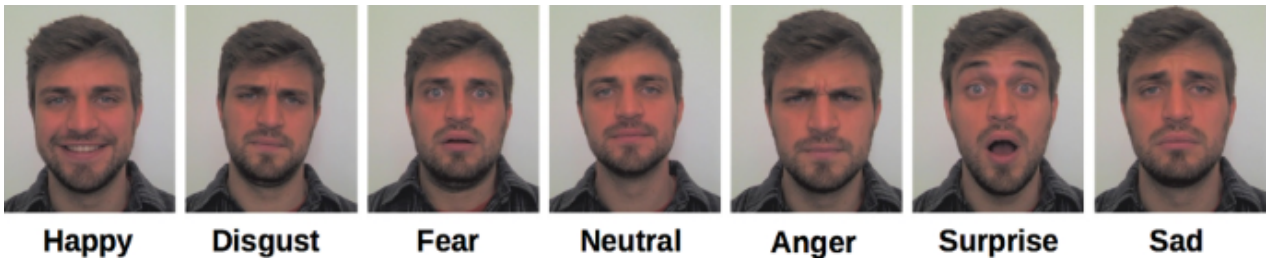

Figure 1.3: [89] The 6 (+neutral) basic emotion categories, according to Paul Eckman.

1.1.3 Expressing and Perceiving Emotions

In this context, our goal is to build computational models that could recognize the affective state, based on a variety of possibly available descriptors. With the term *Affective Computing* we refer to the study and development of systems and devices that can recognize, interpret, process, and simulate human affects [151], with the ultimate goal to decode and make good use of human's emotional intelligence. The computational and learning tools we use to this end will be analyzed in detail in Chapter 2. These tools work on expressive emotion descriptors that we can detect everywhere. Specifically, recognizing emotional information requires the extraction of meaningful patterns from various forms of gathered data. Data collection usually begins with passive sensors that capture data about the user's physical state or behavior, in an analogous way to how we also perceive emotions in others (camera, microphone, EEG etc.). Below we mention the most prominent modalities that are considered for emotion perception.

Facial Expressions

Faces are a ubiquitous part of the human life, affecting it immediately after birth. People greet each other with smiles or nods, have face-to-face conversations on a daily basis, capture their faces with smartphones and tablets and exchange photos of each other on social-media platforms. Our face is a predominant medium of communicating our behavior and particular facial gestures are actually correlated with specific emotion classes (Figure 1.3). As a result, various facial expression recognition (FER) studies have been conducted in the fields of Affective Computing and Computer Vision.

FER systems can be divided into two main categories according to the feature representations: *static image* FER and *dynamic sequence* FER [76]. In static-based methods [81, 100] the feature representation is encoded with only spatial information from a single image, whereas dynamic-based methods [60, 176] also consider the temporal relation among video frames. The majority of the traditional methods have used handcrafted features (e.g., local binary patterns [138], optical flow [29], face action landmarks [36]).

However, since 2014, the community has collected relatively sufficient training data from challenging real-world settings. Additionally, due to the increased chip processing abilities and well-designed model architectures, studies have begun to transfer to deep learning methods, which have achieved state-of-the-art recognition performance [140]. However, there is a large debate on whether our faces actually reflect our true affective states. Some studies show strong evidence on the universality of emotions in specific contexts [33] while others remain critical and emphasize the use of additional modalities (speech, context) to correctly identify the affective state [11]. Based on these views, other modalities, such as audio and biosignals, have also been used in *multimodal systems* [31].

Natural Language

The core of human communication and emotion expression relies on the use of our language, either written or spoken. With the term *Natural Language* we refer to both speech and text modalities that we use to this end and are crucial in expressing and identifying affect. Moreover, language data are more easily recorded than facial expressions, since only a microphone or a typewriter is required. *Speech Emotion Recognition* (SER) systems are used in several applications to enhance Human-Computer Interaction, such as speech synthesis, customer service, forensics and medical analysis [7]. SER is achieved by

the development of methodologies based on Digital Signal Processing and Machine Learning. Research here dates three decades back, however the results are still not applicable in large-scale everyday settings. A speech signal is a complex modality that contains lexical information, speaker-dependent vocal parameters (such as the fundamental frequency, the formants and the prosody of the signal), the elicited emotion and the utilized language, so if one has to recognize particular information in speech, then ideally they should generalize upon all these parameters. Before the prevail of Deep Learning, a large variety of features were used to extract emotion semantics from speech. Examples include statistical measures [72], energy features [9], non-linear and spectral transformations [68], usually processed sequentially by Hidden Markov Models (HMM) [121]. Modern systems though can efficiently analyze raw speech data through deep learning models [175].

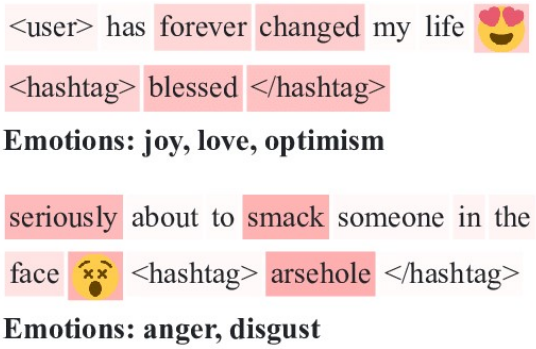

Figure 1.4: Heatmap Visualization of the affect in text. Source: Baziotis et al. [14]

Written language, on the other hand, requires a fairly different processing framework and research has been conducted in the field of *Natural Language Processing* (NLP), usually denoted as *Sentiment Analysis*. Since it is not in the scope of our study, we shall briefly survey the research directions in this field. In terms of data, there are a few annotated datasets for the task (e.g., ISEAR [135], SemEval), however there is a vast amount of unstructured data, like social media posts and opinion articles.

Approaches in literature are generally either rule-based or based on Machine Learning (or hybrid). The rule-based approach outlines major grammatical and logical rules to follow in order to detect emotions, which is insufficient though for large amounts of data. The rule construction approach encompasses keyword recognition (use of dictionaries) [148] and lexical affinity [2]. The Machine Learning approach classifies texts into affective categories using supervised or unsupervised Learning algorithms and has offered comparatively better detection rates. Recently, deep learning models are being adopted as approaches to detect emotions from segments of text, because they are more robust, incorporate vast amounts of data and can extract the intrinsic details text may carry. Most deep models utilize recurrent networks [163] and attention [122] to emphasize critical parts in texts. Applications of Sentiment Analysis include conversation monitoring [119], business-customer interaction, multimedia tagging and more.

Audio and Music

Audio signals that carry emotional information are generally grouped into speech and music, since naturalistic audio data are generally not used for this purpose. While borrowing methodology from Speech Analysis, *Music Emotion Recognition* (MER) is a standalone field that has provided cutting-edge research in the latest years. MER lies in the intersection of Emotion Analysis, which we discuss here, and Music Information Retrieval, a field that drives music-related research in various tasks, e.g., Genre Recognition [44], Automatic Transcription [124], Instrument Classification [66] and more.

MER is a critical task of Music Information Retrieval as well, since, as we will discuss later in this chapter, music is one of the most powerful ways to express and induce emotions. In this thesis, we will analyze emotion induction both by relevant musical features and physiological human responses to music listening. However, most research in MER deals with the first task, while neuroscience and psychology are utilizing computational methods to approach the second. Up to this day, several research works have identified possible correlations between specific musical elements and emotions. One of the most widely accepted is *mode*: major modes are frequently related to positive emotional states, whereas minor modes are often associated with sadness [129]. Other elements include tempo, articulation, timbre, pitch, tonality, rhythm, loudness or more sophisticated, like vibrato [108, 107]. Of course, contemporary methods also incorporate Deep Learning models to automatically extract meaningful features for the task [82]. Most studies in MER use audio data, however some prefer to analyze symbolic data of music transcription (e.g., MIDI files) or even combine and associate these modalities [106]. Emotion detection from music has many emerging application domains, such as in tagging and recommendation systems (e.g., Spotify), while it could also serve as a means to analyze human emotional responses and foster therapeutic methods.

Biomedical Signals

All the aforementioned studies use behavioral signals to determine emotion, having the advantage of easy large-scale data collection. However, the reliability of this approach can't be guaranteed, as the perception of emotions could be highly subjective in some cases and there is not yet a universal guide into specific indicators for specific emotions. Moreover, it is relatively easy for people to control their behavior in order to hide their real emotions, particularly during special social interactions. On the contrary, physiological measurements are induced without our active interference and could thus depict more clearly the actual affective state [24]. Apart from recognition purposes, physiological signals can also be utilized for studying the nature of human emotional responses and our nervous system, that has invaluable importance for Psychology and Medicine studies.

The researched physiological signals are induced mainly by the *Central Nervous System* (CNS) and *Autonomic Nervous System* (ANS), which is why they are largely involuntarily activated and therefore cannot be easily controlled. Examples of such signals are electroencephalogram (EEG), temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP), etc. For a comprehensive review for each one of these modalities, the reader can refer to [139]. The tools that are used for this kind of research are Signal Processing and Statistical Tests or Learning algorithms to handle their properties. Physiological signals, however, inherit a large amount of noise, both from the recording equipment and other interfering physiological processes, whereas their extraction is an expensive procedure as well. In contrast to behavioral signals, biosignals also require extensive domain knowledge to be properly analyzed.

1.2 The Human Brain

Emotion expression and perception are advanced forms of human cognition, governed by remarkable functions of the human brain, the most complex organ in our bodies. Weighting less than 1.5 kg, this jelly-like organ is the seat of intelligence, interpreter of the senses, initiator of the body movement and controller of our behavior. From an

anatomical point of view the brain may be divided into 3 parts: cerebrum, cerebellum, and brain stem. The *cerebrum* consists of both left and right lobes of the brain with highly convoluted surface layers, called the *cerebral cortex*. The *cerebrum* includes the regions for movement initiation, conscious awareness of sensation, expression of emotions and behaviour. The *cerebellum* coordinates voluntary movements of muscles and maintains balance. The brain stem, on the other hand, controls involuntary functions such as respiration, heart regulation and biorhythms. For centuries, science and philosophy have been amazed by the functional complexity of the brain and, until today, it remains largely incomprehensible. As technology and analysis tools are improving though, the pace of research in neuroscience and behavioral sciences is accelerating. Indicatively, the American Congress named the 1990s as the *Decade of the Brain*. Nowadays, powerful computing capabilities and advances in Artificial Intelligence have made it possible to research the brain functionality from a whole new, computational perspective [152], from which we should expect important breakthroughs in the near future.

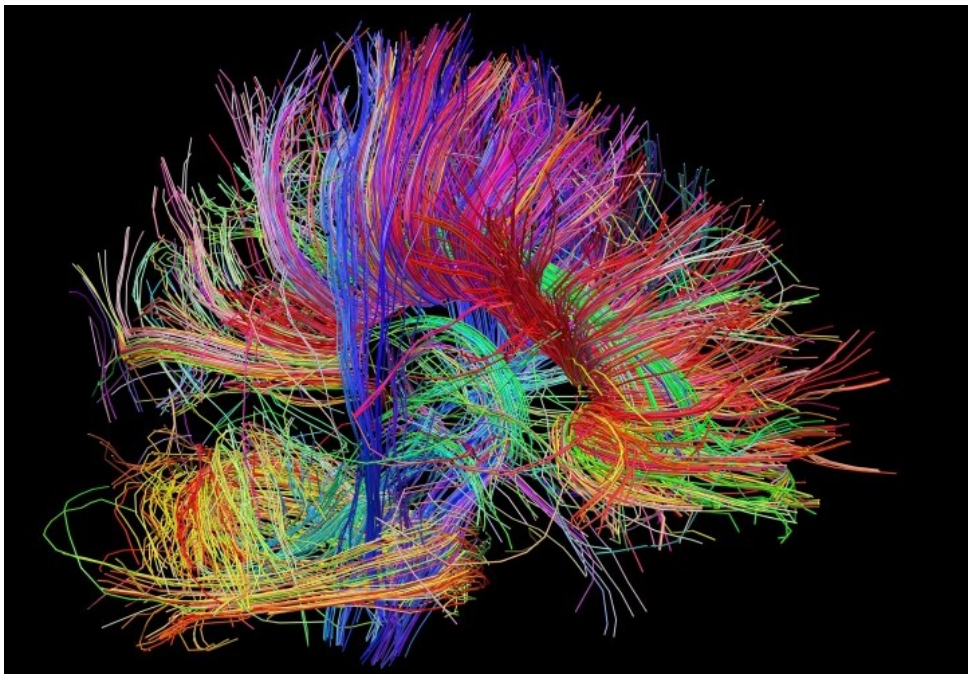

Figure 1.5: Illustration of the white matter fiber architecture of the brain, measured from diffusion spectrum imaging (DSI). Shown are the corpus callosum, cerebellum, and others. Source: https://humanconnectomeproject.org.

1.2.1 Physiology of Neurons

Like all parts of our body, our brain is made up of cells, many of which help regulate the chemistry and structure of the organ. Some cells, called neuronal cells, or just neurons, are specialized to do far more. In specific, they are responsible of much of the work needed for us to think, feel and move. A human brain contains an astonishing number of at least 90 billion neurons [5] that connect through spider-like arms (Figure 1.5) and communicate through electrochemical signals. In particular, each of these neurons may connect to at least 1000 other neurons and, in total, the human brain is estimated to have more than 100 trillion connections! Neurons are the building blocks of the nervous system, in which the brain can be seen as its hub.

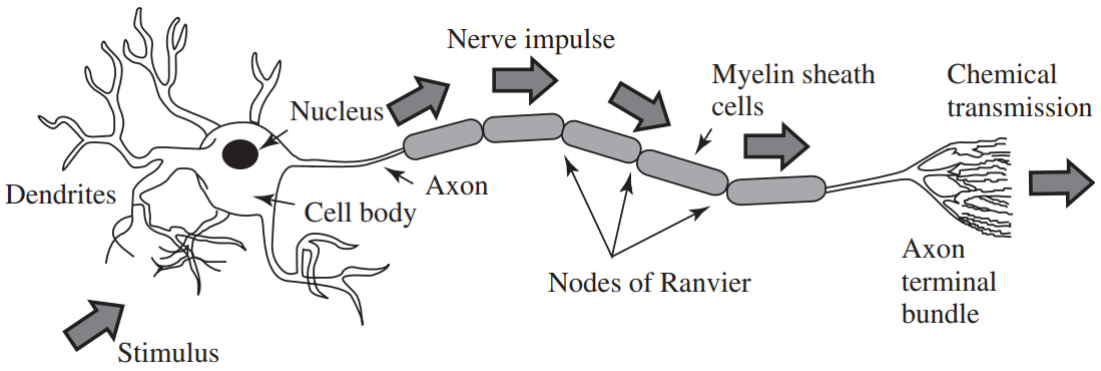

Figure 1.6: The structure of a neuron (adopted from Attwood and MacKay [5]).

As seen in Figure 1.6, a neuron, or nerve cell, consists of axons, dendrites, and cell bodies. A nerve cell body has a single nucleus and contains most of the nerve cell metabolism, especially that related to protein synthesis. The proteins created in the cell body are delivered to other parts of the nerve. An axon is a long cylinder, which transmits an electrical impulse and can be really long (several meters in some animals). In humans the length can be a percentage of a millimetre to more than a metre. Dendrites are connected to either the axons or dendrites of other cells and receive impulses from other nerves or relay the signals to other nerves. In the human brain, each nerve is connected to thousands of other nerves, mostly through dendritic connections. The activities in the CNS are mainly related to the synaptic currents transferred between these connections (called synapses). A potential of 60–70 mV with negative polarity may be recorded under the membrane of the cell body and, if it travels along the fibre to a synapse, a post-synaptic potential occurs in the following neuron. If several action potentials travel simultaneously, there will be a summation of post-synaptic potentials, producing an action potential on the postsynaptic neuron, provided that a certain threshold of membrane potential is reached. This functionality will later lead us to the modeling of perceptrons as artificial neurons. The information transmitted by a nerve is called an action potential (AP). APs are caused by a temporary exchange of ions across the neuron membrane and is transmitted along the axon. It is usually initiated in the cell and lasts between 5 and 10 msec.

1.2.2 The Electroencephalogram (EEG)

The neural activity of the human brain starts between the 17th and 23rd week of prenatal development. It is believed that from this early stage and throughout life electrical signals generated by the brain represent not only the brain function but also the status of the whole body. Understanding of the neural processes and neurophysiological properties of the brain and the mechanisms underlying the generation of these biosignals is, therefore, vital for those who detect and analyze brain functions.

Carlo Matteucci (1811–1868) and Emil Du Bois-Reymond (1818–1896) were the first people to register the electrical signals emitted from muscle nerves using a galvanometer and established the concept of neurophysiology [171]. Richard Caton (1842–1926), a scientist from Liverpool, placed two electrodes over the scalp of a human subject and thereby first recorded brain activity in the form of electrical signals in 1875. Since then, the concepts of electro-encephalo-gram were combined so that the term EEG was henceforth used to denote electrical neural activity of the brain. The history of EEG has been continuous and has brought daily development of clinical and computational studies for

discovery, diagnosis, and treatment of a vast number of physiological abnormalities of the brain and the rest of our CNS. Nowadays, EEGs are recorded digitally, using many delicate electrodes/channels, a set of differential amplifiers (one for each channel) and filters. Fortunately, the effective bandwidth for EEG signals is limited to 100 Hz so sampling is easy. Regarding electrode placement, the International Federation of Societies for Electroencephalography and Clinical Neurophysiology has recommended the conventional electrode setting [58]. According to this, the closest electrode to each ear should be located at 10% the distance between the two ears, and all electrodes should have an equidistance of 20% the same distance. This set-up is commonly called the 10-20 placement system. In special cases only, like in brain computer interfaces, a single channel may be used.

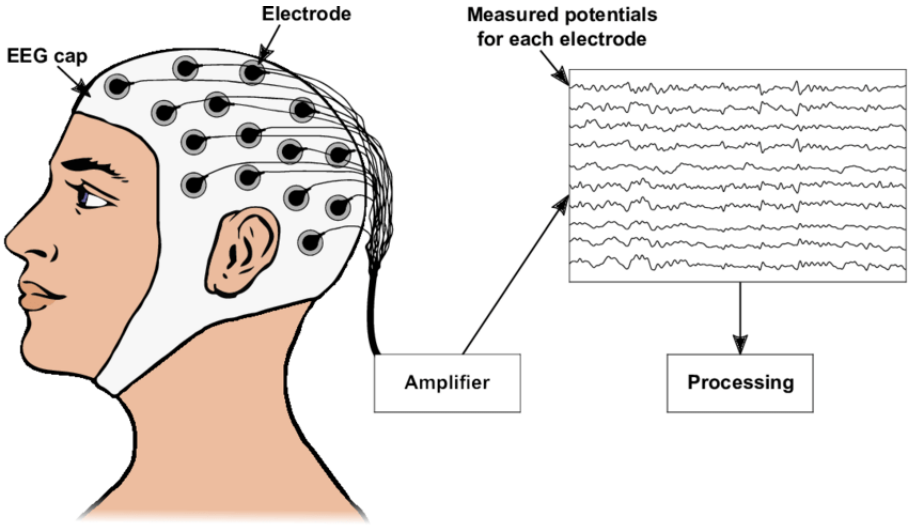

Figure 1.7: Schematic diagram of an EEG recording experiment. Source: [101]

The produced EEG signal is a measurement of currents that flow during synaptic excitations of the dendrites of many pyramidal neurons in the cerebral cortex. When neurons are activated, as we saw, the synaptic currents are produced within the dendrites, generating a magnetic field, measurable by EMG, and a secondary electrical field over the scalp, measurable by EEG systems. Differences of electrical potentials are caused by summed postsynaptic potentials from electrical dipoles between the neuron's body and apical dendrites. However, the human head consists of different layers (scalp, skull etc.) that severely attenuate the electrical signals. Other than that, most of the noise is generated either within the brain (internal noise) or over the scalp (system/external noise). Therefore, only large populations of active neurons can generate enough potential to be recordable, whereas these signals must be greatly amplified in order to be displayed.

Brain Rhythms

In healthy adults, the amplitudes and frequencies of EEG signals change from one state to another, such as wakefulness and sleep, or age. There are five major brain rhythms, distinguished by their different frequency ranges, called delta (δ), theta (ϑ), alpha (α), beta (β) and gamma (γ). The delta rhythm was introduced by Walter (1936) [159] at 0.5-4 Hz and has been primarily associated with deep sleep. Walter (1944) [160] also introduced theta waves as those having frequencies within the range of 4–8 Hz. Theta waves have been associated with access to unconscious material, inspiration and deep meditation, while they have proved crucial in early childhood as well as in arousal detection [131].

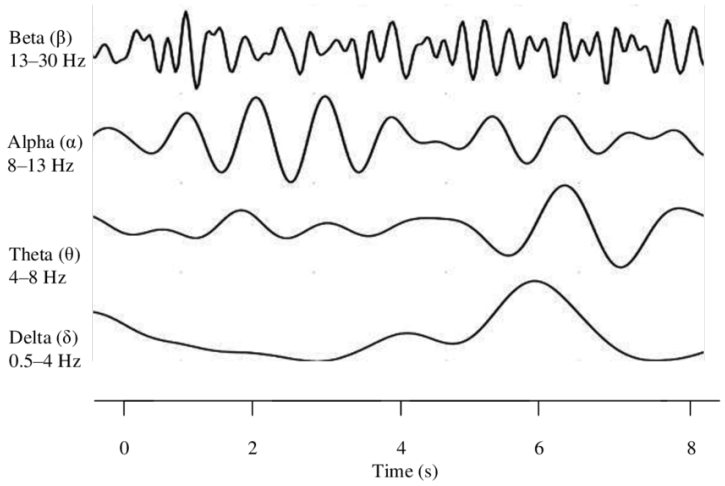

Figure 1.8: Four typical dominant brain normal rhythms. Source: [131]

The alpha and beta waves were introduced by Berger (1929). Alpha waves appear in the posterior half of the head and are usually found over the occipital region of the brain. Their frequency lies within the range of 8–13 Hz, and they commonly appear as round or sinusoidal signals, to indicate a relaxed awareness. The alpha wave is the most prominent rhythm in the whole realm of brain activity, it is however weakened by opening the eyes, by hearing unfamiliar sounds, by anxiety, or at periods of mental concentration. A beta wave is the electrical activity of the brain varying within the range of 13–30 Hz. It is the usual waking rhythm of the brain, associated with active thinking, attention, focus on the outside, or solving concrete problems, found in normal adults. A high-level beta wave may be acquired when a human is in a panic state. Last, Jasper and Andrews (1938) used the term "gamma" to refer to the waves of above 30 Hz. Although the amplitudes of these waves are very low and their occurrence is rare, detection of these rhythms can be used for confirmation of certain brain diseases. The gamma band has also been proven indicator of event-related synchronization (ERS) of the brain.

1.3 Music Perception

When Bob Dylan dared to play an electric guitar at the Newport Folk Festival in 1965, people walked out and many of those who stayed, booed. The Catholic Church banned music that contained more than one musical part playing at a time, fearing that it would cause people to doubt the unity of God. The church also banned the musical interval of an augmented fourth, the distance between C and F-sharp, also known as a tritone. This interval was considered so dissonant that the church named it *Diabolus in musica* [73]. Perceiving music is thus a complex process that involves receiving sound signals, analyzing their structure and eliciting variable psychological or cultural responses.

1.3.1 The Auditory System

The ability to recognize sounds and identify their organization is possible thanks to the auditory system, consisting of two main parts: the ear and the brain. The ear's task is to convert sound energy into neural signals, while our brain has to receive and process the information those signals contain. The human ear has striking abilities of detecting and differentiating sounds. It is sensitive to a wide range of frequencies and intensities,

having also an extremely high temporal resolution [64]. The ear consists of the outer, the middle, and the inner ear. The outer ear acts as a receiver and filters sound waves on their way to the ear drum (tympanic membrane) of the middle ear, amplifying some sounds and attenuating others, depending on their frequency and direction. Sound waves cause the tympanic membrane to vibrate, and these vibrations are subsequently amplified and transmitted on to the oval window of the cochlea, a small membrane-covered opening in the inner ear. The cochlea is fluid-filled and contains thousands of hair cells that react to different tones and pitches. The inner ear then translates vibrations into electrical signals, which are carried into the brain's cerebral cortex via the cochlear nerve system. Many areas of the brain are then joining to analyze different musical elements.

1.3.2 Principles of Music Structure

It was pitch and intervals that had the medieval church in an uproar and it was timbre that got Dylan booed. So before we examine how music affects our brains and emotions, it is helpful to examine what music is made of, what are its fundamental building blocks and how do they give rise to music. As the composer Edgard Varèse famously defined it, "Music is organized sound". The basic elements of any sound are loudness, pitch, contour, duration (or rhythm), tempo, timbre, spatial location, and reverberation. Our brains organize these fundamental perceptual attributes into higher-level musical concepts: meter, harmony, and melody. When we listen to music, we are actually perceiving these multiple attributes and dimensions, which we briefly summarize below [73]:

- A discrete musical sound is called a *tone*. The word *note* is also used, but we usually reserve that word to refer to something notated on a page or music score. The two terms describe though the same abstract concept.
- *Pitch* is a purely psychological construct, related both to the actual frequency of a particular tone and to its relative position in the musical scale.
- *Rhythm* refers to the duration of a series of notes and how they relate. In musical scores we denote that with different subdivisions of a note, as seen in Figure 1.9.

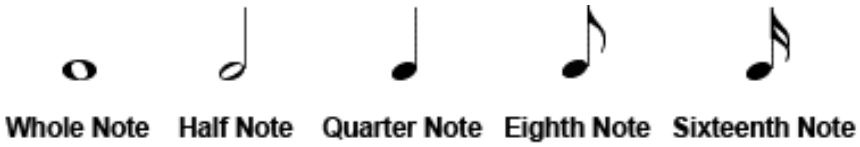

Figure 1.9: Subdivisions of notes' duration. Source: learnpianoforfree.weebly.com

- *Tempo* refers to the overall speed or pace of the piece.
- *Contour* describes the overall shape of a melody, taking into account only whether a note goes up or down, not the amount by which it changes.
- *Timbre* is the tonal color that which distinguishes one instrument from another (e.g., piano from a guitar) when both are playing the same note.
- *Loudness* is a psychological term that relates to the physical amplitude of a tone.

The difference between music and a random set of sounds has to do with the way the above fundamental attributes combine and the relations that are formed between them. When these basic elements combine and form relationships with one another in a meaningful way, they give rise to higher-order concepts like [73]:

- *Meter*: It is created in our brain by extracting information from rhythm and loudness cues, and refers to the way in which tones are grouped across time.
- *Key*: It has to do with a hierarchy of importance that exists between tones in a musical piece; this hierarchy exists only in our minds, as a function of our experiences with a musical style and musical idioms that we develop for understanding music.
- *Melody*: The main theme of a musical piece, the part you sing along with. The notion of melody is different across genres.
- *Harmony*: It has to do with relationships between the pitches of different tones, and with tonal contexts that these pitches set up. Harmony can mean simply a parallel melody to the primary one or it can refer to a chord progression.

The idea of hierarchical building of musical sounds is important since it helps us relate different parts of the human brain to music processing, for example the auditory cortex for the analysis of tones, the cerebellum for movements and rhythm perception and amygdala for emotion induction. Further, this formulation provides us the tools to identify and quantify those features in music data, in both audio and written form, so that we could use them for reasoning and classification purposes [107].

1.3.3 Haunted by Music

As we mentioned above, different aspects of music are handled by multiple neural regions. The most mysterious and deeply researched aspect is the emotional impact of music on humans. Music is said to evoke strong emotions, usually more powerful than, for example, static images, and can be used to investigate a wide variety of emotions, as well as mixed emotions, such as "pleasant sadness" [64]. Due to its temporal structure, music can be used to study the time course of emotional processes, while it can also be viewed under the lens of its social influences and consequences of the induced mood states [73]. Studying brain's responses to music has thus gained a lot of attention, especially nowadays that there is an upsurge in available neuronal data. Researchers are investigating our emotions in order to approach the ultimate question of why do we like music.

One of the most important aspects of music perception is the ability to anticipate future events, in the form of patterns [96]. This is fundamental for our survival and matches to the structured form of music patterns (intervals, chord progressions, tempo etc.). Another important factor is the biological reward system that has been tested to respond not only to basic stimuli, like appetite, but also in various other occasions, such as music [174]. However, it is not clear yet why the reward system is engaged in such stimuli. From the neuroscientific perspective, one of the core findings is the correlation between the frequency and magnitude of neural oscillation patterns and rhythmical patterns in music [103]. Additionally, Event-Related Potentials (ERP) have been utilized to extract brain activity patterns that can relate to the structure of musical events, such as note onsets or pitch [134, 116]. In addition to the well-controlled auditory experiments, modern approaches gather physiological data from listeners as they enjoy or imagine naturalistic music [85, 147], in order for instance to examine correlations in temporal structure [157].

1.4 Thesis Structure & Contributions

This study attempts to offer some further insights into the affective responses of music listening, from a computational perspective. We begin by analyzing the structure of brain electrical signals and providing novel feature extraction algorithms, to later proceed into data-intensive deep learning approaches to correlate brain responses to latent musical features. The remainder of this thesis is organized in 5 chapters as described below. Note that each chapter can be considered self-contained in terms of notation and methodology, however conclusions may be drawn from the results of former chapters.

- **Chapter 2** outlines the necessary signal processing and machine learning background to follow the methods and content of the present study. Specifically, we dive into fundamental properties of signals and systems, probability theory and fractal algorithms, as well as provide a detailed overview of supervised learning classifiers and state-of-the-art deep learning techniques.
- In **Chapter 3** we develop algorithms to analyze the fractal and multifractal properties of EEG signals, as well as investigate to what extent these properties carry emotional information. In the end, we indicate that multifractal analysis could serve for the development of robust models for the purpose of Emotion Recognition.
- In **Chapter 4** we focus on modeling the relationship between pairs of music tracks and corresponding EEG recordings. We propose a framework that can be utilized for emotion recognition both directly, by performing supervised predictions, and indirectly, by providing relevant music samples from EEG input queries.
- **Chapter 5** draws general conclusions on the research sections of our study and discusses possible future work based on our experiments. The measured multifractality of EEG signals and the cross-modal EEG and music framework could be the incentive for further research in music cognition and understanding.
- There are also 2 **Appendices** in which we provide a detailed description of the Dataset we use and the articles that we have published in the context of the thesis.

To give a brief summary of our contributions, these can be divided into two main sections, with respect to our conducted learning experiments. The major theme of our analysis is the affective perception of music signals through EEG responses. The first part focuses on feature extraction algorithms based on multifractal signal analysis, and the second addresses the problem of identifying emotion-related similarities between EEG and music signals through advanced deep learning techniques. More specifically:

Multifractal Analysis on EEG

- We analyzed the structure of EEG signals and demonstrated their multifractal properties. In specific, we investigated the effect of signal's observed stationarity and quantified the signal's complexity through the Hurst Exponent. We derived evidence that EEG signals could be modeled as fractional Gaussian noise realizations.
- We developed two novel algorithms, based on Multiscale Fractal Dimension (MFD) and Multifractal Detrended Fluctuation Analysis (MFDFA) to derive meaningful feature vectors for emotion detection.

- We tested the proposed methods through an SVM classifier, against widely used baseline frequency and fractal features. We showed that the proposed feature sets perform strongly, particularly in the subject-independent setting and in arousal recognition, indicating that arousal is correlated with the structure of the EEG.

- Fractal and multifractal features seem to generalize more easily than frequency-related ones, which perform better in subject-dependent settings. Further improvements are achieved when the fractal features are aggregated. As a result, the observed multifractality should be considered when processing EEG signals.

EEG-Music Cross-Modal Learning

- We presented a robust 3D deep network to efficiently analyze EEG signals or EEG features by preserving their temporal and spatial correlation. We additionally provided ways of dealing with core problems associated with this kind of data, such as the limited sample size and their noisy structure.

- We proposed a multimodal framework to model the correspondence between human brain responses and music stimuli. We trained a bi-stream network on pairs of EEG and corresponding music stimuli, whereas by conditioning the learning process with emotion tags we constructed a common emotion space.

- Through the produced latent space by the aforementioned network, we performed emotion recognition both by predicting output annotations and by ranking music tracks to EEG input queries, based on their cosine distance on the space.

- We performed a qualitative study across 32 subjects by formulating personalized models. This way we could compare 32 model instances and observed significant patterns, such as the visualized latent spaces, the temporal variation of recognition performance and activation patterns on the simulated scalp grid of the EEG network.

Both sets of experiments reveal important affective characteristics of brain signals and illustrate multiple ways that music influences the functioning of the human mind. While it provides empirical answers regarding the nature of music and emotion encoding in the human brain, this study also provides insights for further research in this fascinating field.

Chapter 2

Theoretical Background

2.1 Signal Processing Fundamentals

Signal Processing, and specifically Digital Signal Processing (DSP) is a set of (digital) operations that we apply on signals in order to achieve a particular goal, i.e. extract information. The foundations of DSP lay on Mathematics, Physics, and Computer Science, and the field has been expanding significantly over the last few decades as a result of rapid developments in computer architectures and artificial intelligence algorithms. Research and development in DSP are driving advancements in many areas including telecommunications, multimedia, medicine and human-computer interaction. The 2 main characters in Signal Processing are **Signals** and **Systems** that operate on them.

2.1.1 What is a Signal?

A *signal* is defined as any physical quantity that carries some kind of information. Mathematically, it is merely a function of one or more independent variables, such as time (1D signal) or space (2D or 3D signal). For example, in an electrical system the physical variables of interest might be a voltage or current, whereas in a mechanical system the variables of interest might be the velocity, mass or volume of an object. However, there are many cases where signals cannot be modeled by an explicit mathematical relation, but are better described via statistical models as random signals, i.e. noise.

In the real world, most signals are continuous-time or analog. That is, the independent variable of the function is allowed to take on arbitrary values (perhaps within some interval) and the value of the signal itself is also allowed to take on arbitrary values (again within some interval). While convenient in certain cases, in most situations it is preferable to work with digital signals that can be processed by a computer. The digitization of the signal domain, so that its values at a discrete set of time instants can be stored, is called sampling. Further, the process of digitizing its range is called quantization and the digital signal is only allowed to take a discrete set of values. In the following, continuous-time signals will be denoted using parentheses, such as $x(t)$, while discrete-time signals and generally digital signals will be denoted using brackets, such as $x[n]$.

2.1.2 Time - Frequency Representations

A *system* is defined as a process whose input and output are signals. Common operations that are applied via simple systems include adding and multiplying signals, differentiation and integration, shifts in time and amplitude, compression and reflection.

Fourier Analysis

A signal can be viewed as a vector in the vector space of its independent variables. In particular, each signal can be represented (or expanded) as a linear combination of elementary signals in the vector space. The most fundamental signal expansion is provided by the Fourier Transform, stating that absolutely integrable signals can be redefined in terms of sinusoidal frequencies or complex exponentials:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\omega)e^{j\omega t}d\omega, \quad \text{where} \quad X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t}dt \tag{2.1}$$

This frequency signal representation has been proven extremely important in Signal Processing and Electrical Engineering in general, since many signals are intuitively better represented in the frequency domain, such as music signals (i.e. notes) or neuronal signals (i.e. brain rhythms). Among the variants of Fourier Transforms, we will concentrate on the Discrete Fourier Transform (DFT) that is typically applied to digital signals of finite duration, using the efficient Fast Fourier Transform (FFT) algorithm.

Let $x(t)$ a continuous signal and a sampling of N points, denoted as $x[k]$. We could in principle evaluate this for any frequency ω, but we have only N data points and N significant output points. Additionally, continuous Fourier Transform over a finite-duration signal would be periodic. Since here a finite number of points would be considered periodic, we evaluate the DFT equation for the fundamental frequency $1/NT$ (one cycle per sequence) and its harmonics. Hence we derive the DFT formula $X[n]$ of $x[k]$:

$$X[n] = \sum_{k=0}^{N-1} x[k]e^{-j\frac{2\pi}{N}nk} = \sum_{k=0}^{N-1} x[k]W_N^{nk} \quad (n = 0, 1, \ldots, N-1) \tag{2.2}$$

where $W = \exp(-j2\pi/N)$ and $W = W^{2N}\text{etc} = 1$. The time taken to evaluate any mathematical process on a computer depends principally on the number of multiplications involved, since these are the slowest operations. Since DFT is calculated as a complex matrix-vector multiplication, this number is directly related to N^2; hence computational speed becomes a major consideration. Highly efficient algorithms for estimating DFT have been developed since the mid-60s [30], known as *Fast Fourier Transform* (FFT) algorithms, relying on the fact that DFT involves a lot of redundant calculations. From Eq. 2.2 it is easy to realise that W_N^{nk} is a periodic function with only N distinct values and repeats itself for combinations of n and k. Hence, let us split the single summation over N samples into 2 summations, one for even values of k, and the other for odd:

$$X[n] = \sum_{k=0}^{\frac{N}{2}-1} x[2k]W_N^{2kn} + \sum_{k=0}^{\frac{N}{2}-1} x[2k+1]W_N^{(2k+1)n} \tag{2.3}$$

Note that

$$W_N^{2kn} = e^{-j\frac{2\pi}{N}(2kn)} = e^{-j\frac{2\pi}{N/2}kn} = W_{\frac{N}{2}}^{kn}, \tag{2.4}$$

$$W_N^{(2k+1)n} = e^{-j\frac{2\pi}{N}(2kn)}e^{-j\frac{2\pi}{N}(n)} = W_N^{2kn}W_N^n \tag{2.5}$$

Therefore

$$X[n] = \sum_{k=0}^{\frac{N}{2}-1} x[2k]W_{\frac{N}{2}}^{kn} + W_N^n \sum_{k=0}^{\frac{N}{2}-1} x[2k+1]W_{\frac{N}{2}}^{kn} \tag{2.6}$$

i.e. $X[n] = G[n] + W_N^n H[n]$. Thus, the DFT $X[n]$ can be obtained from two $\frac{N}{2}$-point transforms. Although the frequency index n ranges over N values, only half values of $G[n]$ and $H[n]$ need to be computed since they have period $N/2$. Assuming that N is a power of 2 , we can repeat the above procedure on the two $\frac{N}{2}$-point transforms, breaking them down to $\frac{N}{4}$-point transforms, etc., until we come down to 2-point transforms, that require only 1 complex multiplication and 2 complex additions. Thus, FFT is computed by decimating the sample sequence $x[k]$ into sub-sequences until only 2-point DFTs remain. At each stage of the FFT, $\frac{N}{2}$ complex multiplications are required to combine the results of the previous. Since there are $\log_2 N$ stages, the number of complex multiplications required is approximately $N/2 \log_2 N$, a tremendous improvement for large-scale data.

Locating the Frequencies

As we saw, the Fourier Transforms give us the frequency spectrum of a signal. However, the spectrum contains no additional information about the temporal localization of the various frequencies, therefore we lose the time resolution of the real signal. A straightforward way to deal with this problem is to divide the original signal into several, possibly overlapping, parts and separately apply the Fourier Transform to each of them, a technique called *Short-Time Fourier Transform*. By then concatenating the resulting spectra as a function of time, we end up with a 2D time-frequency representation, called a *spectrogram*. Let us consider a continuous-time function $x(t)$. This function is to be multiplied by a short time window $w(t)$. The Fourier transform of the resulting signal is calculated as the window slides along the time axis. Mathematically:

$$\text{STFT}\{x(t)\}(\tau, \omega) \equiv X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w(t - \tau) e^{-i\omega t} dt \tag{2.7}$$

In the discrete-time case (signal $x[n]$ and window $w[n]$), the data could be broken up into overlapping frames to reduce boundary artifacts. Each chunk is Fourier-transformed like before. In practice, the computation is often done through the FFT algorithm:

$$\text{STFT}\{x[n]\}(m, \omega) \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n] w[n - m] e^{-j\omega\pi} \tag{2.8}$$

One of the pitfalls of the STFT is that it has a fixed resolution, which is not intuitive to how each frequency is represented. In specific, low frequencies require larger time windows of analysis whereas higher frequencies require smaller time windows. From another perspective, a big window would give better frequency resolution but poor time resolution. A narrower window will of course cause the opposite effects.

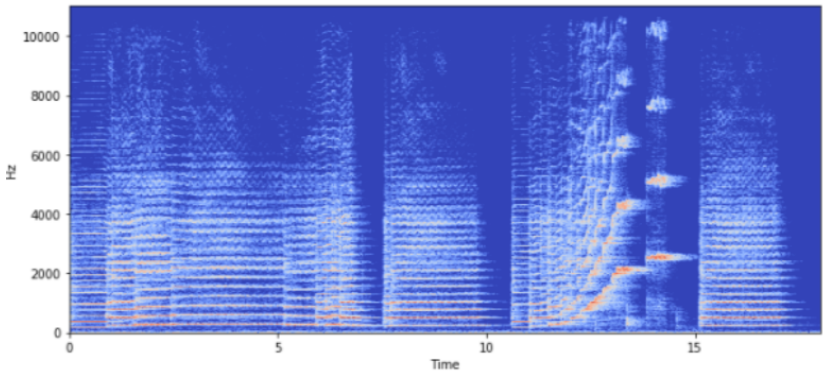

Figure 2.1: An example STFT spectrogram, extracted from a music signal (violin).

An effective solution to the above-mentioned malfunction is achieved using wavelets. A *wavelet* is a wave-like oscillation that is localized in time (Figure 2.2). Wavelets are governed by 2 basic properties: scale (a) defines how stretched a wavelet is and relates to the signal's frequency content, and translation (b) defines where the wavelet is positioned, thus relates to the signal's temporal content. The idea here is to compute how much of a wavelet is in a signal, using convolutions. So, we pick a specific wavelet at a particular scale and slide it across the entire signal. The product of this multiplication gives us a wavelet coefficient at each timestep. We then alter the scale and repeat the process, resulting as well in a 2D time-frequency representation, called *scalogram*.

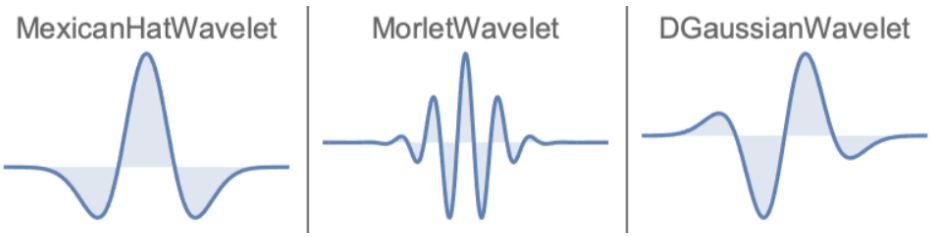

Figure 2.2: Example types of wavelets. Source: reference.wolfram.com/language

Wavelet analysis is applied in two main ways, *Continuous Wavelet Transform* (CWT) and *Discrete Wavelet Transform* (DWT). CWT uses a wavelet function $\psi(t)$ and assumes every possible wavelet in scale and translation, hence we usually focus on the practical case of Discretized CWT that uses a particular set of coefficients and is written as:

$$X_{a,b} = \int_{-\infty}^{\infty} x(t)\psi_{a,b}(t)dt \tag{2.9}$$

On the other hand, DWT decomposes a signal into two components: a lowpass signal, using a scaling function, and a highpass signal using a wavelet function. DWT recursively decomposes the lowpass signal with the same scaling and wavelet functions to the desired level of decomposition. A couple of key advantages of wavelet analysis is that, contrasting to the STFT approach, it can extract local temporal and spatial information at the same time, providing scalograms of better resolution. Moreover, it is an easily customizable approach, since there is a large variety of possible wavelets to try out for each task.

2.1.3 Probability Theory and Statistics

While the methods outlined above are indeed efficient in analyzing the properties of deterministic signals, there is a broad range of other signals that cannot be fully determined by mathematical functions. These random signals are produced by complex physical processes like the movement of air particles or electrons within an EEG acquisition device. Such signals are commonly interpreted as noise or interference to other meaningful signals and are analyzed through statistical metrics and probabilistic models. Probability theory is based on the notion of the random experiment, meaning an experiment whose outcome is random. The set of all possible outcomes of a random experiment S is called the sample space Ω and we call event a set $A \subseteq S$. In that context, we define the probability $P(A)$ as a function of the uncertainty of each possible event A.

An important concept that we will further need is *conditional probability*. Let a random experiment S and two events A, B. We denote $P(B|A)$ the conditional probability of event

B, given that A occurs. The conditional probability is determined by the rule:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \;\rightarrow\; P(A \cap B) = P(B|A)P(A) = P(A|B)P(B). \tag{2.10}$$

We can thus denote the probability of an intersection of events as the product of the conditional probability of the one given the second and the probability of the second. By rearranging the above outcome we end up with the *Bayes' Rule*:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.11}$$

Of course, if A and B are statistically independent, then B does not depend on A:

$$P(B|A) = P(B) \;\rightarrow\; P(A \cap B) = P(A)P(B) \tag{2.12}$$

Random Variables

Sometimes it is preferable to assign a numerical value to each of the events of a random experiment. This kind of function $X : \Omega \rightarrow R$ is called a *random variable*. Now, given a random variable X, we define the probability of the event $X \leq x$ as $F_X(x) = P(X \leq x)$, where $F_X(x)$ is the cumulative distribution function (CDF) of X. The CDF is a non-decreasing monotonous function, bounded between 0 and 1. Its derivative $f_X(x) = F'_X(x)$ is called the probability density function (PDF) of the random variable X. Obviously:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy, \qquad \int_{-\infty}^{\infty} f_X(x)dx = 1 \tag{2.13}$$

Now that we can adequately describe random experiments, we need some metrics to evaluate their behavior. The *expected value* (mean) of a random variable X is given by:

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx \tag{2.14}$$

From this point, we can easily derive the mean of any function of X as follows:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx \tag{2.15}$$

Especially when $g(X) = X^n$, $\mathbb{E}[g(X)]$ defines the n^{th} moment of X. Of particular interest are the so-called central moments, which are computed on the difference between a random variable X and its mean μ_X. The 1st central moment is obviously always zero. The 2nd central moment is known as the variance of X,

$$\text{var}[X] = \mathbb{E}[(x - \mu_X)^2] = \int_{-\infty}^{\infty} (X - \mu_X)^2 f_X(x)dx \tag{2.16}$$

whereas the square root of var$[X]$ represents the standard deviation σ_X of X. Intuitively, the variance and standard deviation measure the degree of randomness in X. These are the most fundamental metrics for a single random variable. However, in our analysis we will need to compare signals and their statistical properties, so we introduce respective metrics for pairs of random variables X, Y. Their joint moments are given by

$$\mathbb{E}[X^i Y^j] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i y^j f_{X,Y}(x, y)dxdy \tag{2.17}$$

The first joint moment $\mathbb{E}[XY]$ represents the *correlation* of X, Y whereas the first joint central moment $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ represents the *covariance* cov$[XY]$ of X, Y.

Random Processes

We have defined as random signals those that cannot be explicitly described or predicted. By extending this definition to probability theory, we could state that each occurence of a random signal is an event in the sample space. Those sample spaces that include signal occurences are called *random processes*. A signal is then a realization (event) of the random process and a point on the signal is itself a random variable. Random processes also inherit a probability distribution that assigns a specific probability to each realization. The metrics defined above generalize easily to random processes, as well.

A crucial attribute of a random process is its stationarity. Specifically, a random process is called *first order stationary* if its mean and variance are constant in time. We also define the *autocorrelation function* of a random process $X(t)$ as:

$$R_X(t_1, t_2) = \mathbb{E}[X(t_1), X(t_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X(t_1),X(t_2)}(x_1, x_2) dx_1 dx_2 \tag{2.18}$$

A random process is called *second order stationary* if its autocorrelation is dependent only on the time difference $\tau = t_2 - t_1$. A process that combines those 2 rules can be fully characterized as *wide-range stationary*.

Power Spectral Density

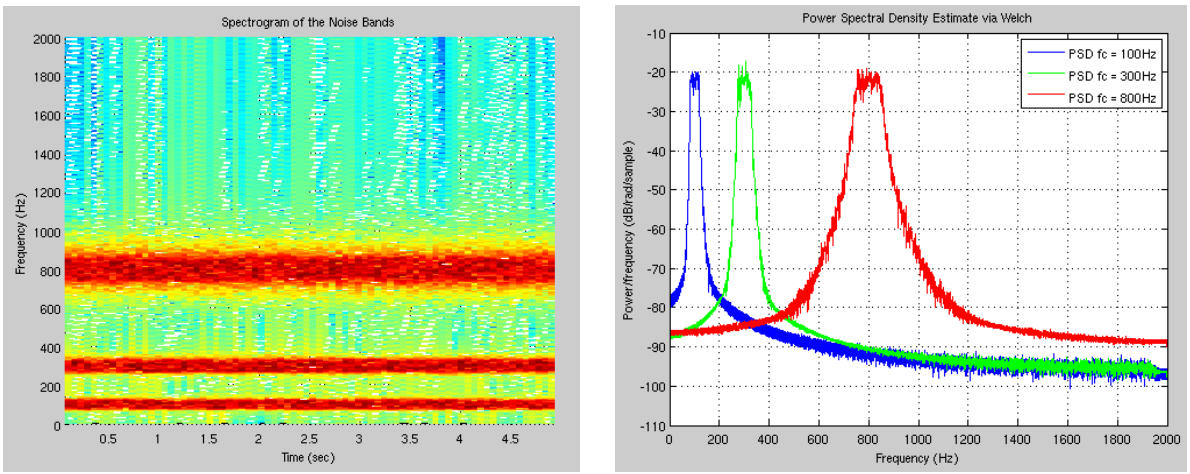

Figure 2.3: Visualization of 3 noisy signals with fundamental frequencies at 100, 300 and 800 Hz. Depicted are their STFT spectrogram (left) and Power Spectral Density (right). Source: https://ccrma.stanford.edu/ jcaceres/yamaha/

We have already seen that the frequency representation of deterministic signals is acquired through the Fourier Transform. How could we generalize this principle to random signals, which, as mere instances of a random process, might not be representative? The answer is to exploit its autocorrelation function (Eq. 2.18), whose Fourier Transform constitutes a powerful signal representation called *Power Spectral Density* (PSD). PSD describes how power is distributed over the frequency content of a random process and has found use in many tasks, in which the frequency content and variation is important, eg. in audio processing [78] and biosignals [105]. In specific:

$$S_X(f) = \mathcal{F}\{R_X(\tau)\} = \int_{-\infty}^{\infty} R_X(\tau) e^{-2j\pi f\tau} d\tau \tag{2.19}$$

If $X(t)$ is a real-valued random process, then $R_X(\tau)$ is an even, real-valued function of τ. From the properties of the Fourier Transform, we conclude that S_X is also real-valued and an even function of f. Also, S_X is non-negative for all f. To compute PSD in our experiments we will exploit the algorithm proposed by Peter D. Welch [169]:

The original N-point signal is split up into K data segments, each of length M, overlapping by D points. The overlapping segments are then windowed in the time domain. Most windowing functions somewhat suppress the edges of a segment, however the information is retained when overlapping the segments. For each segment, the periodogram is computed using the squared magnitude of its DFT. The individual periodograms are then averaged, reducing the variance of the individual power measurements. The end result is an array of power measurements for each frequency bin. Below we visualize an example of 3 noisy sinusoids at 100, 300 and 800 Hz.

2.1.4 Fractal Signal Analysis

Mathematicians Lewis Fry Richardson and Benoit B. Mandelbrot are credited with introducing the notion of fractal shapes and dimensions. The term *fractal* originates from the latin word *fractum*, meaning *broken*, and it was introduced by Mandelbrot [45] to describe "unusual" shapes that cannot be modeled geometrically. What they found by measuring the British coastline [88] is that, by increasing the precision of the measurements, the measured total length appears to increase as well. This reflects the self-similar structure of the coastline across a wide range of length scales. Since traditional geometrical metrics were inadequate in modeling such a behavior, fractal algorithms and metrics were introduced to address the property of self-similarity. Some popular self-similar shapes is the Mandelbrot set, the Koch's snowflake and the Herpinski Triangle, depicted in Figure 2.4. However, other than those ideal fractals, a lot of physical processes and signals demonstrate similar properties. Natural schematic patterns [150], music signals [17, 180] as well as biomedical signals [39] show indeed a complex structure across timescales.

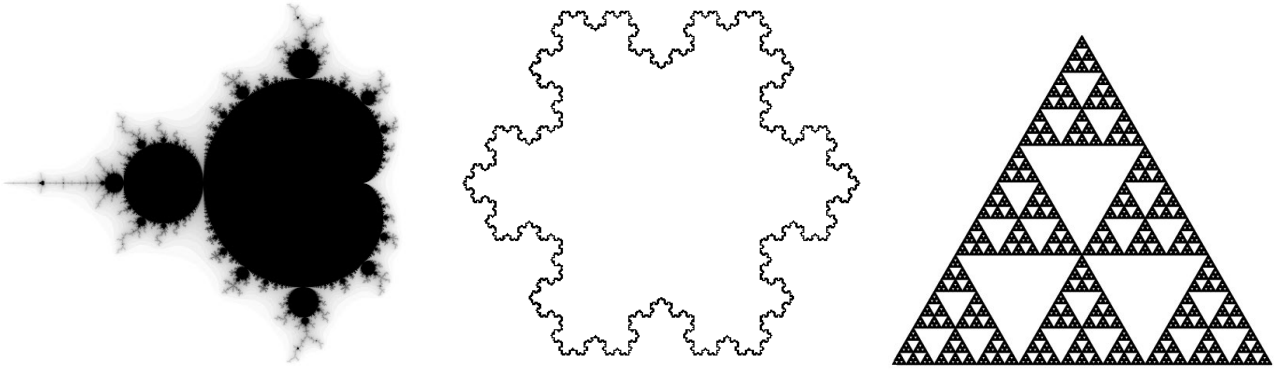

Figure 2.4: Left to right: the Mandelbrot Set, the Koch's Snowflake and the Herpinski Triangle.

The complexity of such signals is typically measured through the fractal dimension, which is higher than their topological dimension. Intuitively, such complex shapes could resemble, and indeed share properties of, shapes of higher dimensionality. However, there is no consensus in determining the fractal dimension of a signal and, as a result, various algorithms have been proposed. Although for some classic fractals all these algorithms coincide, in general they are not equivalent. The most researched ones are the following:

Similarity Dimension

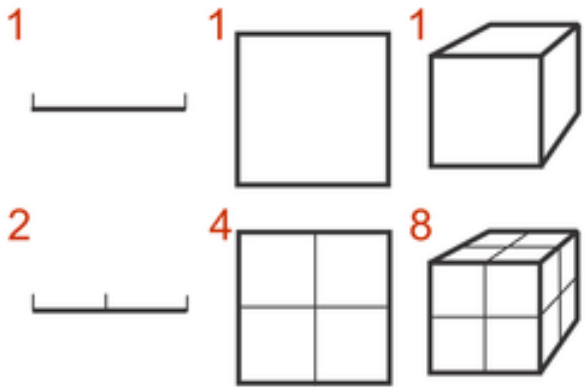

Figure 2.5: Visual Inspection of $N = r^D$ for $r = 1, 2$. Source: [137]

If we take an object residing in Euclidean dimension D and increase its linear size by r in each spatial direction, its measure (length, area, or volume) would increase to $N = r^D$ times the original. We depict this in Figure 2.5. By taking the logarithm of both sides we get $\log N = D \log r$. If we solve for D : $D = \log N / \log r$. Now D does not need to be an integer, in fractal geometry it is actually a fraction. This generalized treatment of dimension is named after the German mathematician, Felix Hausdorff, and has been proven useful for describing fractal objects and trajectories of dynamic systems [41].

Box Counting Dimension

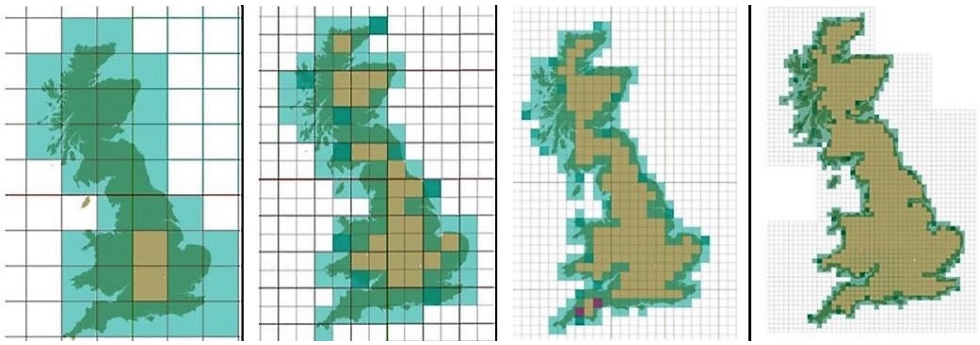

Figure 2.6: Estimating the box-counting dimension of the coast of Great Britain [YouTube].

Also known as the *Minkowski–Bouligand Dimension* [98], the box counting dimension is a way of determining the fractal dimension of a set S in a euclidean space $\mathbb{R}^n$. It is named after the German mathematician Hermann Minkowski and the French mathematician Georges Bouligand. To calculate this dimension for a fractal S, imagine this fractal lying on an evenly spaced grid, and count how many rectangular boxes are required to cover it. The box-counting dimension is calculated by seeing how this number changes as we make the grid finer. Suppose that $N(\epsilon)$ is the number of boxes of side length ϵ required to cover the set. Then the box counting dimension is defined as:

$$D_{\text{box}}(S) := \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)}. \tag{2.20}$$

We depict this method in Figure 2.6. The rectangular is not the sole type of kernel that can be used and, indeed, we will later exploit a more advanced method on this.

Higuchi Fractal Dimension

Higuchi Fractal Dimension (HFD) was proposed by Higuchi et al. [50] as an approximation for the box-counting dimension of the graph of a real-valued function or time series. It has many applications in science and engineering eg. in seismograms [43], clinical neurophysiology [59] and Alzheimer's disease treatment [142]. Given a time-series X, for each $k \in \{1, \ldots, k_{\max}\}$ and $m \in \{1, \ldots, k\}$ we define the length $L_m(k)$ by:

$$L_m(k) = \frac{N-1}{\left[\frac{N-m}{k}\right] k^2} \sum_{i=1}^{\left[\frac{N-m}{k}\right]} |X_N(m+ik) - X_N(m+(i-1)k)| \tag{2.21}$$

The length $L(k)$ is defined by the average value of the k lengths $L_1(k), \ldots, L_k(k)$,

$$L(k) = \frac{1}{k} \sum_{m=1}^{k} L_m(k) \tag{2.22}$$

The slope of the best-fitting linear function through the data points $\left\{\left(\log \frac{1}{k}, \log L(k)\right)\right\}$ is defined to be the Higuchi Fractal Dimension of the time-series X.

2.2 Machine Learning Fundamentals

Artificial Intelligence (AI) describes intelligence demonstrated by machines, in contrast to what is described as natural intelligence displayed by humans and animals. The field of AI is broadly described as the study of *intelligent agents*: any device that perceives its environment and takes actions to maximize its chance of successfully achieving its goals. AI is also widely considered as the attempt to mimic cognitive functions associated with the human mind, and has drawn inspiration from the human brain.

Machine Learning (ML) is a subfield of AI. It enables computers to learn from data and predict outputs without being explicitly programmed. In recent years, AI has experienced a resurgence due to advent of the Deep Learning subfield. Deep Learning utilizes large amounts of data and networks that can model a plethora of high-level concepts, achieving state-of-the-art and even superhuman performance on many tasks. ML algorithms are generally classified into *Supervised Learning*, in which algorithms are trained based on annotated data, and *Unsupervised Learning*, where we want to extract statistical structure from non-labeled data. A third ML variant is *Reinforcement Learning* (RL), which is concerned with how intelligent agents make decisions in order to maximize a type of reward or minimize respective penalties. RL is deeply connected to the notion of human learning and the human brain, however it will remain out of this study's scope.

2.2.1 Supervised Learning Algorithms

The majority of machine learning tasks exploit supervised learning methods. Supervised learning involves a set of input variables x and a set of annotations y for each input sample. The goal is to determine a mapping function $y = f(x)$ such that it generalizes well upon new, non-labeled input. We can define 2 broad categories of supervised learning problems: *Regression*, in which the goal is to estimate a real value ($y \in \mathbb{R}$) and *Classification*, in which the labels are organized into discrete classes.

Linear Regression

Maybe the simplest and oldest ML algorithm is the estimation of the best-fit line for a bunch of data points in a 2D space. In statistics, it is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. The case of one dependent variable is called simple linear regression while if more than one dependent variables exist, the process is called multiple linear regression. Simple linear regression estimates linearly how much the dependent variable y will change when the independent variable $\boldsymbol{x}$ changes by a certain amount (Figure 2.7):

$$\boldsymbol{y} = b_0 + b_1 \boldsymbol{x} \tag{2.23}$$

In Machine Learning we usually consider the matrix notation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}$, where rows of $\boldsymbol{X}$ correspond to data points (samples) and columns to data dimensions (features):

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ 1 & x_{31} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{bmatrix} \tag{2.24}$$

and b coefficients can be estimated by minimizing the sum of squared errors (SSE):

$$\hat{b} = \arg\min_{b} \|\boldsymbol{y} - \boldsymbol{b}\boldsymbol{X}\|_2^2 \tag{2.25}$$

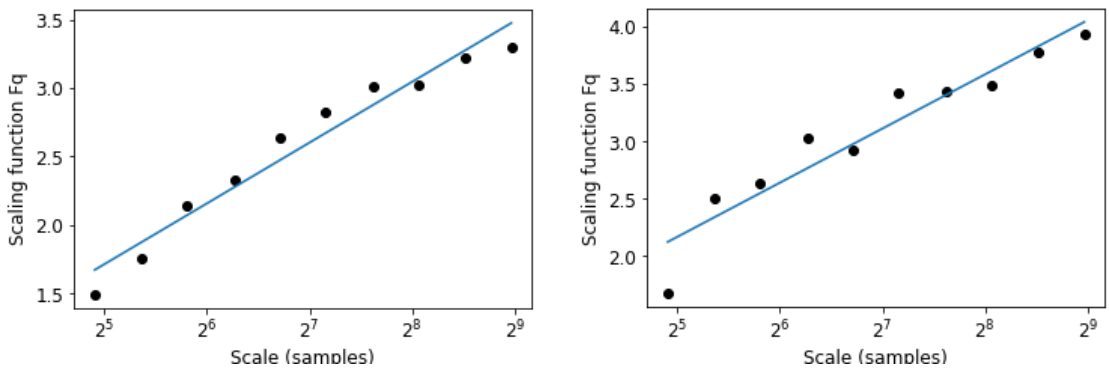

Figure 2.7: Visual examples of a linear regression implementation that will be used in Detrended Fluctuation Analysis (DFA) algorithm (Chapter 3), to determine measures regarding the EEG complexity. Here we depict the DFA result for two sample 30-sec. EEGs.

Trees and Forests

The *Decision Trees* algorithm also belongs to supervised learning, with the goal to create a model that can predict the target variables by learning simple decision rules inferred from training data. In Decision Trees, to predict a class label for a record, we start from the root of the tree, which represents the whole dataset. On the basis of comparison, we follow the branches corresponding to the value in hand and jump to the respective nodes until we reach a target value (leaf). Each node in the tree acts as a test query for some attribute, and each edge descending from the node corresponds to the possible answers to it. This process is repeated for every subtree rooted at the new node.

Now, a *Random Forest*, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's final prediction. The reason that this framework works so well is that, generally, a large number of relatively uncorrelated models (here trees) operating as a committee will outperform any of the individual constituent models. The low correlation is the key: uncorrelated tree models can protect each other from their individual errors. While some trees may be wrong, many other trees will be right, so as a group they move in the correct direction.

Support Vector Machines

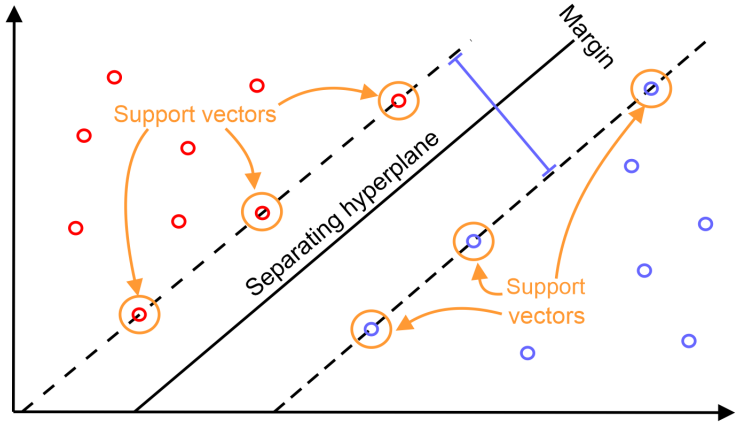

Figure 2.8: Illustration of a Linear SVM functionality. Source: [149]

A *Support Vector Machine* (SVM) [32] is a classification algorithm that is trying to find maximum-margin hyperplanes in order to create efficient classification boundaries for the data classes. In specific, let a training set of N input vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ with corresponding target values $y_1, ..., y_N$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Now, all hyperplanes in $\mathbb{R}^d$ are parameterized by a vector $\boldsymbol{w}$ and a constant b, expressed in the following linear equation:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \tag{2.26}$$

Given such a hyperplane $(\boldsymbol{w}, b)$ that separates the data, this gives the function

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b) \tag{2.27}$$

which correctly classifies the training data and could also generalize well. So we define the canonical hyperplane as the one that separates the data from the hyperplane by a margin of at least 1. That is, we consider those that satisfy:

$$y_i \left(\mathbf{x_i} \cdot \mathbf{w} + b\right) \geq 1 \quad \forall i \tag{2.28}$$

To obtain the geometric distance from the hyperplane to a data point, we must normalize by the magnitude of w. This distance is simply:

$$d\left((\mathbf{w}, b), \mathbf{x}_i\right) = \frac{y_i \left(\mathbf{x_i} \cdot \mathbf{w} + b\right)}{\|\mathbf{w}\|} \geq \frac{1}{\|\mathbf{w}\|} \tag{2.29}$$

Intuitively, we want the hyperplane that maximizes the geometric distance to the closest data points. This is accomplished by minimizing $\|\mathbf{w}\|$ (subject to the distance constraints),

using Lagrange multipliers. We can define the matrix $(H)_{ij} = y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ to provide more compact notation. The problem to solve is eventually transformed into:

$$\begin{array}{ll} \text{min:} & W(\alpha) = -\alpha^T \mathbf{1} + \frac{1}{2}\alpha^T H \alpha \\ \text{subject to:} & \alpha^T y = 0, \quad \mathbf{0} \leq \alpha \leq C \end{array} \tag{2.30}$$

where α is the vector of l non-negative Lagrange multipliers to be determined, and C is a regularization term for configuring the penalty of wrongly classified instances. In addition, the optimal hyperplane can be written as follows:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{2.31}$$

The solution of the constrained equation system (Eq. 2.30) is given by Lagrange multipliers. However, in many learning problems, feature vectors of different classes may be not linearly separable in their original embedding space. Presumably, one cannot easily find a hyperplane to serve as a classification boundary for data belonging to each class. For this reason, it was proposed that the original space be mapped into a higher-dimensional one, in which the separation would be easier. We could thus define the mapping $\mathbf{z} = \phi(\mathbf{x})$ that transforms the d-dimensional input vector $\mathbf{x}$ into a higher d'-dimensional vector $\mathbf{z}$. Given a mapping $\mathbf{z} = \phi(\mathbf{x})$, to set up our new optimization problem, we simply replace all occurences of $\mathbf{x}$ with $\phi(\mathbf{x})$. The problem now becomes:

$$\text{min:} \quad W(\alpha) = -\alpha^T \mathbf{1} + \frac{1}{2}\alpha^T H \alpha \tag{2.32}$$

with $(H)_{ij} = y_i y_j (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$ Then, the optimal hyperplane would be

$$\mathbf{w} = \sum_i \alpha_i y_i \phi(\mathbf{x}_i) \tag{2.33}$$

In all necessary computations, $\phi(\mathbf{x}_a)$ appears in a dot product with some other $\phi(\mathbf{x_b})$. That is, given the kernel for the dot product in the feature space:

$$K(\mathbf{x_a}, \mathbf{x_b}) = \phi(\mathbf{x_a}) \cdot \phi(\mathbf{x_b}) \tag{2.34}$$

the matrix would be $(H)_{ij} = y_i y_j (K(\mathbf{x}_i, \mathbf{x}_j))$, whereas the classifier would be

$$f(\mathbf{x}) = \text{sign}\left(\sum_i \alpha_i y_i (K(\mathbf{x}_i, \mathbf{x})) + b\right). \tag{2.35}$$

We can easily extend the previous formulation of binary decision SVMs in multi-class problems by simply training separate binary classifiers for all the classes available in the training data and choosing the one with the highest confidence.

2.2.2 Neural Networks and Optimization

An *Artificial Neural Network* (ANN) is a biologically inspired computational model that is modelled after the network of neurons in the human brain. The area of ANNs was initially developed to model biological neural systems, but has since diverged and become a matter of engineering and achieving state-of-the-art results in Machine Learning tasks. The basic computational unit of an ANN is the neuron, or the *Perceptron* [127].

From Neurons to Networks

Let us recall the definition of the biological neuron from Section 1.2. Each neuron receives input signals from its dendrites and produces aggregated output signals along its axis. If the final sum is above a certain threshold, the neuron can fire, sending a spike along its axis to the dendrites of other neurons, via synapses. For the Perceptron, we model the firing rate of the neuron with an activation function (eg. the Sigmoid function). Aside from that, its functionality resembles the biological neuron at a great extent.

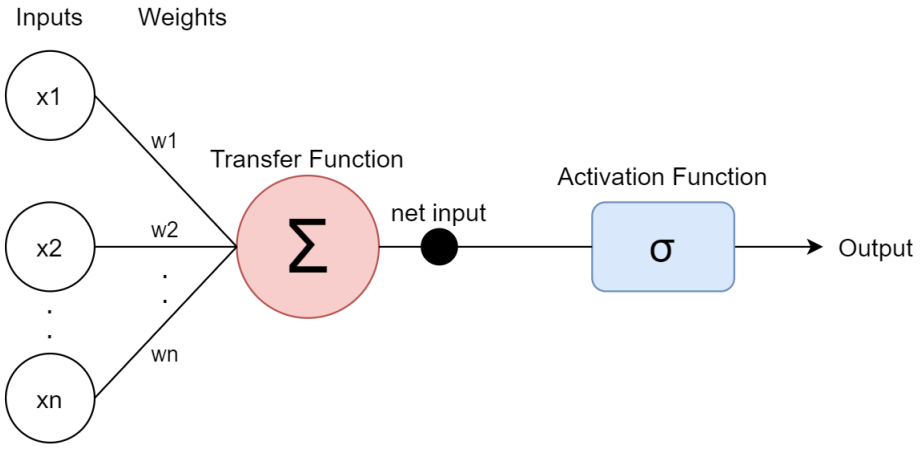

Figure 2.9: The Rosenblatt's Perceptron [127].

The perceptron consists of 4 parts: inputs, weights and bias, net sum and activation function. As shown in Figure 2.9, the input (data) vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is multiplied with the weight vector of the perceptron $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ and we refer to the aggregated value $\boldsymbol{x}^T\boldsymbol{w}$ as the weighted sum. The output of the perceptron is the value of the activation function, evaluated at this sum. Weights are used because they determine the strength of the respective node of the perceptron, whereas the bias value allows us to control the influence of the activation function. The purpose of this function is to apply a non-linear transformation to data points, in order to discriminate them into distinct categories.

The perceptron is comprehended as a binary linear classifier and separates data using a straight line. The *Perceptron Algorithm* was proposed to determine that line. The algorithm has been proven to converge and can adequately implement *linearly separable* functions. This is however a relatively narrow set of modeling functions.

Algorithm 1: The Perceptron Algorithm

Result: Weight parameter vector $\boldsymbol{w}$ of the separating line.
Initialize $\boldsymbol{w}$ randomly;
while *not converged* **do**
  Pop a sample $(\boldsymbol{x}, y)$ from the input set;
  if $y = 1$ *and* $\boldsymbol{x}^T\boldsymbol{w} < 0$ **then**
    $w \leftarrow w + x$;
  end
  if $y = 0$ *and* $\boldsymbol{x}^T\boldsymbol{w} \geq 0$ **then**
    $w \leftarrow w - x$;
  end
end

In pursuance of learning complex non-linear functions, architectures that combine several artificial neurons have been designed and are called *Multi-Layer Perceptrons* (MLPs). Instead of MLPs, *Feed-Forward Neural Networks* (FFNNs) have been implemented, where each neuron connects with all neurons of the previous layer and there are no connections between the neurons of the same layer. The network is composed of an input layer, one or more hidden layers and an output layer as depicted in Figure 2.10. A crucial component of these neurons is the utilized activation functions. These incorporate non-linearities in data modeling and help the network adapt and relate linearly invisible dependencies. The most common activation functions are the Sigmoid function, the Softmax function (usually on outputs) and the Rectified Linear Unit (ReLU). The question is how could we train such a network on our data, since the Perceptron algorithm would be insufficient in fine-tuning so complex networks with so many parameters.

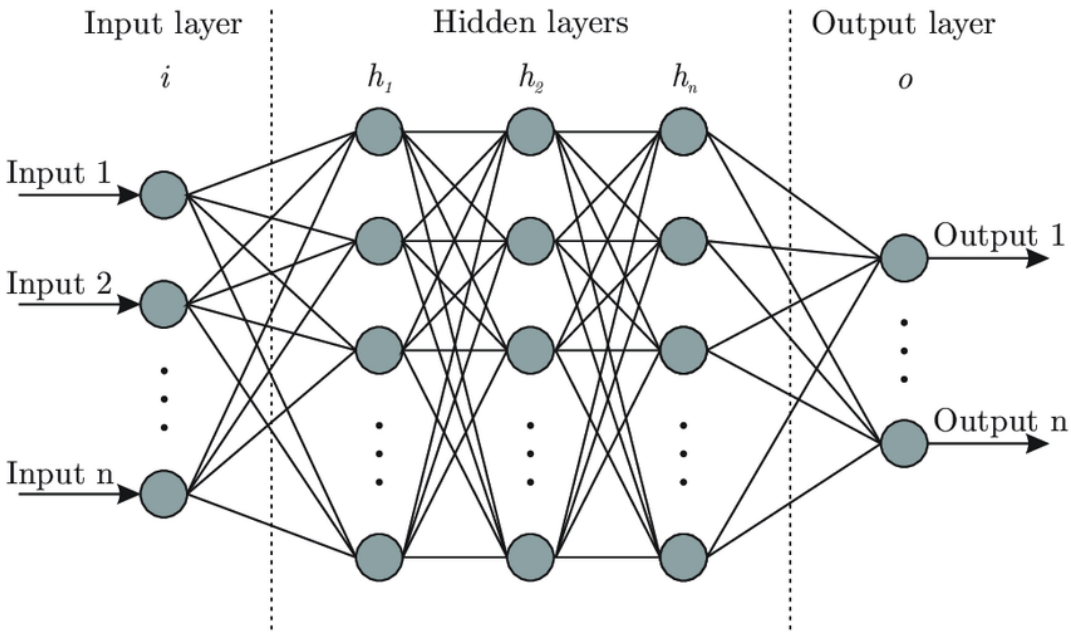

Figure 2.10: A 5-layer Neural Network containing n inputs, 3 hidden layers and an output layer with n outputs. It is an example of a Deep Neural Network (DNN). Source: [18].

Back-Propagation Algorithm

The objective of a Neural Network can vary (generative networks, autoencoders etc). However, in the baseline case of supervised learning, the network will try to minimize

$$\hat{\Theta} = \arg\min_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\left(f\left(\boldsymbol{x_i}; \theta\right), \boldsymbol{y_i}\right) \tag{2.36}$$

hence, match the output predictions to the respective labels of the input vectors. Training a neural network translates into minimizing such an objective, something that inevitably includes the computation of gradients for the fluctuation of the network parameters (weights). Iterative gradient-based optimization schemes can be summarized by 2 steps, namely computing the gradients and updating the weights. These steps are performed iteratively until convergence is guaranteed (gradients close to zero).

When training deep networks with many stacked layers, the gradient computation at each layer is not straightforward. The solution is to compute them layer-wise, starting

from the loss function and moving backwards, hence the Back-Propagation term. Let us consider the example of an intermediate fully-connected layer described by $y = \sigma\left(\mathbf{w}^T\mathbf{x}\right)$, where $\mathbf{x}$ is the input, y the 1D output, $\mathbf{w}$ the layer's weights and σ a non-linear activation. Given the output gradient $\vartheta J/\vartheta y$, we want to compute the gradients $\vartheta J/\vartheta \mathbf{w}$ and $\vartheta J/\vartheta \mathbf{x}$ in order to update the weights and propagate the gradient error, respectively. The first one, using the simple chain rule, can be expressed as:

$$\frac{\vartheta J}{\vartheta \mathbf{w}} = \frac{\vartheta J}{\vartheta y}\frac{\vartheta \sigma\left(\mathbf{w}^T\mathbf{x}\right)}{\vartheta\left(\mathbf{w}^T\mathbf{x}\right)}\frac{\vartheta\left(\mathbf{w}^T\mathbf{x}\right)}{\vartheta \mathbf{w}} = \frac{\vartheta J}{\vartheta y}\sigma'\left(\mathbf{w}^T\mathbf{x}\right)\mathbf{x} \tag{2.37}$$

while the second one as follows:

$$\frac{\vartheta J}{\vartheta \mathbf{x}} = \frac{\vartheta J}{\vartheta y}\frac{\vartheta \sigma\left(\mathbf{w}^T\mathbf{x}\right)}{\vartheta\left(\mathbf{w}^T\mathbf{x}\right)}\frac{\vartheta\left(\mathbf{w}^T\mathbf{x}\right)}{\vartheta \mathbf{x}} = \frac{\vartheta J}{\vartheta y}\sigma'\left(\mathbf{w}^T\mathbf{x}\right)\mathbf{w} \tag{2.38}$$

The above utilization of gradient chain rule can be used sequentially to the input in order to calculate every parameter gradient with respect to the objective function 2.36. After computing the gradients and the aggregated score of the objective function with respect to the parameters θ of the entire training dataset, we commonly update them using the following rule: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta J(\theta)$, where η is the learning rate.

2.2.3 Training and Evaluation Issues

Every type of neural network realizes a nonlinear function $\hat{y} = f_\theta(x)$, where θ is the set of all the weights comprising the network. This function can be trained by iteratively processing the available data, where a single iteration is called an *epoch*. One should first define a training set of input samples and a rather small validation set to monitor training. This data should not influence the training process by any means, however it is useful for adjusting critical network hyper-parameters. Of course, we also need an independent test set that will be used only to test the performance of the final trained model. Sometimes though, and this is the case in our study as well, the quantity of data is insufficient to properly define the above 3 subsets. In this case, the common practice is to split the available data in k subsets (folds) and perform k training sessions, in each of which a single set would be used as the validation set. In the end of this procedure, called *k-fold Cross-Validation*, we consider their average scores as indicative of the model's capabilities. Below we visually depict the concept of the Cross-Validation method.

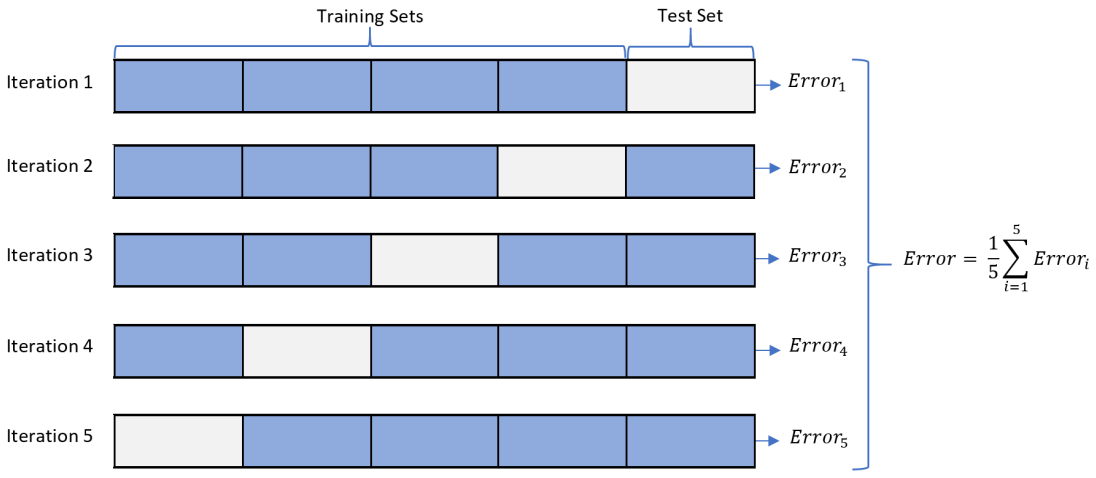

Figure 2.11: The Cross-Validation concept diagram for $k = 5$. Source: towardsdatascience.com

Optimization

Apart from data splitting, we should define the form of the utilized objective, or loss function, $L(\hat{y}_i, y_i)$, that must quantify the proximity of the prediction $\hat{y}_i$ to the target y_i. The loss function is task dependent and its minimization corresponds to having perfect predictions, i.e. $\hat{y}_i = y_i$. There is a large variety of proposed loss functions in the literature and the selection is critical for the effectiveness of the trained model. First and foremost, it should reflect the task's goal. For regression tasks, the most widely used loss function is the *Mean Squared Error* (MSE), as it can handle float differences:

$$L_{\text{MSE}}(\hat{y}_i, y_i) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2.39}$$

where N is the number of evaluated samples. On the other hand, for classification tasks we commonly utilize the -categorical- *Cross Entropy* (CE) loss:

$$L_{CE}(\hat{y}_i, y_i) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{2.40}$$

Of course, losses can be complex and can consist of multiple terms when considering multi-task problems. The only restriction is that the selected loss should be differentiable, so that a gradient-based optimization scheme (or *optimizer*) could be considered.

Convergence

The formulation to update the network parameters presented before is impractical as it is, since calculating the gradients over the entire dataset can introduce significant computational costs. To address this problem one can use the *Stochastic Gradient Descent* (SGD) optimization algorithm. SGD performs a parameter update for each training sample. The pairs of inputs and targets are fed to the SGD in a different sequence at each epoch (hence "Stochastic"). What is being used in practice, however, is the mini-batch alternative of SGD, where the gradients are computed over batches of samples. Additionally, SGD can be accompanied by a momentum strategy, where a history of previous gradients is used in order to avoid extreme oscillations or even getting stuck in local minima. A number of alternate optimization algorithms have been introduced to ensure or even accelerate convergence, the most popular of them being *Adaptive Moment Estimation*, or Adam [63]. Adam computes adaptive learning rates for each parameter and keeps an exponentially decaying average of past gradients, resulting in extremely faster convergence. Last but not least, *Batch Normalization* has been introduced [56] to assist convergence. It is used as a separate layer that constrains the range of input/output values by computing the running mean value and standard deviation, updated at each batch. The input is then normalized to approximate a standard normal distribution.

What happens, however, when the model does not converge as it should? There are cases where the class separation is highly dependent on the noise from training data, making it harder for the model to generalize in test time. This is called *overfitting* and has been a common pitfall of many machine learning algorithms. We can avoid overfitting by keeping well-defined data splits, as mentioned before, by data augmentation and the introduction of random noise into the model. A form of such noise in neural networks is the random zeroing of neurons, called Dropout [145]. It assists the creation of multiple information paths and avoids correlating a neuron with a specific input sample, thus enhancing generalization. Dropout is commonly used in many state-of-the-art models.

Evaluation Metrics

Once we have trained a machine learning model, the question is how well does that model behave and how we could quantify this performance. Depending on the task, one could utilize a bunch of different metrics. Here we will just review the fundamental *confusion metrics* that are used in our study and in evaluation protocols in general. Their concept arises from the following *confusion matrix* for a binary experiment:

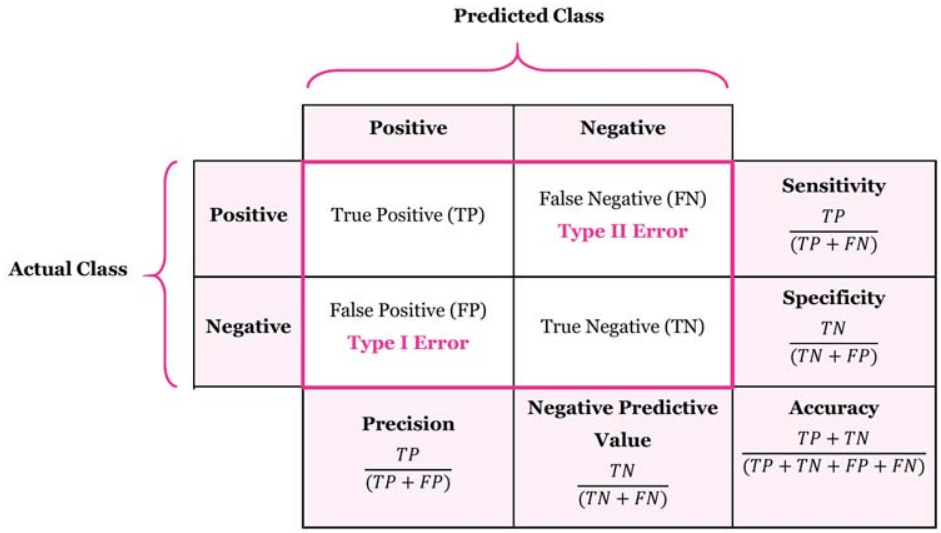

Figure 2.12: Confusion Matrix of a binary classification experiment, along with the definitions of most common metrics (Precision, Recall / sensitivity). From manisha-sirsat.blogspot.com

True Positives are the correctly predicted positive values and True Negatives (TN) are the correctly predicted negative values. Together they form the set of correct predictions. False Positives (FP) and False Negatives (FN) are the wrong predictions made for each class respectively. These 4 terms form the basic evaluation metrics:

- **Accuracy**: the ratio of the correct predictions to the total number of predictions. It is the most straightforward indicator, however misleading in class imbalance.
- **Precision**: the ratio of the correct positive predictions to the total number of positive predictions, measures the ability to correctly identify a class.
- **Recall**: the ratio of the correct positives to the total number of positive samples.
- **F1 Score**: the weighted average of Precision and Recall, used in the place of Accuracy in the case of imbalanced data for classification.

2.3 Deep into Neural Networks

Figure 2.10 depicts a FFNN comprised of more than a single hidden layer, three in specific. These networks that incorporate > 1 hidden layers in their structure are called *Deep Neural Networks* (DNNs) and the large subfield that is equipped with their analysis is called *Deep Learning*. While it has been long suggested, Deep Learning only recently emerged in the literature, taking advantage of the high computational capabilities of modern Graphical Processing Units (GPUs) to accelerate the training of heavy networks by distributing training data in multiple cores and processing them in parallel.

2.3.1 From Feature Engineering to Deep Learning

Deep Learning has revolutionized the common practice in Machine Learning Research and Applications, mainly due to the ability of multiple network layers to extract powerful features from input data. Hence they emerge as efficient feature extractors and help us not only by avoiding the costly procedure of handcrafted feature extraction, but also by providing us with features of even better quality. This feature extraction process is sometimes referred to as *Representation Learning*. Then, the bottleneck and output layers transform these features into class predictions, performing the classification stage. However, it soon became evident that plain DNNs could not analyze efficiently all types of raw data and especially complex forms like images, videos and time-series.

2.3.2 Convolutional Neural Networks

One of the breakthroughs in Image Analysis, and Machine Learning in general, was the introduction of *Convolutional Neural Networks* (CNNs) to efficiently handle image data and learn representative spatial features of high quality. Until then, researchers used multiple Computer Vision techniques, eg. SIFT [86], SURF [13], or simple MLPs to handle such data. However, the detectors were either too general or too over-engineered and hard to generalize. In addition, the amount of MLP weights rapidly became unmanageable when processing large images, while local information is not retained in them at all.

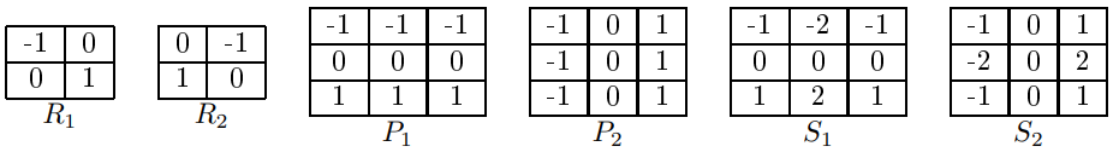

Figure 2.13: Kernels for Edge Detection: Roberts (R_1, R_2), Prewitt (P_1, P_2), Sobel (S_1, S_2)

CNNs employ representation learning based on convolutional kernels to automatically find relevant features that are spatially correlated. Such local kernels (or filters) were known to the research community and were applied by convolutions in order to provide handcrafted features like edges or corners (Figure 2.13). CNNs introduced specialized network layers and used trainable filters, which can generate discriminative feature maps that are optimized with respect to the task in hand. Such networks are commonly comprised of stacked convolutional layers, which perform the convolution operation $\mathbf{Y} = \mathbf{X}\mathbf{W}$, where $\mathbf{X}$ and $\mathbf{Y}$ are the input and output tensors respectively, while $\mathbf{W}$ is the kernel-weight tensor. Convolution is defined using the 2D cross-correlation operation:

$$\mathbf{Y}[m] = \sum_{n=1}^{C_{in}} \mathbf{X}[n] \star \mathbf{W}[m, n], \quad m = 1, \ldots, C_{out} \tag{2.41}$$

$$\left(\mathbf{Y} \in \mathbb{R}^{C_{out} \times H \times W},\ \mathbf{X} \in \mathbb{R}^{C_{in} \times H \times W},\ \mathbf{W} \in \mathbb{R}^{C_{in} \times C_{out} \times k_H \times k_W}\right)$$

The spatial dimensions $H \times W$ and $k_H \times k_W$ correspond to the the feature map and the kernel size respectively, while C_{in} and C_{out} correspond to the number of 2D feature maps on the input and output of the convolution (commonly called channels). Essentially, convolution layers transform an image feature map, taking into account contextual information about each pixel's neighborhood. Layers close to the input generate low-level features (eg. edges), while layers close to the output generate high-level features,

like complex shapes or texture. Research on CNN architectures has grown rapidly and various forms have been proposed (eg. VGG [140], ResNet [48]).

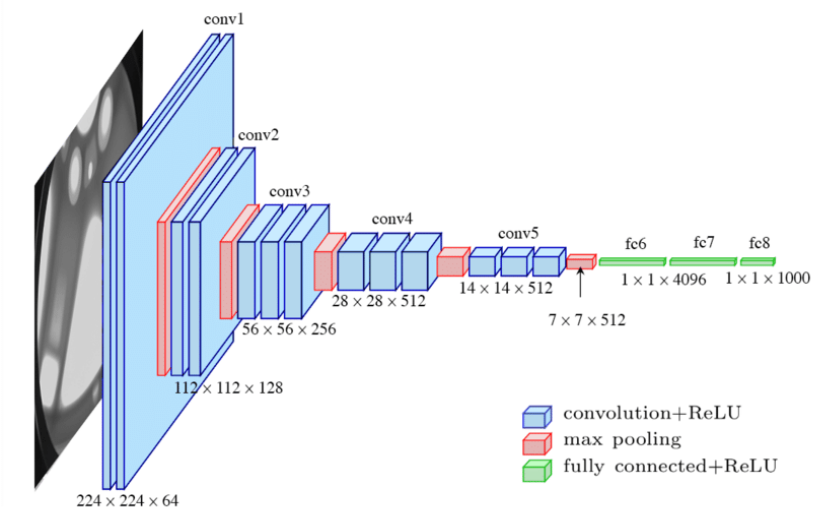

Figure 2.14: The architecture of an example VGG-16 [140] network for Image Analysis.

2.3.3 Recurrent Neural Networks

A *Recurrent Neural Network* (RNN) is a class of networks where connections between nodes form a directed graph along a temporal sequence. They are particularly useful where the underlying time dependencies are inherent in the nature of the input data. They are called *recurrent* because they perform the same task for every element of a sequence, depending the output on previous inputs. Thus, they demonstrate the ability to have a "memory" which captures the information calculated so far. As seen in Figure 2.15, RNNs' nodes are organized in successive layers. Given the input $\boldsymbol{x}_{t_{t=0}^{N}}$, where N is the length of the input sequence, each layer processes every input vector at time step x_t, outputs h_t (called the hidden state) and forwards both to the next step. Formally, at each time step t, the equations that describe the RNN function are:

$$\begin{aligned} h_t &= q\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right) \\ y_t &= r\left(W_{hy}h_t + b_y\right) \end{aligned} \tag{2.42}$$

where y_t is the output vector at timestep t, b_h is the bias for h, b_y the bias for y and q, r the activation functions for x and h respectively. Finally there are three parameter matrices: W_{xh} (input-to-hidden weights), W_{hh} (hidden-to-hidden), and W_{hy} (hidden-to-output).

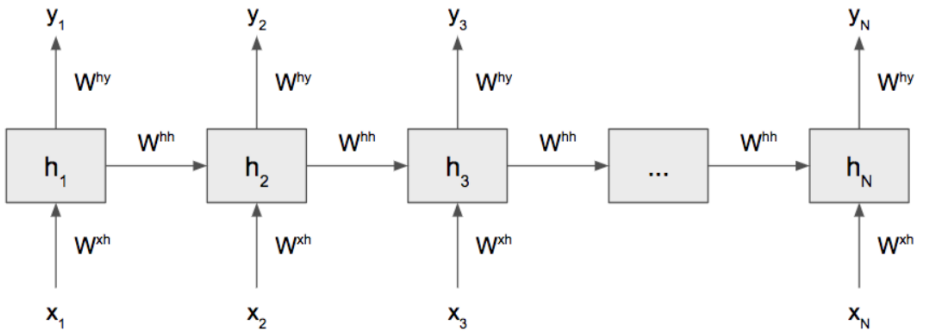

Figure 2.15: Diagram of an RNN cell's structure.

Theoretically, RNNs are able to model arbitrarily long dependencies between the input data. However, the nature of back-propagation training yields the problem of vanishing or exploding gradients. Precisely, the gradient computation over multiple timesteps tends to vanish or explode due to the finite-precision calculations when the error is propagated backwards. LSTM [51] and GRU [26] modules have been proposed to tackle this problem. Their core functionality is to control the magnitude of gradients via a forget gate that controls the informational flow from the networks' memory. The block diagrams of an LSTM and a GRU cell are displayed in Figure 2.16. These modules have been thoroughly successful, so vanilla RNN networks are now only rarely employed.

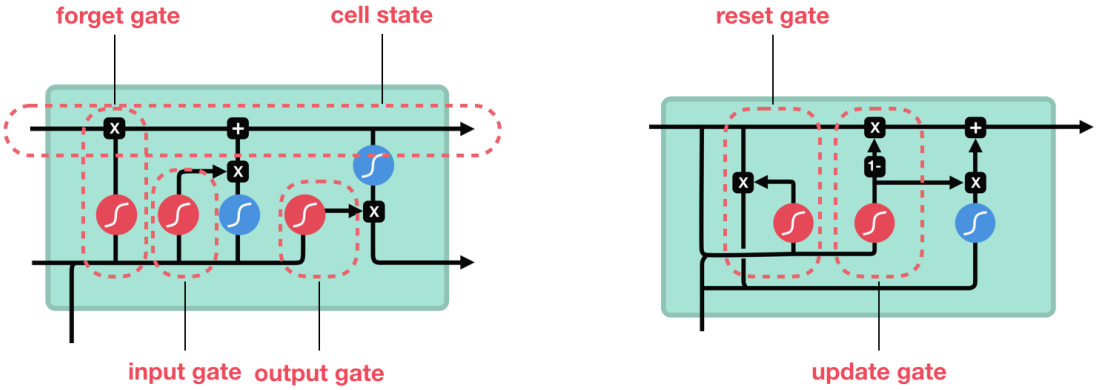

Figure 2.16: Structure of LSTM (left) and GRU cells (right). Red: Sigmoid, Blue: Tanh. From towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation

2.3.4 Multimodal and Metric Learning

Our experience of the world is rather multimodal: we see objects, hear sounds, feel the texture, smell odors and taste flavors, so that, in the end, we come up to a decision. Multimodal Learning, in its broader sense, suggests that when a number of our senses are being engaged in the processing of information, we understand and remember more. By combining these modes, learners can combine complementary information from different sources. For example, an image depicting a football match would be complemented by a transcription of a speaker's description of the highlight.

Deep neural networks have been successfully applied to feature learning for single modalities, as we have already analyzed. Here, we aim to fuse information from different modalities to improve our network's predictive ability. The overall task can mainly be divided into three phases: individual feature learning, information fusion and testing. A first step is learning how to represent input modalities and summarizing the data in a way that expresses the multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representations. A second step is to address how to translate one data modality to another. Not only are the data heterogeneous, but their relationship is often subjective. To tackle this challenge, we need to measure the similarity between different modalities and deal with possible long range dependencies and ambiguities. One of the most popular similarity metrics applied in such tasks is called *Canonical Correlation Analysis* (CCA). Given a pair of sample vectors (x, y), CCA aims to find a pair of linear projections of the two views $(\boldsymbol{w}_x^T x, \boldsymbol{w}_y^T y)$ that are maximally correlated, i.e.,

$$(\boldsymbol{w}_x^*, \boldsymbol{w}_y^*) = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \text{corr}(\boldsymbol{w}_x^T x, \boldsymbol{w}_y^T y) = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T \Sigma_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T \Sigma_{xx} \boldsymbol{w}_x \boldsymbol{w}_y^T \Sigma_{yy} \boldsymbol{w}_y}} \tag{2.43}$$

where Σ_{xx}, Σ_{yy} the covariance matrices of the samples and Σ_{xy} their cross-covariance matrix. Inspired by deep representation learning, *Deep Canonical Correlation Analysis* (DCCA) [4] was introduced to learn complex nonlinear transformations for different modalities, such that the resulting representations would be highly correlated. DCCA computes the latent representations of the two views by passing them through multiple stacked fully-connected layers, optimizing them through a CCA objective.

We then need to build best-suit models to extract features from individual modalities. Feature extraction from each source is usually independent from the others, at least initially. For example, in image-to-text translation, the features extracted from images are in the form of finer details, like edges and environmental surroundings, while corresponding features extracted from text are in form of tokens. After all the important features are extracted from all data sources, we then fuse them into a shared representation. This step can take several possible forms, from simple feature concatenation to complex feature-wise transformations, that can also be learnable (FiLM layers [110]).

Metric Learning

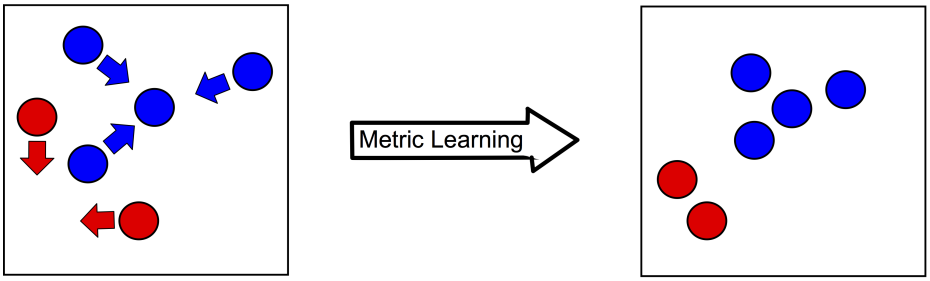

Figure 2.17: Concept of a contrastive objective on multimodal output embeddings. Source: laboratoirehubertcurien.univ-st-etienne.fr/en/teams/data-intelligence

However, information fusion is not achieved only through feature fusion. Instead, one could parameterize the loss function of a neural network to force an enhanced latent representation for all modalities by quantifying and extracting their similarity. The field that incorporates similarity objective functions is referred to as *Metric Learning* or, commonly, Deep Metric Learning, since it is applied predominantly on deep models. Metric Learning aims to learn feature embeddings in a way that reduces the distance between individual feature vectors that exhibit (commonly label) similarity (Figure 2.17) or as well increases the distance between dissimilar ones. The most prominent metric losses are contrastive losses, which evaluate how similar is a multimodal pair of samples, and triplet losses, that work on triplets of samples. We will make extensive use of triplet losses in Chapter 4, where we analyze this concept in further detail. The distance metric used can be the Euclidean distance, but, because we most times deal with high-dimensional vectors, other metrics are commonly used, such as the cosine or the Mahalanobis distance.

Chapter 3

Multifractal Analysis on EEG

In this chapter we examine novel algorithms in order to investigate the fractal and multifractal properties of EEG signals, as well as to what extent these properties carry emotional information. Emotion Recognition from EEG signals has been densely researched (Section 1.1), their complex and noisy structure however has proven to be a barrier for traditional modeling methods. In the end, we indicate that multifractal analysis could serve as basis for the development of robust models for Emotion Recognition.

3.1 Literature Review

Although Machine Learning has made overwhelming progress in modeling rational intelligence, there are still many challenges in approaching emotion-driven intelligence, a fundamental aspect of human's perception and decision-making processes. The reason for this is that emotions are highly subjective, and thus really difficult to be labeled when expressed, as analyzed in Chapter 1. Nevertheless, there is a growing interest in emotion tagging through physiological signals [24], since those are induced without our active interference and thus depict more clearly the actual affective state.

The electroencephalogram (EEG) is the most widely researched signal of its kind and has been highly effective in detecting affective states. A variety of time, frequency and joint-domain features have been extracted from EEG for that purpose. Indicatively, Petrantonakis et al. [111] introduces an adaptive filtering method to efficiently extract emotion-related characteristics from decomposed EEG signals. In addition, Higher Order Crossings of the signals were employed as feature vectors, with the overall framework achieving robust performance. Wang et al. [167] utilizes the power spectrum of separate EEG rhythms, along with statistical features, to recognize 4 emotion states: joy, relaxation, sadness, and fear. Other studies incorporate time-frequency and often wavelet features, like wavelet energy and entropy [57], achieving competitive results on the DEAP Dataset. Particular attention has also been given to channel connectivity features: Piho et al. [112] uses mutual information to extract informative EEG segments for Emotion Recognition, whereas in [38] the authors compare differential entropy features and their combination on symmetrical electrodes with traditional frequency and energy features, reporting high scores in literature. Nowadays, various types of deep neural networks have exceeded the performance of traditional feature-oriented methods. One of the most prominent efforts in sequential modeling is the bi-hemispheric model [79], proposed to process asymmetrical features through directed recurrent networks. CNNs have also been employed to capture both temporal and inter-channel spatial information [168].

However, processing EEG signals and extracting useful features remain core challenges, since EEG, like most biological signals, is chaotic, nonlinear and incorporates a large amount of noise, both from the recording equipment and interfering physiological processes [67]. Because of the nature of such signals, several nonlinear fractal methods have been proposed, one of them being the Higuchi Fractal Dimension (HFD) [50], which has been used extensively in emotion recognition. Liu et al. [83] utilizes the HFD spectrum in conjunction with statistical measures, while the authors of [156, 80] include HFD among several non-linear features, to classify emotions using random forests. Yet, due to the complexity of the EEG [131], such signals do not always share the same structure over every time scale, hence the fractal characteristics may vary and change dynamically or accordingly to the examined scale. For this reason, we propose the Multiscale Fractal Dimension [90] and Multifractal Detrended Fluctuation Analysis [62] to examine the EEG signals and determine emotional information buried in their fragmented structure. Until now, several studies have proposed multiscale fractal features for speech extraction[93, 113] as well as for identification of the speaker's affective state [23]. These take into consideration the inherent turbulence during speech production [93].

3.2 Multifractal Algorithms

3.2.1 Multiscale Fractal Dimension

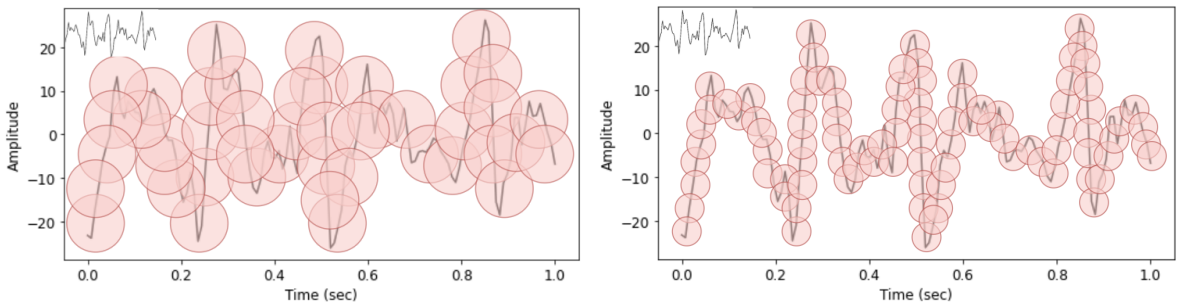

Figure 3.1: Minkowski cover simulation of a sample EEG signal from DEAP. At the up-left of both diagrams we depict the raw EEG sample waveform.

While many ways have been proposed to measure the fractal dimension (Section 2.1.4), chaotic and non-linear signals like the EEG and other physiological signals cannot be adequately modeled. Due to their complexity, such signals do not always share the same structure over every time scale, hence the fractal characteristics may vary and change dynamically or accordingly to the examined scale. Multiscale analysis should be thus considered to model these signals. Maragos [90] developed an efficient algorithm to measure a Multiscale Fractal Dimension (MFD), based on the Minkowski-Bouligand Dimension. This algorithm measures the multiscale length of a curve by covering it with disks of varying radius, whose center lies on the curve, referred to as the *Minkowski cover*. The developed algorithm is known as the *morphological covering method*:

1. Create the Minkowski cover (Figure 3.1) by using 2D morphological set dilations of the graph F of the signal by multiscale versions $sB = \{sb : b \in B\}$ of a unit-scale convex symmetric planar set B, where $s \geq 0$ represents the scale parameter:

$$F \oplus sB = \{z + sb \in \mathbb{R}^2 : z \in F, b \in B\} \tag{3.1}$$

2. Compute the cover area $A_B(s) = area(F \oplus sB)$. The Fractal Dimension is then:

$$D = \lim_{s \to 0} \frac{\log[A_B(s)/s^2]}{\log[1/s]} \tag{3.2}$$

Ideally B is a unit disk. However, D remains invariant as long as B is compact, convex and symmetric. In the discrete case, we select as B an approximation to the disk by a unit-radius convex symmetric subset of $\mathbb{Z}^2$. Now, Maragos has shown that the above limit will not change if we approximate $A_B(s)$ with the area of the difference between the morphological dilation and erosion of the N-sample discrete signal $F[n]$ by a function $G_s[n]$ that is the upper envelope of the discrete set sB:

$$A_B(s) = \sum_{n=0}^{N-1} ((F \oplus G_s) - (F \ominus G_s))[n] \tag{3.3}$$

for $s = 1, ..., s_{max} \leq N/2$. This greatly reduces the complexity by introducing one-dimensional signal operations, that are simple nonlinear convolutions, instead of complex two-dimensional set operations. Further reduction to linear complexity is accomplished by performing the above operations scale-recursive.

Fractal Dimension D can be estimated by least-squares fitting a straight line to and measuring the slope of the plot $\log[A_B(s)]$ versus $\log(s)$ because

$$\log[A_B(s)] = (2 - D)\log(s) + \text{constant} \tag{3.4}$$

assuming the power law $A_B(s) \approx s^{2-D}$ as $s \to 0$. Our signals however, and most real-world signals, do not have the same structure over every scale, hence the exponent in the dominant power s^{2-D} may vary. We therefore compute the slope of the data over a small scale window of w scales that move along the scale axis s $\{s, s+1, ..., s+w\}$, creating a profile of local MFDs $D(s,t)$ (or *fractogram*) at each time location t. The local slope is now an estimate of $2 - D$ and gives us the fractal dimension. D ranges between 1 and 2 and the larger it gets, the greater the amount of geometrical fragmentation of the signal. Figure 3.2 depicts sample EEG fractograms of a certain trial across three signal windows, where we indeed observe D to appear in the expected interval. It is also evident that D is high along every scale and even approaches $D = 2$ at the largest scales.

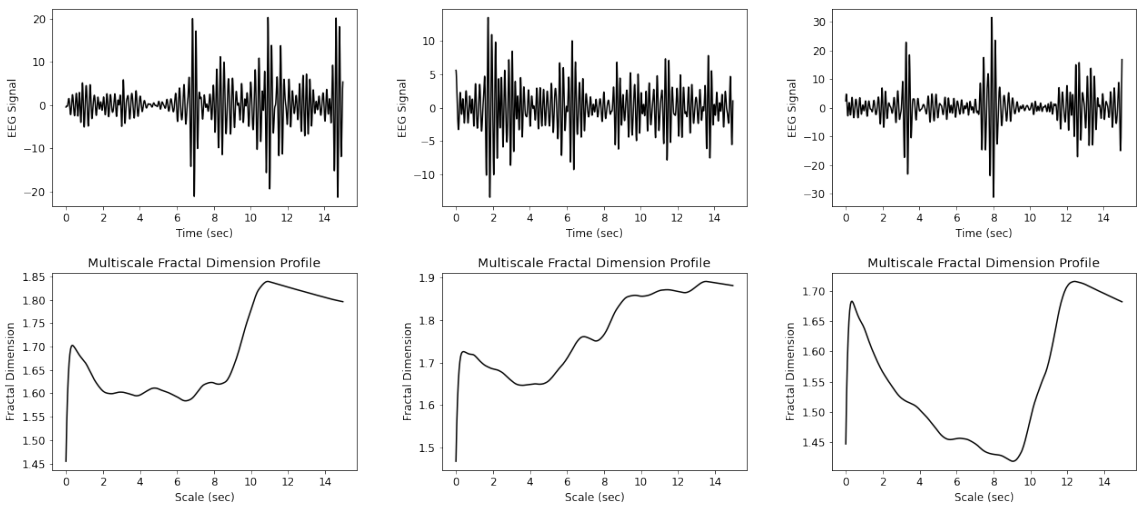

Figure 3.2: Fractogram of a sample EEG trial (Subject 8, Track 6, Channel CP1, α band, 0-15, 15-30 and 30-45 sec. windows). The utilized configuration is the one described in Section 3.4.2.

3.2.2 Multifractal Detrended Fluctuation Analysis

Detrended Fluctuation Analysis (DFA) estimates the Hurst exponent H in time series data instead of the fractal dimension. In general, the fractal dimension D presents local features of the signal whereas the Hurst exponent reflects on global properties of the time series. Additionally, if the time series is self-similar, the fractal dimension is easily derived from the Hurst exponent. The method takes a time series $x[n]$ of length N as input and consists of the following steps:

1. Initially, $x[n]$ is replaced by its centered cumulative sum: $y[n] = \sum_{m=1}^{N} (x[m] - \mu_x)$.
2. $y[n]$ is divided into N_s non-overlapping windows $y_N[k, n],\ k = 1, ..., N_s$ of length s.
3. For every window, the local trend $r[k, n]$ is obtained through linear regression.
4. $y_d[k, n] = y_N[k, n] - r[k, n]$ is the detrended version of the k-th profile segment.
5. RMS value of each detrended segment is computed and averaged across segments:

$$F(s) = \sqrt{\frac{1}{N_s} \sum_{k=1}^{N_s} F_k^2(s)}, \qquad F_k(s) = \sqrt{\frac{1}{s} \sum_{n=1}^{s} y_d[k, n]^2}. \tag{3.5}$$

The result of the above operations is a vector of s values, one for each chosen scale. The relationship between $F(s)$ and s is described by the power law $F(s) \propto s^H$, which determines H. In order to acquire its value, we plot these vectors in a (log-RMS, log-scale) plot. According to the power law, the points should form a line, which is estimated via linear regression and its slope determines the generalized Hurst exponent.

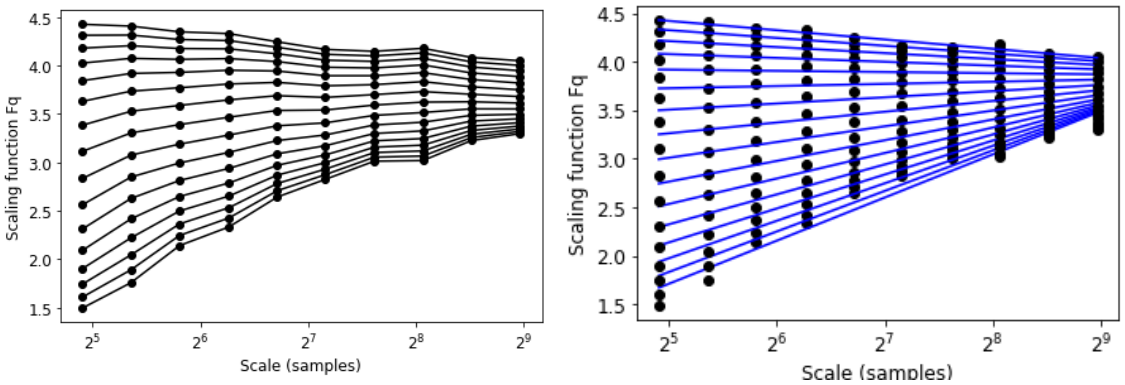

Figure 3.3: MFDFA on an EEG: depicting 16 DFA graphs for $F_q(s)$ along with the linear regression lines, the slopes of which determine the generalized Hurst Exponent $H(q)$.

Multifractal Detrended Fluctuation Analysis (MFDFA) [62] is essentially a generalization of DFA, where $F(s)$ is estimated over multiple moments of $F_k(s)$:

$$F_q(s) = \sqrt[q]{\frac{1}{N_s} \sum_{k=1}^{N_s} F_k^q(s)}. \tag{3.6}$$

As a result, a separate line is computed for every value of the factor q, with $q = 2$ being the reduction to classical DFA. MFDFA could prove especially useful in cases where the

scaling exponents and complexities are dependent on the scale, or change dynamically, in the context of time series. Figure 3.3 shows an EEG example for 16 q values.

Detrended Cross Correlation Analysis (DXA) [115] and Multifractal DXA (MFDXA) [179] are generalized versions of the above methods in the context of a pair of time series that we need to examine their structure and fractal correlations. Two time series $x_1[n]$, $x_2[n]$ of the same length are given as input and the sole change is the detrended covariance that is used in place of the squared detrended signal:

$$F_k(s) = \sqrt{\frac{1}{s}\sum_{n=1}^{s} y_{d1}[k,n]y_{d2}[k,n]} \tag{3.7}$$

Of course DXA reduces to DFA when $x_1[n] = x_2[n]$. Lots of other variants have been proposed regarding the MFDFA algorithm, one of the most prominent being the replacement of polynomial detrending with Empirical Mode Decomposition (EMD) [52], having proved to be efficient for biomedical time series that have oscillatory or ramp-like trends [77]. EMD is a way to decompose a signal into a small number of components that form a nearly orthogonal basis for the original signal, called Intrinsic Mode Functions (IMF).

3.3 EEG as a Multifractal Signal

Before presenting our analysis, we mention that we work on the DEAP Dataset [65] which we describe in detail in the Appendix. DEAP is a widely used data source for Emotion Recognition from physiological signals, including preprocessed data from 32 subjects. Each subject is exposed to forty 60-sec. long music videos as stimuli, while having their EEG recorded, along with other physiological signals. After watching each video, the subject rated their induced emotion in valence and arousal. Each video has been also separately annotated by the authors, this is however not yet among our concerns.

3.3.1 Stationarity of EEG Signals

Physiological signals and EEG are widely researched as noisy and non-stationary signals and commonly demand heavy pre-processing, since they normally exhibit time-varying oscillations. The observed structure is partly due to external stimuli or other physiological operations and mainly indicates the complexity and the states of neural assemblies during brain functioning [67]. In our experiments it is crucial to determine the stationarity of the signals in order to interpret their multifractal properties. We use the Augmented Dickey Fuller (ADF) Test [37] for that purpose, a test that depends on the concept of *unit root*. Suppose we have a time series $y_t = ay_{t-1} + e_t$ where y_t is the value at time t and e_t the error. If we solve the recursive formula we get:

$$y_t = a^n y_{t-n} + \sum_i e_{t-i}a^i \tag{3.8}$$

For $a = 1$ (unit root) the equation implies that the variance will monotonically increase in time. Thus, the signal will be non-stationary. The ADF Test is used to determine the presence of unit root in time series, the presence being the Null Hypothesis.

By applying the test to a limited but randomly sampled set of DEAP signals, to our surprise, we derived evidence of strict stationarity. Specifically, the examined signals

appeared non-stationary only at very low scales, up to windows of 100 samples or 0.8 sec. The same holds when we test the signal profiles, i.e. their cumulative sums. However, a few signals exhibit non-stationarities at their major frequency bands. In order to find the source of this stationarity, we reproduced the pre-processing applied in [65] to a sample raw waveform. This included downsampling to 128 Hz, eye-artefact removal, filtering at 4–45Hz and averaging to the common reference channel. After this procedure, it was found that the cause of stationarity was the performed bandpass filtering. As a second experiment we checked the stationarity of the frequency bands of the EEG (theta, alpha, beta, gamma). We found that, despite the bandpass filtering, a few signals indeed exhibit non-stationarities and, the lower the band, the more signals appear non-stationary. A significant difference is that by taking here cumulative sums as input, the vast majority of signals in every band turns non-stationary, whereas most of them exhibit a trend.

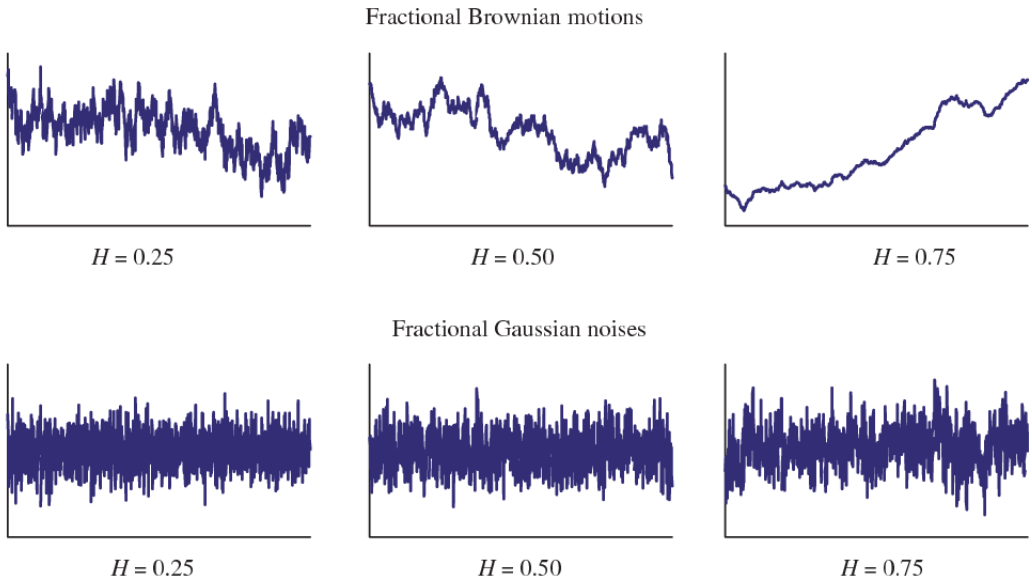

Figure 3.4: Graphical examples of fractal time series. The upper graphs represent fractional Brownian motions (fBm) and the lower graphs, the corresponding fractional Gaussian noises (fGn), for three typical values of the Hurst Exponent H: $0.25, 0.5, 0.75$. Source: [35]

3.3.2 Hurst Exponent Estimation

Scale-free stationary processes, like EEG signals, can be viewed as *fractional Gaussian noise* (fGn) while their increments typically construct non-stationary processes in the form of *fractional Brownian motion* (fBm) of the same Hurst Exponent. Thus, the exponent estimation is crucial in characterizing EEG signals for multifractal analysis [55] and can be determined by monofractal DFA. If the estimated exponent is less than $H = 1$, then it characterizes a stationary process, which can be modeled as fGn with that exponent. Otherwise, it is assumed to be produced by a non-stationary fBm process with an exponent of $H - 1$. The EEG signals of the DEAP dataset provide a very low Hurst Exponent value that approaches 0, while their profiles and separate bands provide an increased DFA-estimated exponent, still though below 0.2 at most cases. The results however alter when we examine the profiles of the filtered bands, particularly theta and gamma, in which the exponent estimation shows a steady increase (Table 3.1).

These values confirm the evidence from the ADF Test that EEG signals are negatively correlated and their fluctuations are smaller in larger time windows, which is the

Signal Resolution	Raw Signal	Theta	Alpha	Beta	Gamma
600 samples	0.83	0.24	0.90	0.57	0.36
whole signals	0.01	0.03	0.01	0.00	0.00
600 (cumsum)	1.28	0.73	1.33	0.92	0.41
whole cumsum	0.02	0.26	0.13	0.09	0.27

Table 3.1: Hurst Exponent of 27 randomly sampled DEAP EEG trials, computed through monofractal DFA and averaged, with respect to EEG channels and the number of trials.

typical behavior of fGn processes having Hurst exponents below 0.5 [35]. Only when considering short signal intervals we could measure actual high exponent values. As seen in Table 3.1, random signal windows of 600 samples introduce very different measures of complexity, especially the raw signal and the alpha band. This finding indicates that the EEG fragmented structure can vary significantly according to the examined scale, a desired property in order to apply multifractal algorithms. In Figure 3.4 we can see how sample fGn and fBm processes vary their structure according to their H Exponent.

3.4 Extraction of Fractal Features

Each 60-sec. EEG segment is partitioned into its main bands through band-pass filtering with a 10th order Butterworth filter. We include alpha (8-13 Hz), beta (14-29 Hz), and gamma (30-45 Hz) rhythms in our analysis, as well as raw signals, since those have been acknowledged as the most emotion-sensitive [178] and have shown the largest multiscale variability. We select 12 left (Fp1, AF3, F7, F3, FC5, FC1, T7, C3, CP5, CP1, P3, P7) and 12 right (Fp2, AF4, F4, F8, FC2, FC6, C4, T8, CP2, CP6, P4, P8) channels that have shown competitive performance, particularly when their asymmetry is examined, and we assess the proposed features on each set separately.

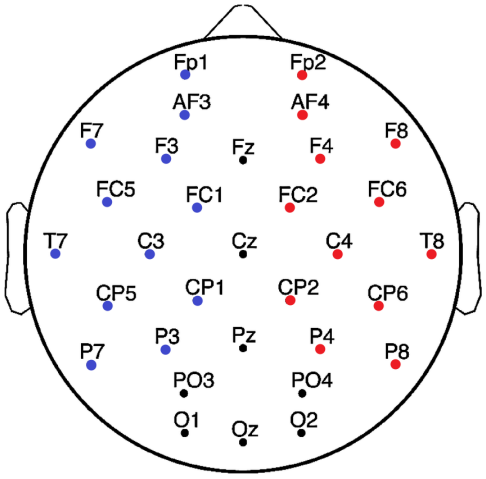

Figure 3.5: The configuration of the selected left (blue) and right (red) DEAP channels.

3.4.1 Baseline Features

A set of widely used baseline features is extracted for comparative reasons and to assess the combined efficiency of the proposed feature set. These features are

- **Power Spectral Density (PSD)**: As mentioned in Chapter 2, PSD denotes the Fourier Transform of the EEG signals' autocorrelation function. It has been widely exploited in feature extraction algorithms from EEG signals [105] since it is able to incorporate insights from the available EEG frequency bands. We compute PSD through the Welch [169] method, resulting in $N = fs/2 = 64$ features per signal.
- **Higuchi Fractal Dimension (HFD)**: As mentioned in Chapter 2, HFD [50] is an alternative derivation of the Fractal Dimension and it has been widely exploited to

analyze neuronal signals. We compute HFD using the PyEEG library [10], resulting in a scalar feature. In order to derive a feature vector for this case, we first split each signal into windows of 15 sec. (1920 samples) with 50% overlap and then we determine the HFD for each of the 7 windows, resulting in a 7D vector.

3.4.2 MFD Features

Since EEG signals show multifractal properties in rather short windows and, also, Multiscale Fractal Dimension has been mainly used for short-time analysis, we choose again to split each 60-sec. EEG signal into 7 windows of 15 sec., with 50% overlap. The proposed feature set includes 30 linearly sampled features, extracted out of each window's MFD. The respective per-window features are then summarized using 3 statistical metrics: their mean, median and standard deviation across the scale of measurement. In this way, we end up with a final 90D feature vector that incorporates the signal's temporal variance. Every signal is analyzed at discrete scales of $s = 1, ..., 274$ samples, thus the maximum scale is at $s = 1/7$ of the signals' length. The fractograms of sample signals along with the variance of their 7 windows are shown in Figure 3.6. The EEG fractograms reveal a highly fragmented structure and a high multiscale fractal dimension $D > 1.5$. This finding is consistent with the low Hurst Exponent estimations we got from the monofractal DFA trials, especially in the case of large signal scales. A common characteristic among these signals is that their MFD shows a steep peak at the first scales of 1-2 sec.

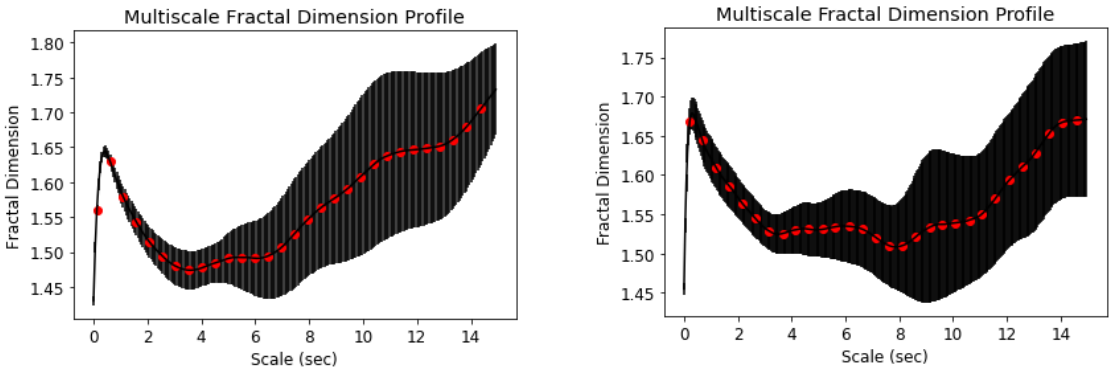

Figure 3.6: Sample MFD profiles of 2 EEG signals (Subjects 5, 20) along with the mean and standard deviation features extracted from their 7 subsignals.

3.4.3 MFDFA Features

We additionally acquire 30 features from processing the last half of each EEG waveform through the computationally expensive MFDFA. We select 10 scales ranging from 30 to 500 samples in a logarithmic scale, along with 16 q-moment values ranging linearly from -5 to 5. The resulting representation is a set of 16 linear-like graphs of 10 values, as shown in Figure 3.3. Sixteen (16) Hurst Exponent values are determined through linear regression, one for each moment. The mass exponent t is then derived through $t(q) = qH(q) - 1$. A monofractal signal with constant H would produce a linear graph, since H remains constant, the EEG instead produces a curve that we utilize to produce the signal's *multifractal spectrum* D, characterized by its cap scheme:

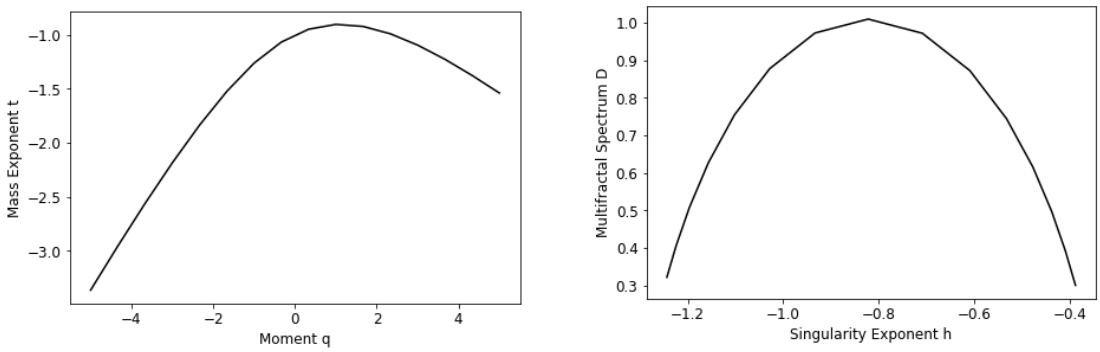

Figure 3.7: $t(q)$ and $D(q)$ MFDFA components for a sample EEG signal.

We derive the multifractal spectrum $D(q)$ through the following equations:

$$D(q) = q'h(q) - t(q'), \qquad h(q_n) = \frac{t(q_n) - t(q_{n-1})}{q_n - q_{n-1}}, \tag{3.9}$$

where $n = 1, ..., 15$, q' excludes the largest moment value, and $h(q)$ is the singularity exponent. The resulting curve, determined by 15 $h(q)$ and 15 $D(q)$ values, represents the MFDFA feature set. In Figure 3.7 we depict $t(q)$ and $D(q)$-$h(q)$ for a sample EEG.

3.5 Experimental Evaluation

We evaluate the features extracted from the multifractal analysis on the emotion recognition task. The experimental protocol can be divided into two categories: *Subject Dependent*, where a unique classifier is trained and tested on the trials of each participant, with the final score being the average per-subject score, and *Subject Independent*, where a classifier is trained on several participants and tested against unseen trials.

3.5.1 Model Formulation

We make use of a single classifier unifying features from all available EEG channels. The model consists of a Standard Scaler, that standardizes training features by removing their mean and scaling them to unit variance, and a Support Vector Machine (SVM) with a radial basis function (RBF) kernel. Experiments consider single labels, i.e. valence or arousal, in binary format by setting the threshold for binarization in the median score 5. We perform 5-fold cross validation on stratified splits of the available data: approximately 56.5% of all samples are of high valence and 59% of high arousal annotations.

3.5.2 Results & Discussion

The classification results for all features at the 2 distinct settings are summarized in Tables 3.2 and 3.3. We notice the accuracy difference between subject dependent and independent tasks, supporting the claim that brain responses inherit mainly subjective characteristics. The EEG PSD is shown to be efficient in the subject-dependent setting, where the raw signal modality achieves 64.2% in valence and 65.2% in arousal. Interestingly, despite the raw signal features are more efficient when taken from the left selected channels, the per-band scores are actually slightly better for the right channels. All these scores significantly drop in the subject-independent setting, where the PSD emerges as the

least efficient feature set, achieving only chance-level scores in arousal, 6% below the top recorded accuracy of MFD. We can therefore assume that the within-subject variability is concentrated more on separate spectral characteristics of each participant and therefore, fractal analysis can prove to be more robust across different subjects.

Features	Channels	Raw Signal	Alpha Band	Beta Band	Gamma Band	Combined
PSD		**0.642** — 0.652	0.598 — 0.645	0.629 — 0.639	**0.635** — 0.620	0.631 — 0.648
HFD	Front	0.615 — 0.638	0.605 — 0.655	0.591 — 0.643	0.601 — 0.634	**0.638** — 0.645
MFD	Left	0.620 — 0.661	**0.626** — **0.669**	0.591 — **0.653**	0.594 — 0.636	0.612 — **0.661**
MFDFA		0.577 — **0.662**	0.571 — 0.643	0.577 — 0.649	0.592 — **0.651**	0.586 — 0.658
PSD		0.627 — 0.644	0.616 — 0.645	**0.637** — 0.641	0.623 — 0.627	0.623 — 0.646
HFD	Front	0.606 — 0.644	0.604 — 0.655	0.595 — 0.633	0.572 — 0.627	0.623 — 0.644
MFD	Right	0.607 — 0.655	0.605 — 0.652	0.566 — 0.652	0.602 — 0.641	0.597 — 0.657
MFDFA		0.587 — 0.655	0.573 — 0.641	0.603 — 0.650	0.573 — 0.620	0.586 — 0.652

Table 3.2: Subject Dependent Task Accuracy in the form: Valence — Arousal

Features	Channels	Raw Signal	Alpha Band	Beta Band	Gamma Band	Combined
PSD		0.554 — 0.569	0.547 — 0.564	0.549 — 0.562	0.553 — 0.570	0.546 — 0.564
HFD	Front	0.541 — 0.601	0.552 — 0.588	0.541 — 0.616	0.545 — 0.584	**0.585** — **0.621**
MFD	Left	0.553 — 0.606	**0.566** — **0.631**	0.545 — **0.618**	**0.554** — 0.580	0.559 — 0.615
MFDFA		**0.569** — **0.630**	0.546 — 0.600	0.545 — 0.598	0.532 — 0.545	0.553 — 0.608
PSD		0.553 — 0.580	0.557 — 0.560	**0.558** — 0.573	0.552 — 0.579	0.555 — 0.575
HFD	Front	0.525 — 0.573	0.566 — 0.582	0.544 — 0.595	0.549 — 0.567	0.571 — 0.605
MFD	Right	0.552 — 0.601	0.556 — 0.605	0.547 — 0.587	0.545 — **0.588**	0.560 — 0.607
MFDFA		0.555 — 0.619	0.552 — 0.580	0.549 — 0.591	0.539 — 0.584	0.544 — 0.599

Table 3.3: Subject Independent Task Accuracy in the form: Valence — Arousal

Multifractal methods show indeed strong performance in both experiments, surpassing chance levels and the baseline features in most cases. In contrast to spectral features that are sensitive to valence, these features prove efficient mainly in recognizing the arousal state, in which they achieve 5% to 7% higher scores, in the Subject Dependent setting, and approximately 3-4% higher scores in the Subject Independent one, compared to the PSD features. Here, the left hemisphere is shown clearly as more effective than the right one. An important distinction, compared to other studies, is that the gamma band does not play a dominant role in recognition performance. At least in the present study, gamma band features perform well in predicting valence from PSD features and arousal from MFDFA features. The HFD feature, on the other hand, while being effective in both settings, it falls behind the multiscale dimension MFD in most experiments. Despite reporting the same high-level concept of a fractal dimension, HFD inherits a significantly lower dimensionality as a feature vector and also lacks the analysis of multiscale variablity that we have observed on the EEG signals, a type of structure variability that is adequately modeled through multiscale versions of their fractal dimension.

Our results are in accordance with those reported in [80] for PSD and HFD, while the top scores obtained by MFD and MFDFA surpass most of the ones reported there (Table 3.4). At the SD experiment particularly, MFD of the alpha band and MFDFA at the raw signal yield 66.9% and 66.2% respectively, whereas their highest subject-independent accuracy yields 63%. These scores are among the state-of-the-art results in the specific dataset [28], considering feature-oriented studies, although we recognize the additional difficulty of eliminating all of the trials of a tested participant from training.

Proposed Features	SD Valence–Arousal Scores
Power Spectral Density [80]	0.617 – 0.621
Higuchi Fractal Dimension [80]	0.632 – 0.622
Higher Order Crossings [80]	**0.647** – 0.661
Multiscale Fractal Dimension (ours)	0.626 – **0.669**
Multifractal DFA (ours)	0.603 – **0.662**
Pearson Correlation Coefficient [80]	0.688 – 0.682
Rational Asymmetry [80]	0.611 – 0.626
Mutual Information [80]	0.708 – 0.687

Table 3.4: Performance Accuracy at the Subject Dependent Setting for various feature types, presented in [80]. Importantly, features that aggregate the inter-channel correlations seem to capture the most discriminating emotional information.

Aggregating Features

As shown in Table 3.4, inter-channel correlations seem to be highly discriminative on the affective state and provide the best feature-wise accuracy scores. While the intent of our study is not to concentrate on such correlations, but rather on the fragmented structure of the EEG, an interesting experiment would be the examination of the inter-channel multifractal correlations through the Multifractal DXA algorithm. However, the high complexity of the algorithm, having to quantify nearly hundreds of channel combinations, as well as the length of the DEAP EEG signals made such an experiment computationally inefficient. It would certainly be though a promising future direction to examine. Other than that, correlation measures between the extracted features of the left and the right hemisphere (e.g., differential asymmetry, rational asymmetry) confirm the finding that no further emotional information can be obtained, as their performance falls between the scores of the independent experiments, still substantially higher (by 2-5%) than plain asymmetry measures, utilized in [80].

Another interesting experiment is the aggregation of the different fractal features. MFD and MFDFA clearly outperform the Higuchi baseline in arousal and perform comparably in valence, indicating that the multiscale variability of the EEG can capture latent emotional information. However, their combined usage reduces performance towards MFDFA-reported scores, indicating lack of complementary information. To this end, we performed additional MFDFA experiments to determine the source of the insufficiency, compared to MFD features. Specifically, we tried out input types that alleviate the effect of the very low Hurst Exponent of the DEAP signals, since it has been stated [55] that MFDFA can be significantly harmed by extreme exponent values. The most prominent input type is the cumulative sum of the EEG, that we use to provide random walk transformations of the EEG signals, in order to increase their H exponent. The results, shown in Table 3.5, do not indicate a specific pattern, however we can deduce that brain rhythms of higher frequency tend to benefit from these configurations.

Features	Experiment	Raw Signal	Alpha Band	Beta Band	Gamma Band
Left	Subject	0.566 — 0.652	0.566 — 0.642	0.581 — 0.640	**0.601** — **0.659**
Right	Dependent	0.580 — **0.655**	0.577 — 0.632	**0.606** — 0.650	0.576 — 0.655
Left	Subject	**0.556** — **0.639**	0.548 — 0.612	0.540 — 0.600	0.550 — 0.581
Right	Independent	0.541 — **0.610**	0.548 — 0.595	**0.555** — 0.598	0.540 — 0.584

Table 3.5: MFDFA on signals' cumulative sum Valence–Arousal Accuracy

Returning to our base results, although the two kinds of fractal dimensions, HFD and MFD, attempt to record the same signal quantity, they differ in their analysis of the signal complexity in multiple scales, thus provide different measurements. While MFD outperforms HFD and performs strong in arousal, the two features provide better scores in arousal when combined. As we depict in Table 3.6, in both subject dependent and independent settings we record higher accuracy than the one we obtained from the individual features in the base study, mainly when testing raw signals or the alpha band. The differences are significant in the subject independent setting, whereas even the top scores obtained previously are improved. The model can now predict arousal at 67% and 64% at subject dependent and independent experiments, respectively.

Features	Exp	Raw	Alpha	Beta	Gamma	Comb
Left	Subject	**0.663**	**0.670**	**0.657**	**0.637**	0.656
Right	Dependent	0.654	**0.662**	0.618	0.640	0.655
Left	Subject	**0.613**	**0.641**	0.612	0.580	0.614
Right	Independent	**0.604**	**0.610**	0.591	0.582	**0.615**

Table 3.6: MFD-HFD Combined Arousal Accuracy

Other than the above mentioned selected cases, it seems that the selection of a single feature type could be adequate and preferable for affective state recognition, since aggregated sets between the mentioned feature types do not provide a statistically significant improvement in recognition. Moreover, our indications regarding the optimal brain rhythms and channels for the task should be taken into account, since we do not observe any substantial improvement for the "Combined" classification category, in which we measure the aggregated performance of the three bands and the raw signal.

Summary

In this section we analyzed the multiscale fractal structure of EEG signals and proposed a feature extraction method utilizing two multifractal algorithms for emotion recognition, that can meet the needs of the observed multiscale variability in the EEG structure. The proposed features perform strongly against the baselines, particularly in the challenging subject-independent setting and in arousal recognition, indicating that arousal is correlated with the fragmented structure of the EEG. Further improvements are achieved when the fractal dimension features are aggregated, while the efficiency of the alpha frequency band is underlined in all experiments. Our analysis showed that multifractality and the anti-correlation properties could be considered when processing EEG signals.

Chapter 4

EEG & Music Cross-Modal Learning

Cross-modal Learning aims to extract the semantic correspondence between different types of data and project it onto a common space so that an input from one modality can retrieve information about the other. Here we focus on modeling the relationship between pairs of music tracks and corresponding EEG recordings. We propose a framework that can be utilized for emotion recognition both directly, by performing supervised predictions, and indirectly, by providing relevant music samples from EEG given inputs. By applying this system independently to all 32 subjects of the DEAP Dataset we extract useful insights regarding emotion perception and how the human brain processes music.

4.1 Literature Review

4.1.1 Music Cognition

Studying the human brain's responses to music stimuli has always been a lively field of research in neuroscience and signal processing [136] that aims to answer fundamental questions regarding our enjoyment of music (Section 1.3). The field has gained a lot of attention in recent years, with the upsurge in available neuronal data. Most studies in the field rely on electroencephalography (EEG) recordings as they provide better temporal resolution than other techniques, such as functional magnetic resonance imaging (fMRI). In addition to the traditional, well-controlled auditory experiments, modern approaches gather physiological data from music listeners as they enjoy or imagine naturalistic music. Examples include the NMED-T Dataset [85] that aims, along with other studies (Stober et al. [146], Vinay et al. [157]), to capture beat information, and OpenMIIR [147], that includes sessions of subjects imagining made-up and naturalistic music samples.

One of the core findings that lead the research on music cognition is the correlation between the frequency and magnitude of neural oscillation patterns and rhythmical patterns in music [103]. In specific, alpha band [158] and beta band oscillations [42] have been thoroughly examined for this task. Toiviainen et al. [153] suggest that the auditory cortex is involved in the processing of musical features during continuous listening to music. Others found that musical features related to timbre and rhythm are also processed in the superior temporal gyrus (STG) [130] and Heschl's gyrus [141]. This specific study has also examined the unfolding of musical emotions and their temporal attributes, something that will be of our interest in the upcoming experiments. Brain connectivity patterns are important in this process, as also shown by Menon et al. [97], in correlating brain structures with music listening. Additionally, Event-Related Potentials (ERP) have

been utilized to extract brain activity patterns that can relate to the structure of musical events, such as note onsets or pitch [134, 116]. Last, although in the following we avoid cross-subject experiments, it should be noted that inter-subject correlations of listeners have been useful as well for brain mapping during listening [170].

Several approaches have been taken to predict emotions encoded in and conveyed by music [54]. As mentioned in Section 1.1.3, researchers have used chords and other auditory features [107, 108], and have included multimodal approaches, like combining audio features with information from the lyrics [47], in Music Emotion Recognition (MER). In parallel to feature-oriented approaches, there has also been a shift towards deep learning based approaches for information retrieval from music stimuli [53], in which we concentrate in the present study. Undoubtedly, the most powerful impact of music on humans is the induced emotions, thus Emotion Recognition is deeply researched both by Neuroscience and Behavioral Signal Processing [132]. Several works have also studied affective musical features [144] (Section 1.3).

4.1.2 Cross-Modal Learning

The task of learning a shared embedding space from different datasets or modalities is being studied through a variety of approaches, which are predominantly applied to image and text modalities [164]. A widely used baseline is Canonical Correlation Analysis (CCA). CCA, as mentioned in Section 2.3.4 is non-probabilistic and enables the extraction of linear components to optimize the correlation of pairs of vectors. One can find in the literature various non-linear CCA-based frameworks and neural networks, utilized to learn inter-modal similarities. The most prominent examples of this include Deep CCA (DCCA) [4], which utilizes processing through neural networks as an intermediate step to calculating the correlation of a pair of data, and Deep Canonically Correlated Autoencoders (DCCAE) [165], a similar technique which further enables cross-signal reconstruction. Recently, a new variant has been proposed in order to assist cross-reconstruction tasks between modalities, called Variational CCA (VCCA) [166]. This study utilizes the concept of a Variational Autoencoder, that attempts to form a meaningful embedding space, in order to enhance this space by CCA objectives.

Besides CCA, other methods that have been used include an HGR-based maximal correlation metric [75], that aims to provide correlation measures based on multiple, non-linear views of the data, and adversarial training [161], focusing mainly on the optimization function of the respective model. Moreover, there have been proposed additional methods to construct a binarized space for the modalities [164], using techniques from another major category of Representation Learning that has to do with learning binarized (or Hammming) latent spaces. CCA in combination with deep networks has been used to model a shared semantic space between audio and EEG signals [125]. In another study, Li et al. [74] incorporated music to co-train a shared space with images using a contrastive loss. Further, in [177] the authors use a state-of-the-art framework for the cross-modal task and indirectly manipulate the latent space via label supervision, a key concept that we also follow in our study, to provide more comprehensive measures.

4.2 Bridging the Semantic Gap

We choose to study brain responses to music by employing a cross-modal system to identify and analyze the correspondence between music and EEG modalities. We also constrain the learning process with emotion labels, therefore aiming to derive important insights regarding the affective role that music can play on humans. To conduct this experiment we exploit multimodal optimization strategies to extract EEG and music features in a common latent space, from which we could assess their similarity. By providing an EEG input as a query, the model should retrieve the most similar music embeddings.

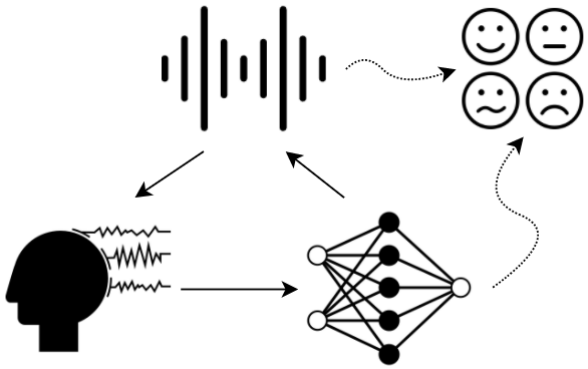

Figure 4.1: Concept of the proposed model: By using EEG data of music listening we attempt to derive embeddings that could resemble the stimulus and the music-induced affect.

4.2.1 Problem Formulation

Let us assume a collection of n instances of EEG-music pairs, denoted as $T = \{(x_i^a, x_i^b)\}_{i=1}^n$ where x_i^a is the input EEG sample of the ith instance and x_i^b the input music stimulus of that sample. Each instance has been assigned an affective annotation $y_i \in \mathbb{R}^2$ for valence and arousal dimensions. Instances of the same pair do not need to have the same affective labels. For each instance, i, we aim to learn an embedding form $u(i) = f(x_i^a, Y^a) \in \mathbb{R}^d$ for the EEG and $v(i) = g(x_i^b, Y^b) \in \mathbb{R}^d$ for the music modality, where d is the dimensionality of the common representation space and Y^a, Y^b the trainable parameters of the two functions, that satisfy the following properties: a) the similarity of samples from the same category is larger than the similarity of samples from different categories, and b) the intra-pair similarity is also larger than the similarity of other random pairings. The EEG latent representation, the music representation and the label matrices for all instances in T are denoted as $U = [u_1, u_2, ..., u_n],\ V = [v_1, v_2, ..., v_n],\ L_a = [l_{a1}, l_{a2}, ..., l_{an}],\ L_b = [l_{b1}, l_{b2}, ..., l_{bn}]$ respectively.

4.2.2 The Issue of Input Representations

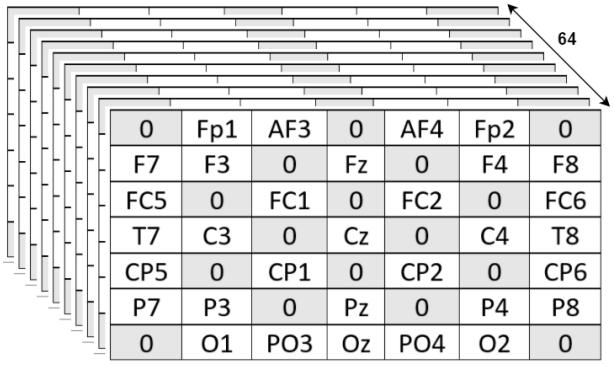

Figure 4.2: EEG input shape that resembles the channel topology on the scalp.

The algorithm to extract input features for each modality is crucial, in order to derive meaningful embeddings that could be correlated in a common EEG-Music embedding space. For the EEG modality we take advantage of the data dimensionality for each trial (channels, timesteps) by arranging the available channels in a two-dimensional grid form that resembles their topology on the human scalp, as shown in Figure 4.2. This way we can employ a compatible network to analyze both spatial and temporal information through 3D inputs of the form (grid-x, grid-y, timesteps). On the other hand, the available music signals are rather limited, being only the stimuli of the corresponding EEG trials. Hence, we choose to provide pre-trained music embeddings as input to our framework and to this end we utilize the MusiCNN [118] model.

MusiCNN is a robust CNN network that is pre-trained on the Million Song Dataset (MSD) [16], takes as input log-mel spectrograms of 3-sec. music signals and produces high-quality music embeddings. The MSD Dataset considers a 50-tag label vocabulary, with tags including rock, pop, dance, metal, male/female vocals, 80s, instrumental, indie, happy etc. Pre-training on these labels using a large amount of music data will compensate for the limited size of our track set and will further assist the cross-modal task. The utilized model contains a musically motivated CNN [117] that consists of a convolutional layer with several filter shapes and receptive fields to capture musically relevant context. It also includes a set of 64 densely connected layers, in charge of extracting higher-level embeddings from the low-level CNN features. These layers incorporate residual connections, in order to aggregate information from different hierarchical levels. Finally, MusiCNN has a temporal pooling module of 200 units, responsible to produce the output tags (taggram) from the extracted features. We specifically use the "max-pool" embeddings, produced just before the reduction to the output tags. A thorough description of MusiCNN model can be found at github.com/jordipons/musicnn/blob/master/musicnn_example.ipynb.

4.2.3 The Proposed Framework

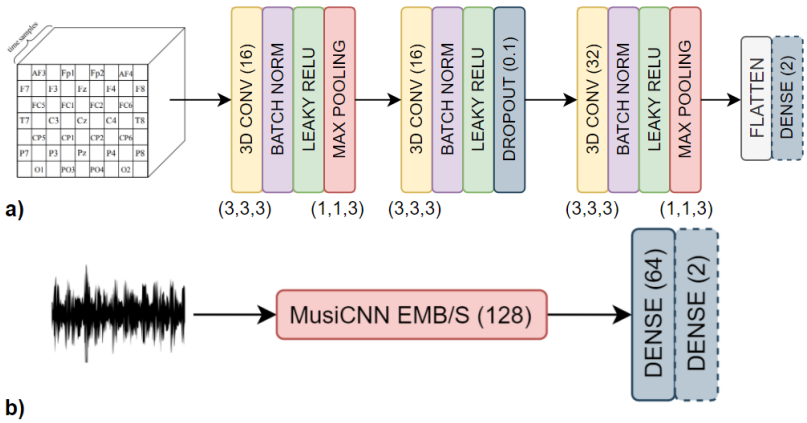

Figure 4.3: Framework Architecture: a) EEG net Architecture b) Music net Architecture.

We will use a bi-stream Neural Network with one stream corresponding to each modality. Both network streams are pre-trained individually on emotion labels and afterwards we fine-tune the whole network by concatenating their final embeddings, and using them as input to a common layer of shared weights. From that point, a linear classifier will be used to predict emotion tags. The EEG branch is a hand-crafted 3D CNN that takes as input an EEG trial in the shape (channels, timesteps) and converts it into the 3D form we

described in Section 4.2.2. The network is lightweight in order to better handle the limited size of the available data and avoid overfitting. It consists of 3 convolutional blocks and a dense layer as shown in Figure 4.4a. The network is pre-trained with supervision on Valence and Arousal separately, as binary classification tasks.

For the music branch we utilize the MusiCNN model [118] to extract high-level embeddings from the available audio stimuli. These are then fed into a simple 2-layer DNN to fine-tune on their respective emotion tags, as shown in Figure 4.4b. The bi-modal network emerges by substituting the 2 dense layers of each of the previous networks with a new common pair of layers (Figure 4.4) – an 64D embedding that constitutes the common latent space for the 2 modalities, followed by a linear classifier to supervise the co-learning process through emotion tags.

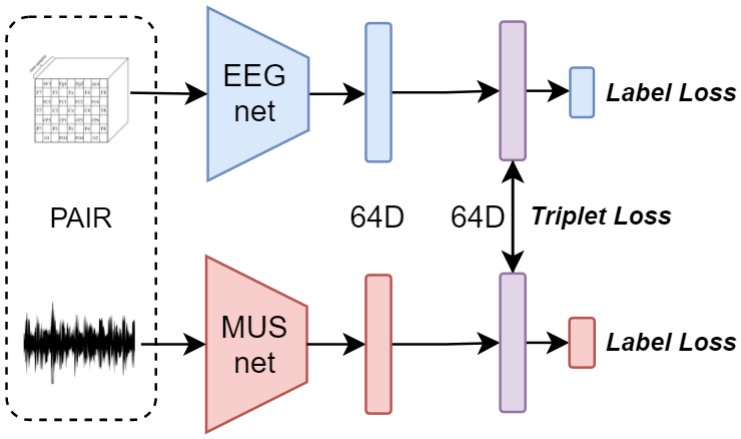

Figure 4.4: The proposed bi-stream network. The 2D dense layers shown in Figure 4.3 are substituted by 64D dense layers and are then connected to the 64D common space.

4.3 The Multimodal Training Procedure

Our goal is to learn a common space where the samples from the same semantic category should be similar, even though they come from different modalities. To learn discriminative features we want to minimize the discrimination loss in both the label and representation space. Simultaneously, we want to reduce the cross-modal discrepancy.

4.3.1 Optimization Methods

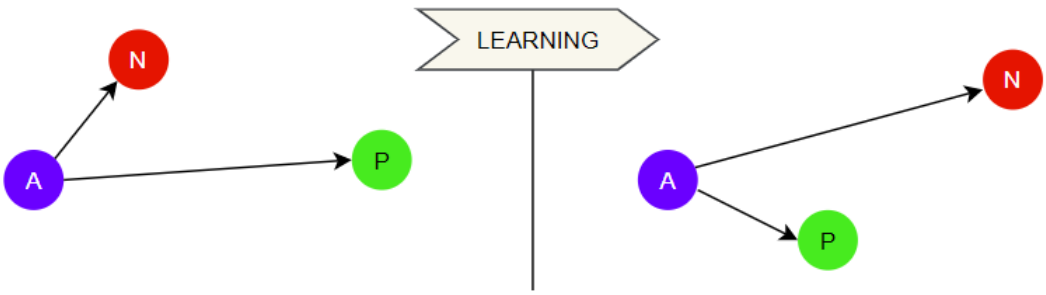

Figure 4.5: The function of triplet losses. Here the arrows correspond to cosine distances. A: anchor, P: positive sample, N: negative sample.

Our framework will incorporate a combination of several loss terms to fulfill the learning requirements of the task. Specifically, we use a linear classifier to predict the emotion labels of the samples projected in the common space. The outputs of each modality are passed through a softmax activation and a binary cross-entropy loss is computed. This loss is applied also at the pre-training sessions of each modality, whereas for the cross-modal task we apply a weighted linear combination of them:

$$\mathcal{J}_1 = \lambda_{11}\mathrm{CE}_a + \lambda_{12}\mathrm{CE}_b \tag{4.1}$$

We measure the metric loss of all samples from both modalities in the common representation space through 2 separate triplet losses. These take as input a triplet of samples: an EEG anchor a, a positive music example p, that is the stimulus sample of a, and a negative EEG-music pair example n that contrasts the emotion label of the anchor. The triplet losses then compare the cosine distances between the three embedding vectors and apply an objective that minimizes the anchor's distance to the positive example, while maximizing its distance to the negative one. The 2 objectives can be denoted as follows:

$$\mathcal{J}_2 = \max\left(\cos(u_a - v_p) - \cos(u_a - u_n), 0\right) \tag{4.2}$$

$$\mathcal{J}_3 = \max\left(\cos(u_a - v_p) - \cos(u_a - v_n), 0\right) \tag{4.3}$$

By combining the above terms we obtain the objective function of our proposed model:

$$\mathcal{J} = \mathcal{J}_1 + \lambda_2\mathcal{J}_2 + \lambda_3\mathcal{J}_3 \tag{4.4}$$

The hyper-parameters λ_i control the contribution of each separate component and are determined through trial and error. Our selected final configuration is:

$$\lambda_{11} = 0.56, \quad \lambda_{12} = 0.24, \quad \lambda_2 = \lambda_3 = 0.1 \tag{4.5}$$

4.3.2 Feature Extraction

For this set of experiments, we will again use DEAP [65] to design subject-dependent training sessions, for each of its 32 subjects. We get the EEG signals in their preprocessed form. The 40 one-minute music stimuli of DEAP are not included in the dataset, whereas the YouTube links provided are in most cases corrupted. We therefore proceeded into locating the video clips of the corresponding tracks and isolating the minute of interest for each one, according to the metadata provided by DEAP. The task of deriving a common space for EEG and music faces a crucial challenge: the semantic gap between the "subjective" affective responses of participants and the "objective" emotion tags of the songs. Ideally, we need musical stimuli that are in accordance with the participants' annotations and independent evaluations as well. The DEAP stimuli have been selected for this purpose and have been separately annotated by the experimenters, as mentioned in Section 3.3. Nearly every song received average ratings from the participants that were in accordance with those annotations. We found that only 8/40 songs had such an inconsistency, from which we only keep 2 of them by altering their "objective" label to match the average annotation of the subjects. The resulting set of tracks is used to extract MusiCNN embeddings. In Table 4.1 the reader can inspect the details for the specific 8 inconsistent tracks, along with the way we chose to process each of them.

ID	Artist	Title	Inconsistency	Operation
8	Lily Allen	**** You	High Arousal Average Rating	Discarded
9	Queen	I Want to Break Free	Low Arousal Label	Changed Label
10	Rage Against The Machine	Bombtrack	High Valence Average Rating	Discarded
11	Michael Franti & Spearhead	Say Hey (I Love You)	Arousal Average Rating nearly 5	Discarded
16	The Submarines	Darkest Things	Valence Average Rating nearly 5	Discarded
21	Diamanda Galas	Gloomy Sunday	High Arousal Label	Changed Label
34	Dj Paul Elstak	A Hardcore State of Mind	Valence Average Rating nearly 5	Discarded
36	Sepultura	Refuse Resist	Valence Average Rating nearly 5	Discarded

Table 4.1: Inconsistent Stimuli of the DEAP Dataset and how we handle them.

EEG & Music Features

EEG and music signals are processed differently in order to end up with an embedding form suitable for multimodal training. DEAP EEG signals are first cut to 1-sec. chunks. We choose this temporal resolution, since it can adequately capture emotion-related characteristics and in order to augment the quantity of our dataset. However, it has been shown that not all seconds of an EEG experiment are important. Qing et al. [120] have studied the temporal variation of the induced emotion on DEAP, suggesting that participants tend to use the first seconds of the experiments for a kind of "emotional calibration". Therefore, from the entire 63-sec. duration we discard the 3-sec. preparation phase along with the first 7 and last 3 seconds of each trial, to avoid periods when emotions are not fully expressed. We end up with 50 signals of 1 sec. for each trial.

Before extracting the input feature vectors, we apply noise augmentation to the EEG chunks, in order to further increase the data quantity. Noise augmentation has been proven useful in assisting the convergence of deep learning models, either by applying additive white noise [162] or by utilizing generative networks [87]. For each chunk, we produce 10 noisy copies by adding white noise with 0.5 variance. For feature extraction purposes we tried out several different baselines, like higher order crossings and instantaneous energy features, ending up selecting the one-sided DFT signal magnitude, using 64 FFT points (see also Section 2.1.2). Each input vector is then given in 2D form (channels, features), since a single EEG segment includes samples from all 32 EEG channels. In Figure 4.6 we depict the extracted features of a sample EEG segment.

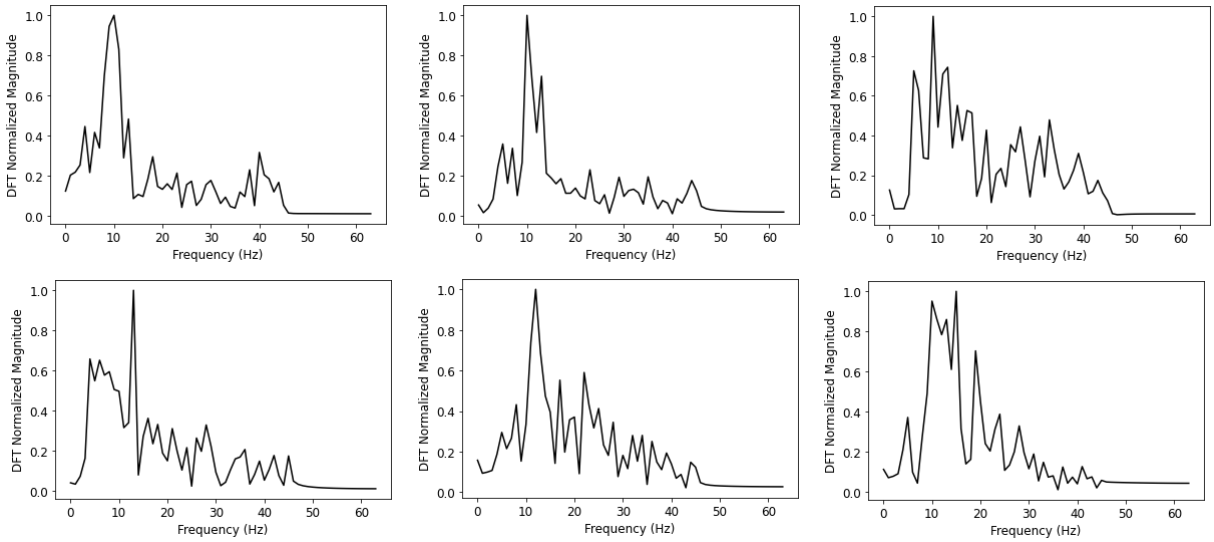

Figure 4.6: Feature vectors of 2 sample EEG trials at different seconds, for channel Fp1. Up: Subject 8 at the 20th, 30th and 40th seconds (left to right). Down: Subject 8 (same time).

Music tracks are trimmed as well into the same 50-sec. window by being cut to 3-sec. chunks with an overlap of 1 sec. This is the input resolution required by the pre-trained MusiCNN model and it also makes an intuitive match to the corresponding EEG. This is because when we listen to music, our affective response does not only correspond to what we hear at that moment, but also takes past stimuli into consideration. Hence, each EEG segment of the interval $[a, (a+1))$ sec. would match to its music stimulus at the interval $[(a-2), (a+1))$ sec, constituting the positive pairs in the networks. On the other hand, negative pairs are mined online, during training: at each epoch, an anchor EEG sample of a certain batch is matched to a random music and a random EEG sample, corresponding to either $\mathcal{J}_2$ or $\mathcal{J}_3$. That sample should be of the opposite label.

4.3.3 Evaluation Metrics

We evaluate our proposed method using accuracy to assess the supervised predictions for each modality and the Precision@10 (P@10) and mean Average Precision (mAP) metrics for the retrieval of music tracks given EEG queries. Those two metrics have been widely used to assess retrieval tasks in the literature [172, 177] as they evaluate the response's distance-based ranking to each query. In particular, P@10 considers the top 10 ranked tracks whereas mAP evaluates the whole ranking. To further clarify the usage of these metrics we provide the following examples in Figure 4.7. Imagine that we are given an EEG query of high valence and we want to rank 10 available music samples based on their cosine distance to the query, from which only 6 are of high valence (dark). To compute P@n we only consider the first n ranked samples and compute their precision as shown in the figure. For $n = 5$ we get $P_{\#1}@5 = 0.8$ and $P_{\#2}@5 = 0.4$. Average Precision is acquired by averaging the precision of the correctly retrieved samples. In that case we would have $AP_{\#1} = (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.68$ and $AP_{\#2} = 0.52$, ending up in the mean Average Precision metric: $mAP = (0.68 + 0.52)/2 = 0.6$.

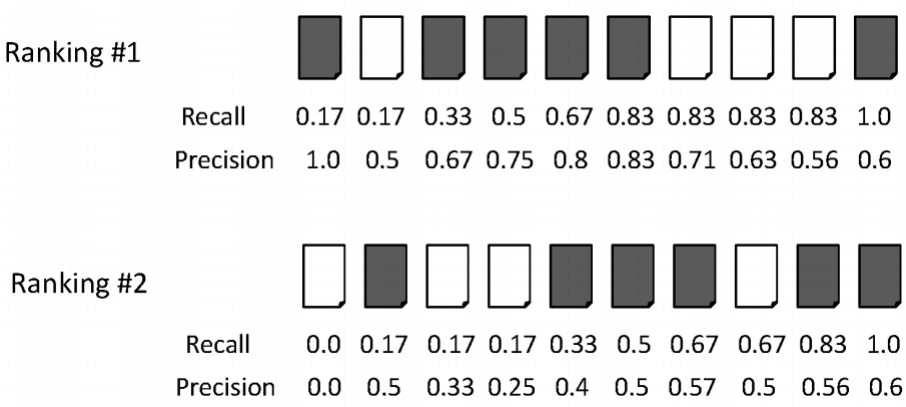

Figure 4.7: Two ranking examples to clarify the usage of the information retrieval metrics P@10 and mAP. Adapted from slides of Rada Mihalcea: web.eecs.umich.edu/mihalcea

Results are also presented after trial aggregation: Each pair of music and EEG trials is composed of 50 segments. For the accuracy we simply denote a prediction as correct if more than half of the segment-wise predictions are correct. For the retrieval metrics, since no such voting can be made, we consider the median of the segment-wise scores. In this way we manage to extract an insightful score for each trial and avoid outliers.

4.4 Experimental Evaluation

The training procedure can be divided into 3 sessions: a) Pre-training of the EEG branch b) Pre-training of the Music branch c) Multimodal fine-tuning. This workflow is applied to the data of each of the 32 available subjects, considering supervision on either valence or arousal. To compensate for possible annotation noise, we binarize the labels by setting the threshold to the median score 5. In order to deal with the issues of limited data size and noisy input samples, we apply 5-fold stratified cross validation, on a per-track basis, on both EEG and music data, where each fold holds 20% of the total data size.

4.4.1 Predicting Emotion Tags

First we present an evaluation of the supervised training sessions on valence and arousal tags (Table 4.2). We observe that, despite the fact that our music set contains only 34 tracks, the emotion recognition performance is substantially high, 74.3% average on valence and 83.3% on arousal, something that indicates the robustness of our transfer learning module and further assists the metric learning process. On the other hand, EEG average scores show high variance per participant, reaching up to 63.3% average on valence and 64.5% on arousal. Generally, aggregating scores per individual track gives a clearer view of the models' interpretation of a track, hence the increased scores.

Dimension	Non-Aggregated	Aggregated
Valence EEG	0.610 → 0.604	**0.633** → 0.632
Arousal EEG	0.645 → 0.641	0.645 → **0.662**
Valence MUS	0.680 → 0.646	**0.743** → 0.689
Arousal MUS	0.838 → 0.837	0.833 → **0.838**

Table 4.2: Emotion Prediction from pre-trained to fine-tuned models - means over 32 subjects.

Moving on to the fine-tuned models and taking trial aggregation into consideration, we observe a decrease in valence scores for both modalities. We will see below that most models trained on valence tend to trade prediction scores, especially those of music tracks, to enhance their performance on the retrieval task, implying poorer correlation between the modalities. On the other hand, co-training on arousal provides supplementary insights for EEG predictions, resulting in nearly 2% improvement, preserving at the same time the efficiency of music embeddings. We now move on to retrieving similar tracks.

4.4.2 Retrieving Tracks from EEG Queries

Table 4.3 summarizes the retrieval scores from the fine-tuned models, acquired by querying the common embedding space of each model with a test EEG sample and then evaluating the ranking of music samples based on their distance to the query. Regarding valence, we have seen that prediction accuracy drops slightly with fine-tuning. On the contrary, retrieval metrics provide more robust results in both cases, indicating that the EEG samples are better situated in the common space and the majority of them are capable of retrieving tracks that are emotionally consistent. In specific, in the case of induced valence, a P@10 value of 65.9% is achieved. We note that this percentage is higher than the supervised prediction accuracy, while the mAP reported over the whole ranking is significantly lower (57.6%), implying that the learned valence space is fragmented into local subspaces of high similarity. We will further explore this observation in Section 4.5.1.

Valence	**Accuracy**	**P@10**	**mAP**
Non-Aggregated	0.610 → 0.604	**0.617**	0.577
Aggregated	0.633 → 0.632	**0.659**	0.576
Arousal	**Accuracy**	**P@10**	**mAP**
Non-Aggregated	0.645 → 0.641	0.653	**0.674**
Aggregated	0.645 → 0.662	0.677	**0.679**

Table 4.3: Retrieval Scores from fine-tuned models - mean values over 32 subjects.

Arousal on the other hand seems to be more consistently represented and this is reflected to the improvement in all available metrics. In specific, both mAP and P@10 median retrieval scores indicate that the majority of tested tracks can derive emotionally consistent music rankings, in contrast to valence where the emotional response similarity seems concentrated to the top-ranked elements. As a result, the correct retrieval percentage conditioned on arousal approaches 68% on average across subjects.

4.4.3 Ablation Studies

In our study we have incorporated a complex objective function (Eq. 4.4) combining 4 terms that aim to minimize the discrimination loss in both the label space and in the common latent space. To investigate the impact of each of those on the performance, we trained separate sessions, each time removing a single term from the objective. The same optimization procedure is followed in all cases and the results are shown in Tables 4.4 and 4.5 for Valence and Arousal respectively, for the aggregated case.

Metric	$\mathcal{J}$	$\neg \mathrm{CE}_a$	$\neg \mathrm{CE}_b$	$\neg \mathcal{J}_2$	$\neg \mathcal{J}_3$
Acc.	0.632	0.562	**0.645**	0.630	0.628
P@10	**0.659**	0.631	0.628	0.625	0.645
mAP	**0.577**	0.527	0.541	0.571	0.573

Table 4.4: Ablation on the utilized Objective Function on Valence (aggregated scores).

Metric	$\mathcal{J}$	$\neg \mathrm{CE}_a$	$\neg \mathrm{CE}_b$	$\neg \mathcal{J}_2$	$\neg \mathcal{J}_3$
Acc.	**0.662**	0.614	0.611	0.661	0.659
P@10	**0.677**	0.505	0.661	0.661	0.632
mAP	**0.679**	0.622	0.564	0.673	0.666

Table 4.5: Ablation on the utilized Objective Function on Arousal (aggregated scores).

From the results we deduce that in general, our full objective $\mathcal{J}$ leads to higher performance on both emotion dimensions, indicating that all the 4 utilized terms contribute to the final scores. We can also see that the supervised losses affect greatly the accuracy score in a contrasting manner in Valence. The absence of supervision on EEG inputs decreases the respective accuracy by 7% while the absence of supervision on music increases it slightly by 1.5%. However, both seem to equally contribute in Arousal and both are important for the retrieval tasks as well. On the other side, metric losses cause slighter modifications to the final representations, since their contribution in $\mathcal{J}$ is rather limited compared to the supervised binary cross-entropy losses. They manage though to bring slight improvements to all metrics, especially in P@10.

Function	Without	Proposed
3D Input Representation	0.614 — 0.646	
White Noise Augmentation	0.600 — 0.620	**0.632 — 0.662**
Pre-training EEG Sessions	0.591 — 0.601	

Table 4.6: Ablation on critical choices in building the EEG model. We depict accuracy scores in their aggregated format and after the fine-tuning Sessions. Form: Valence – Arousal

Finally, we tested the impact of various critical choices in modeling the proposed framework (Table 4.6). To ease the procedure, we concentrate on the more complex EEG model and consider its accuracy scores after fine-tuning. First, the usage of 3D input representations proves to be beneficial, compared to baseline 2D inputs, by 2% on average, something that indicates the efficiency of placing EEG channels in an intuitive topological grid. This form will later enable the extraction of brain map-like activations, as well. Our second experiment considers the usage of noise augmentation applied to EEG samples. While bypassing data augmentation during pre-training leads to lowered accuracy (by 4% on average), we should mention that it does not lead to major changes when discarding it only during fine-tuning. This points to the pre-training sessions being crucial to the convergence of the network – and indeed, we notice an average drop in both accuracy and retrieval scores by more than 5% when omitting them.

4.5 Qualitative Analysis

Each of the 32 included subjects provides different scores so our results are prone to variation. We depict that in the following Figure 4.8, in which the reader can inspect the accuracy scores over the entire set of available subjects, after the fine-tuning. In order to extract qualitative insights from our experiments we consider selected subjects or trials that reveal interesting observations, which we analyze below.

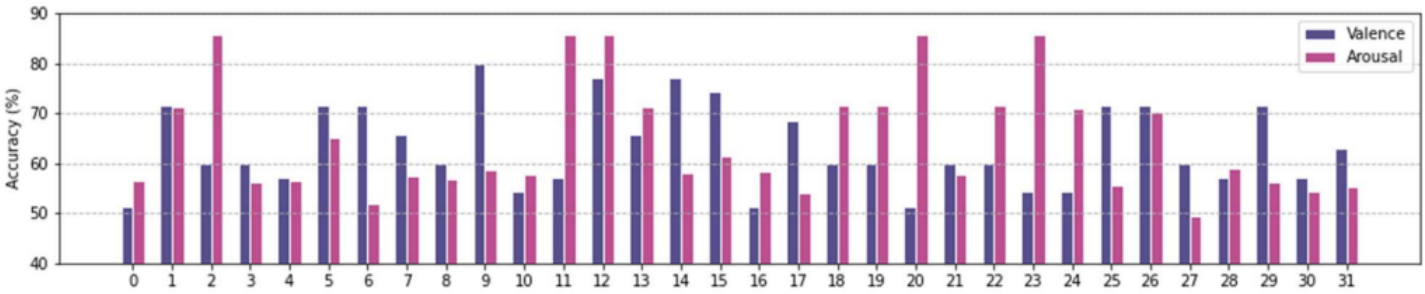

Figure 4.8: Per-Subject Accuracy after the fine-tuning Sessions.

4.5.1 The Common Latent Space

The learning process of the proposed framework aims at constructing a common embedding space where EEG and corresponding music samples could be represented. We visually inspect the produced latent space using the t-SNE [155] to reduce its 64D dimension to 2D. T-SNE is a tool for dimensionality reduction, converting similarities between data points to joint probabilities, so as to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the initial data.

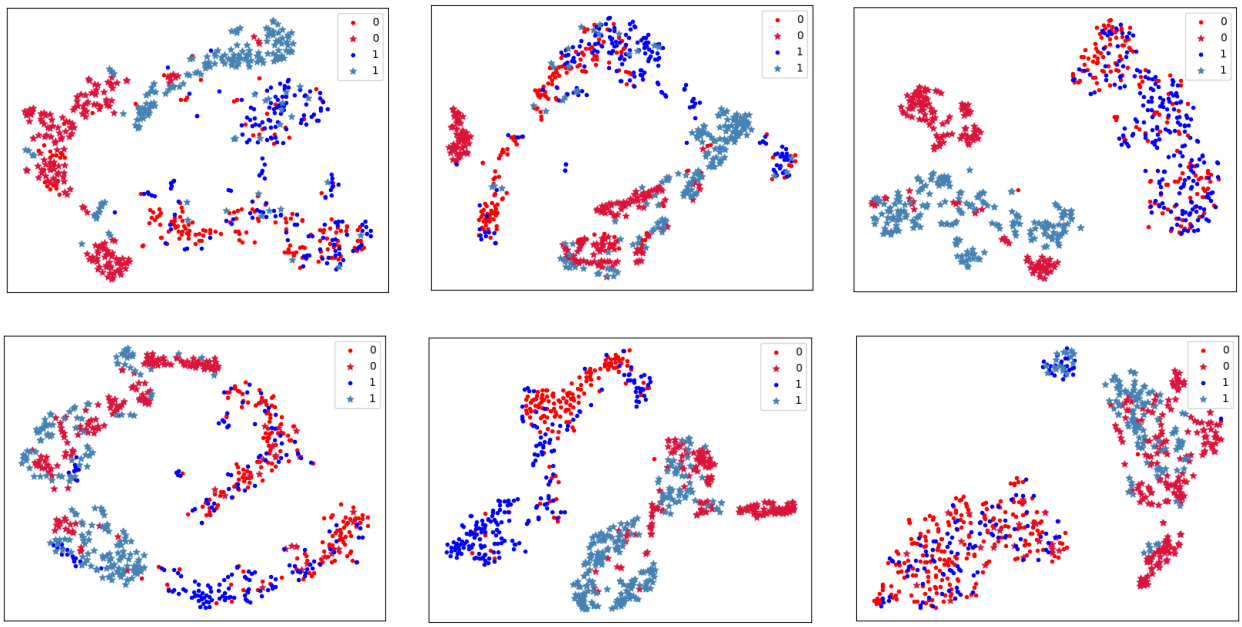

Figure 4.9: Latent Space t-SNE visualisation for the test data of subjects for Valence. Dots denote EEG samples (bright colors) while asterisks denote music samples (dim colors).

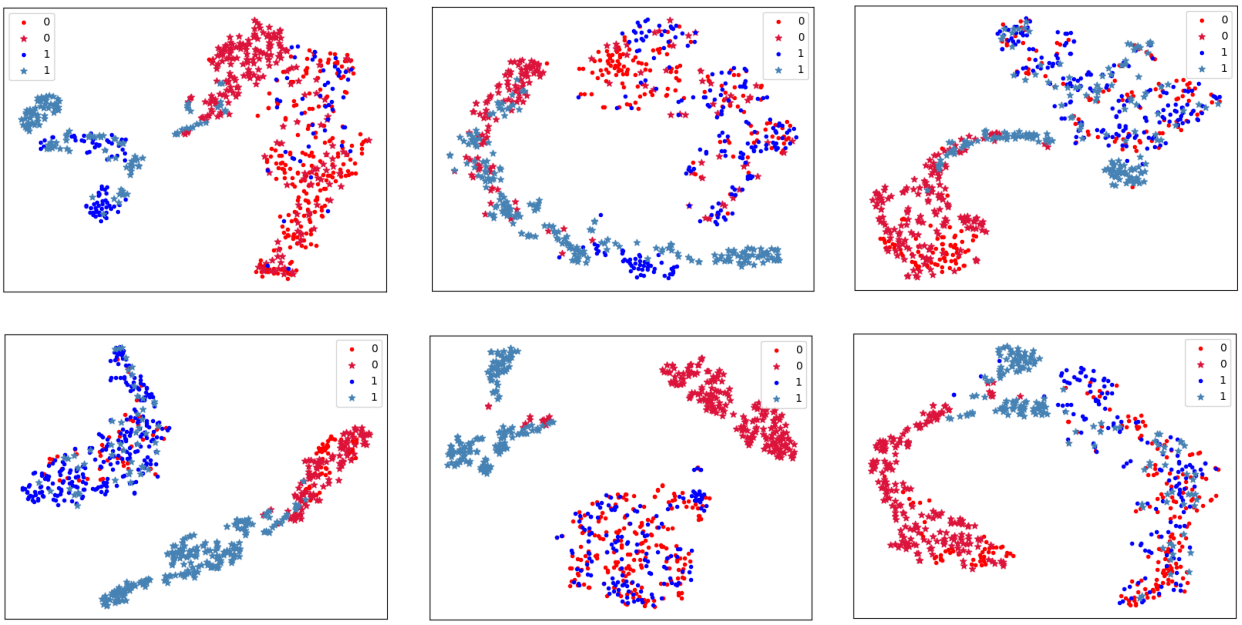

Figure 4.10: Latent Space t-SNE visualisation for the test data of subjects for Arousal. Dots denote EEG samples (bright colors) while asterisks denote music samples (dim colors).

We select one of the 5 trained models for four (4) subjects in Valence and Arousal to display their results in Figure 4.9 and Figure 4.10 (similar trends are observed for most subjects). It is evident that both modalities still form separate groups and cannot homogenize their embeddings. This is especially visible in the case of valence, with cohesive local clusters appearing in the learned common space. This provides an explanation towards the discrepancy we observed between P@10 and mAP metrics, since the top-ranked track retrievals originate from the corresponding local subspace, but there is no coarse bisection between high- and low- valence samples, in contrast to the case of arousal. Nevertheless,

the induced emotion is visually discriminated, especially considering music samples. EEG samples have a rather weak tendency towards the correct music samples, but there are cases as well where the inter-modal sample embeddings are almost indistinguishable.

4.5.2 Temporal Variation of Recognition

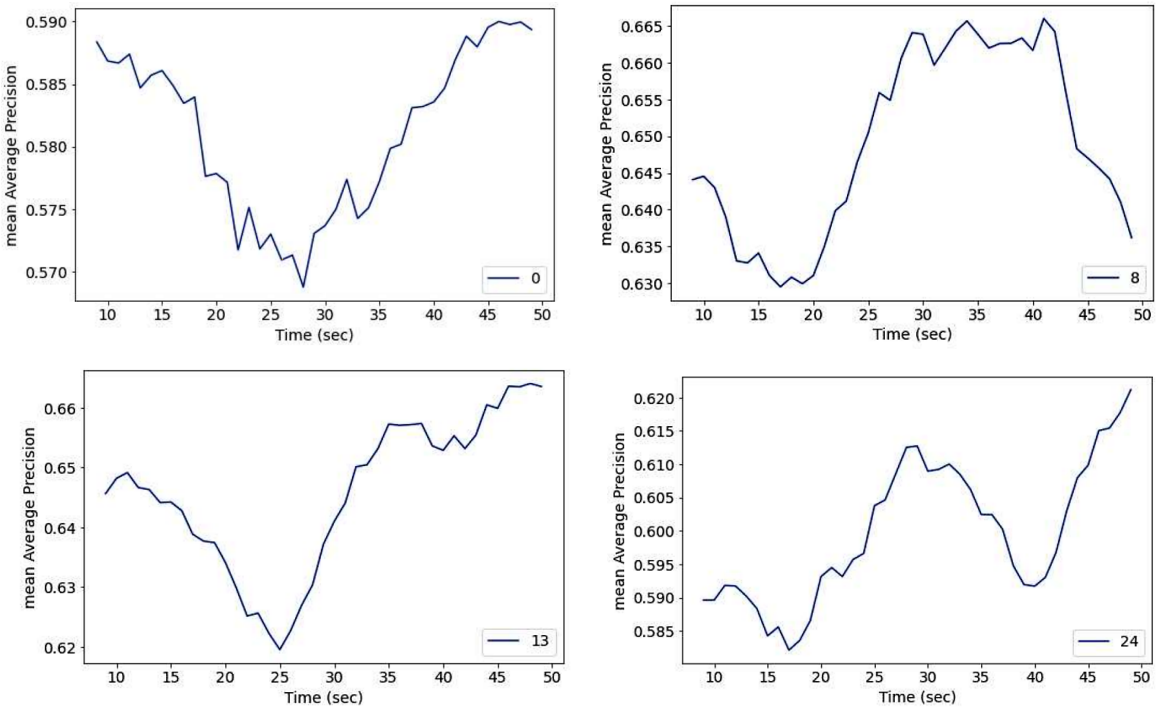

Figure 4.11: Valence mAP scores over the 50 time samples for specific numbered tracks. The scores have been averaged across all 32 subjects, for each time sample separately.

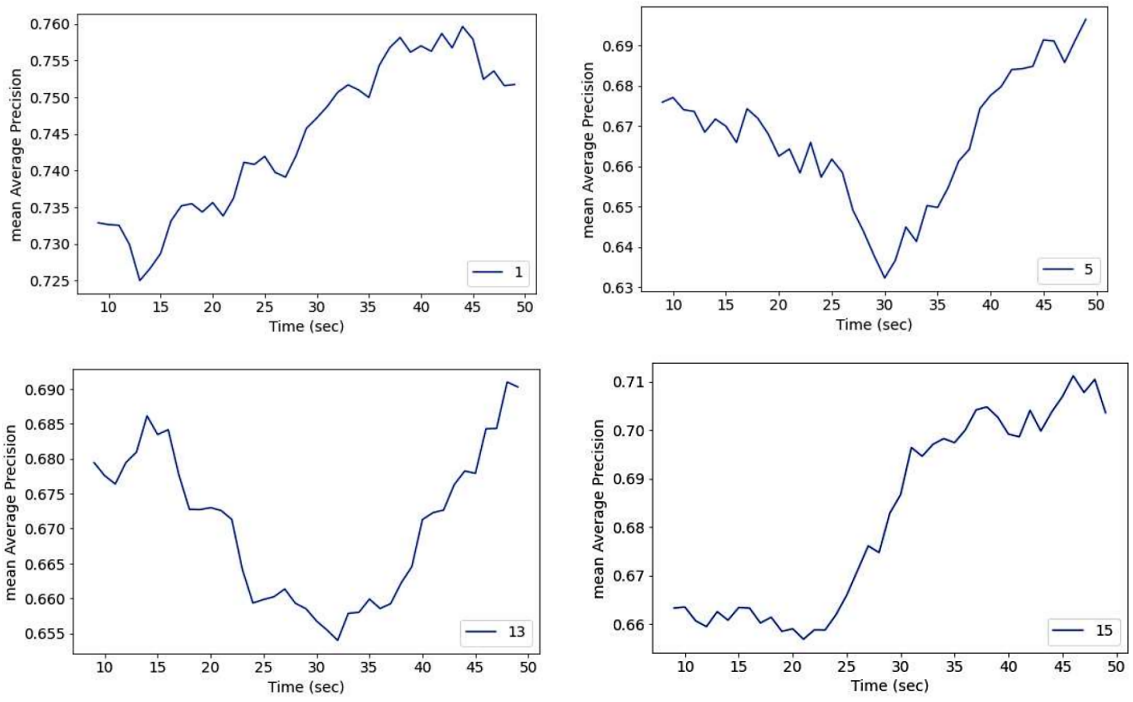

Figure 4.12: Arousal mAP scores over the 50 time samples for specific numbered tracks. The scores have been averaged across all 32 subjects, for each time sample separately.

Since each track is segmented into 50 samples of 1 sec, it is expected that the emotion is not elicited at the same pace throughout the duration of each stimulus. While we perform aggregation, the temporal variation of these scores is of interest and could indicate important moments in the track. In Figures 4.11 and 4.12 we present the temporal evolution of the mAP scores for selected music tracks, averaged across all subjects. While the raw plots are noisy over time, each song individually exhibits a pattern of variation, which we depict by applying a 7-sample moving average filter. Scores typically oscillate on the time axis and emotions are discriminated either at the first moments (10-15 sec) of the stimulus or gradually as it unfolds (45-50 sec), something that has been observed in related studies [120]. Additional experiments are needed though to verify this trend and the use of temporal attention is a plausible research direction for this task.

4.5.3 Scalp Network Activations

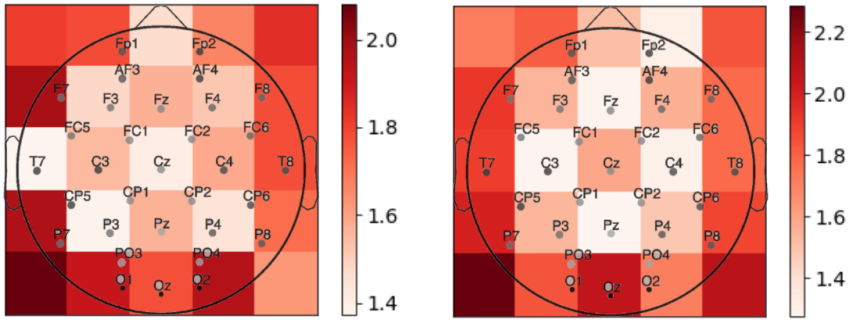

Figure 4.13: Post-ReLU Activation of the first CNN block for subjects 5, 9 in Valence. We present the activation averaged on time and feature axes.

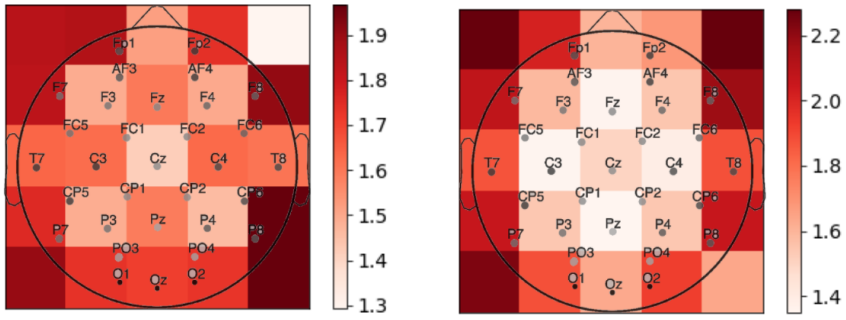

Figure 4.14: Post-ReLU Activation of the first CNN block for subjects 5, 9 in Arousal. We present the activation averaged on time and feature axes.

As displayed in Figure 4.2, the input form of EEG data is a 3D structure that resembles the topological orientation of the 32 available electrodes. Besides its intuitive efficiency during the training sessions, the specific format allows us to derive insights about the dominant brain regions that are involved in emotion elicitation. We evaluate these regions by extracting an intermediate representation of the networks' EEG branch, specifically the activation of the first CNN block that retains a 5×5 topological structure. In Figures 4.13 and 4.14 we depict 4 representative cases. Each plot is derived for a specific track by averaging the activations of its 50 samples. From the plots it is evident that the peripheral channels capture the most discriminating information, which is consistent

with the literature in emotion cognition [178]. We observe that the back-left regions are more important in valence, while front and right regions are mostly activated in arousal. Central regions show low means and low variance as well, hence their role proves limited.

Summary

In this section we presented a novel approach to analyze EEG recordings of music listening and proposed a cross-modal framework to learn common representations for EEG and music data and retrieve consistent music rankings from EEG queries. The proposed approach further indicates that distilling information from processed musical stimuli to the respective EEG signals can lead to interesting insights in personalized emotion analysis. To the best of our knowledge, this is the first study to propose a complete framework to model the specific task and dataset, thus our results can be viewed as a concrete baseline.

Moreover, we provided an extended qualitative evaluation of the framework, that helped us derive interesting evidence regarding critical aspects of music listening, like the actual latent similarity of EEG and music embeddings, the temporal variation of the induced emotions, as depicted by the accuracy scores, and the brain regions that are activated in each case, as simulated by the 3D EEG network. By conducting subject-dependent experiments for 32 different subjects, we reveal important patterns that should be taken into account in future studies.

Chapter 5

Conclusions

This Thesis can be divided into two main sections with respect to the use of machine learning tools, while the major theme of our analysis is the affective perception of music signals through EEG responses. The first part focuses on feature extraction algorithms based on multifractal signal analysis, and the second addresses the problem of identifying emotion-related similarities between EEG and music signals through advanced deep learning techniques. The main contributions of this study are summarized below.

5.1 EEG Affective Features

- We analyzed the structure of EEG signals and demonstrated their multifractal properties. In specific, we investigated the effect of signal's observed stationarity and quantified the signal's complexity through the Hurst Exponent. We derived evidence that EEG signals could be modeled as fractional Gaussian noise realizations.
- We developed 2 novel algorithms, based on Multiscale Fractal Dimension (MFD) and Multifractal Detrended Fluctuation Analysis (MFDFA) to derive meaningful feature vectors for emotion detection. In the first case we performed short-time analysis of rather long EEG sequences and built feature vectors using statistical measures of the acquired fractograms. In the second case, we utilized the signals' multifractal spectrum to recognize emotions.
- We tested the proposed methods in classical Machine Learning settings and an SVM classifier, against widely used baseline frequency and fractal features. We also performed both intra-subject and inter-subject experiments to identify the features' generalizability. We showed that the proposed feature sets perform strongly against the baselines, particularly in the subject-independent setting and in arousal recognition, indicating that arousal is correlated with the structure of the EEG.
- Fractal and multifractal features seem to generalize more easily than frequency-related ones, which perform better in subject-dependent settings. Further improvements are achieved when the fractal features are aggregated. As a result, the observed multifractality should be considered when processing EEG signals.
- We performed an extensive ablation study regarding the EEG channels' position and the frequency rhythms that are prone to elicit emotional information. For the DEAP Dataset, we identified stronger affective characteristics on the left hemisphere, while the discriminability of the alpha frequency band is underlined in all experiments.

5.2 Cross-Modal Learning

- We presented a robust 3D deep network to efficiently analyze EEG signals or EEG features by preserving their temporal and spatial correlation. That is, we provided sequential input features and arrange the EEG channels in a topological grid. We additionally provided ways of dealing with core problems associated with this kind of data, such as the limited sample size and their noisy structure.

- We proposed a multimodal framework to model the correspondence between human brain responses and music stimuli. We trained a bi-stream network on pairs of EEG and corresponding music stimuli, whereas by conditioning the learning process with emotion tags we constructed a common emotion space. To the best of our knowledge, this is the first study to propose such a framework, seen as a baseline reference.

- Through the produced latent space by the aforementioned network, we performed emotion recognition both by predicting output annotations and by ranking music tracks to EEG input queries, based on their cosine distance on the space. We tried out various forms of prediction aggregation and performed extensive ablation studies on our choices in building and evaluating this framework.

- We performed a qualitative study across all 32 subjects of DEAP [65], by formulating personalized models with data from a single participant in training and testing sessions. This way we could compare 32 model instances and observed significant patterns, such as the visualized latent spaces, the temporal variation of recognition performance and activation patterns on the simulated scalp grid of the EEG network.

5.3 Suggestions for Future Work

Our work contributes to the research community in the above-mentioned fields, however it also paves the way for further investigation of cognitive aspects of music listening. Regarding the first section, further work could consider feature extraction algorithms for determining asymmetrical multifractal properties, whereas an interesting direction would be the examination and comparison of multi-band energy EEG features, obtained through energy separation algorithms. The Teager-Kaiser operator [92, 61] has provided intuitive results in fields like speech analysis [91] and one of our next research efforts will be the analysis of its application to EEG signals for recognizing affect. The Energy Separation Algorithm (ESA), proposed in [91], could be used to analyze the contribution of each EEG band, in comparison with other empirical decomposition methods.

Regarding the second section, our field of study is relatively underexplored and, to the best of our knowledge, this is the first study to propose a complete framework to model the specific task and dataset, thus our results can be viewed as a concrete baseline. In particular, future work should incorporate more sophisticated feature extraction methods to reach state-of-the-art performance and optimally correlate with music stimuli. An underexplored dimension in this study concerns the temporal dependencies of the induced emotions. As we saw, in certain time intervals the model performance is relatively higher, something that could be further researched through a recurrent or attention module. The combination of fractal feature input vectors is a straightforward proposal. Another interesting direction would be to enhance the common representation space with a music database and examine its impact on the retrieved rankings.

Appendix

DEAP Dataset

The *Database for Emotion Analysis using Physiological Signals* (DEAP) is a multimodal dataset for the analysis of human affective states. The electroencephalogram (EEG) and peripheral physiological signals of 32 participants were recorded as each watched 40 one-minute long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence and other dimensions. The dataset has been made publicly available by its curators and can be obtained upon request. Music video clips are used as the audiovisual stimuli to elicit different emotions. To this end, a relatively large set of music video clips was gathered using a novel stimuli selection method. A subjective test was then performed to select the most appropriate test material. For each video, a one-minute highlight was selected automatically. 32 participants took part in the experiment and their EEG and peripheral physiological signals were recorded as they watched the 40 selected music videos. Participants rated each video in terms of arousal, valence, like/dislike, dominance and familiarity. The database contains all recorded signal data, frontal face videos for a subset of the participants and subjective ratings from the participants. Also included are independent track annotations, made by the authors.

Stimuli Selection

Eliciting emotional reactions from test participants is a difficult task and selecting the most effective stimulus material is crucial. The stimuli used in the experiment were selected in several steps. First, 120 initial stimuli were selected, half of them using the Last.fm music website. Last.fm allows users to track their music listening habits, receive recommendations for new music and assign tags to individual songs. Last.fm offers an API, allowing one to retrieve tags and tagged songs. A list of 304 emotional keywords was yielded from psychological studies. Next, for each keyword, corresponding tags were found in the Last.fm database. For each affective tag, the 10 songs most often labeled with this tag were selected. This resulted in a total of 1084 songs and, in order to ensure balance, 15 were selected manually for each quadrant in the Valence-Arousal space, according to various criteria (tag accuracy, video availability etc). The rest of the tracks were selected manually, again 15 for each of the quadrants. The goal here was to select those videos expected to induce clear emotional reactions.

For each of the 120 initially selected music videos, a one-minute segment was extracted for use, so that it contains emotionally stimulating content. To this end, a highlighting algorithm was used, originally proposed by Soleymani et al. [143], performing linear regression on the content-based features. The music videos were then segmented into one minute segments with 55 seconds overlap. Content features were extracted and provided the input for the regressors. For each video, the segment with the highest emotional score,

aggregated on valence and arousal, was chosen for the experiment. For few well-known tracks only, the selection was manual.

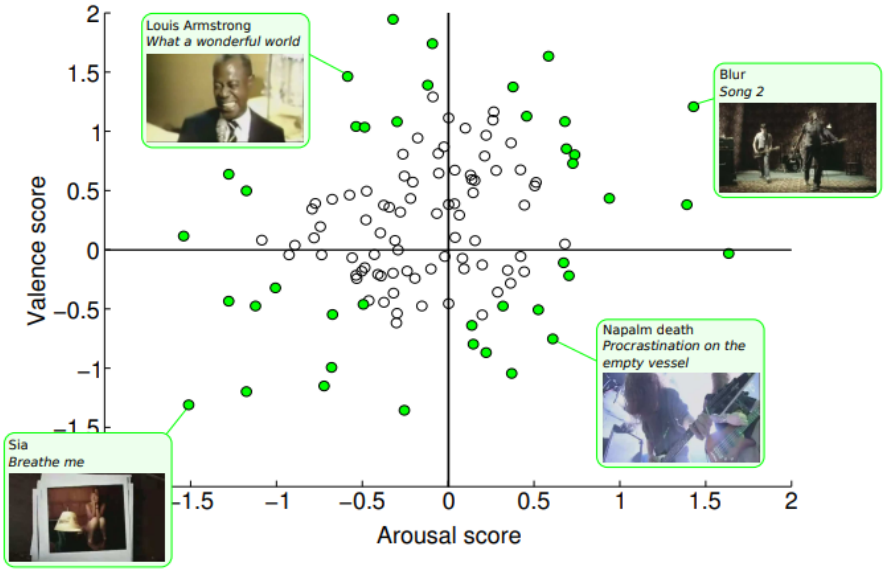

Figure 5.1: [65] Online Assessment Ratings. Selected 40 Videos are highlighted in green.

Given the 120 one-minute music video segments, the final 40 test video clips were chosen using a web-based subjective emotion assessment interface. Participants watched music videos and rated them on a discrete 9-point scale for valence, arousal and dominance. The order of the clips was randomized, but preference was given to the clips rated by the least number of participants, which ensured a similar number of ratings for each video (14- 16 assessments per video). To maximize the strength of elicited emotions, the authors selected as final those videos that had the strongest volunteer ratings with small variation. Figure 5.1 shows the score for the ratings of each video and the selected videos highlighted in green. The video whose rating was closest to the extreme corner of each quadrant is mentioned explicitly. Of the 40 videos, 17 were selected via Last.fm tags.

EEG Recording Experiment

The experiments were performed in two laboratory environments with controlled illumination. EEG and peripheral physiological signals were recorded using a Biosemi ActiveTwo system on a PC. Stimuli were presented using a dedicated stimulus PC , while for presentation of the stimuli and recording the users' ratings, the "Presentation" software by Neurobehavioral systems was used. The music videos were presented on a 17-inch screen, at 800 × 600 resolution. Stereo Philips speakers were used and the music volume was set at a relatively loud level. EEG were recorded at a sampling rate of 512 Hz using 32 active AgCl electrodes, according to the international 10-20 system. 13 peripheral physiological signals were also recorded. Additionally, for the first 22 participants, frontal face video was recorded using a Sony DCR-HC27E consumer-grade cam recorder.

32 Healthy participants (50% female), aged between 19 and 37, participated in the experiment. Prior to the experiment, each participant signed a consent form and filled out a questionnaire. Next, they were given a set of instructions regarding the experiment protocol. Afterwards, the sensors were placed and their signals checked, the participants

performed a practice trial to familiarize themselves with the system. Next, the experimenter started the physiological signals recording and left the room, after which the participant started the experiment by pressing a keyboard key. The experiment started with a 2 minute baseline recording, during which a fixation cross was displayed to the participant. Then the 40 videos were presented in 40 trials, with a break at 20 completed, each consisting of the following steps:

1. 2-sec display of the current trial number and progress
2. 5-sec baseline recording (fixation cross)
3. 1-min music video stimulus display
4. Self-assessment on arousal, valence, liking and dominance

Self-assessment manikins (SAM) were used, with the numbers 1-9 printed below. For the liking scale, thumbs down and thumbs up symbols were used. Participants just clicked the respective manikin to indicate their induced emotion. The valence scale ranged from sad to joyful, while the arousal scale from calm to excited. The dominance scale ranged from submissive to dominant. The fourth scale asked for participants' personal liking of the video and should not be confused with the valence scale. Finally, participants were asked to rate their familiarity with each of the songs on a scale of 1 ("Never heard it before") to 5 ("Knew the song very well").

As a side note, since we do not incorporate them in our study, the peripheral physiological signals that were recorded are: GSR, respiration amplitude, skin temperature, electrocardiogram, blood volume by plethysmograph, electromyograms of Zygomaticus and Trapezius muscles, and electrooculogram (EOG). GSR provides a measure of the resistance of the skin by positioning two electrodes on the distal phalanges of the middle and index fingers. This resistance decreases due to an increase of perspiration, which usually occurs when one is experiencing emotions such as stress or surprise. Skin temperature and respiration were recorded since they vary with different emotional states. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear. Regarding the EMG signals, the Trapezius muscle (neck) activity was recorded to investigate possible head movements during music listening.

Subjective Ratings Analysis

Stimuli were selected to induce emotions in the four quadrants of the valence-arousal space (LALV, HALV, LAHV, HAHV). The stimuli from these 4 conditions generally resulted in the elicitation of the target emotion aimed for when the stimuli were selected, ensuring that large parts of the arousal-valence plane (AV plane) are covered (see Figure 5.2). The emotion elicitation worked specifically well for the high arousing conditions, yielding relatively extreme valence ratings for the respective stimuli. The stimuli in the low arousing conditions were less successful in the elicitation of strong valence responses, something that is however commonly observed in related studies [49]. The distribution of the individual ratings per conditions shows a large variance within conditions, possibly associated with stimulus characteristics or inter-individual differences in music taste, general mood, or scale interpretation.

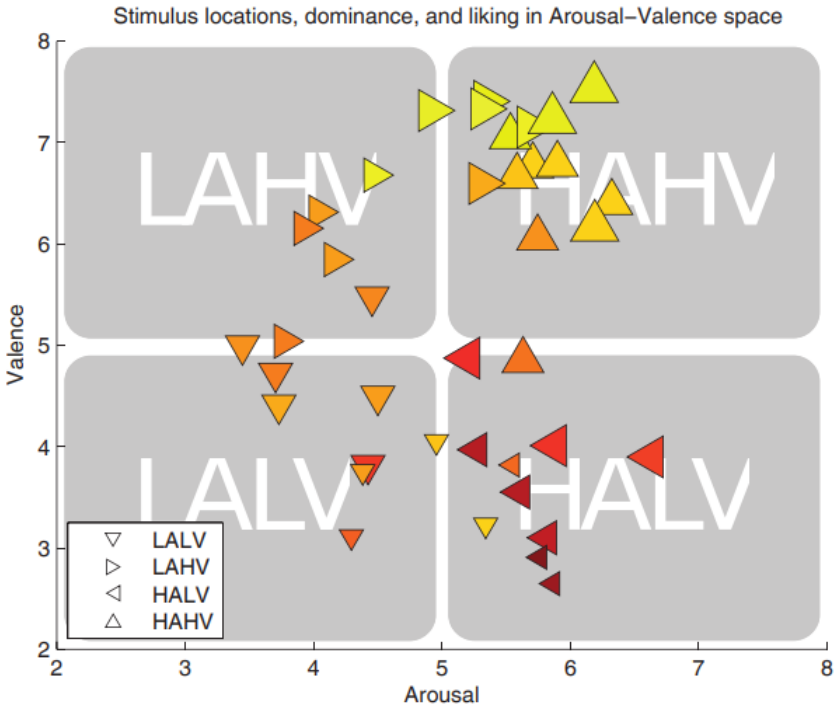

Figure 5.2: [65] The mean locations of the stimuli on the arousal/valence plane. Liking is encoded by color: dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: small symbols stand for low dominance and big for high dominance.

The distribution of ratings for the different scales and conditions suggests a complex relationship between ratings. By exploring the scale inter-correlation over participants (see Table 3), the authors observed high positive correlations between liking and valence, and between dominance and valence. Seemingly, people liked music which gave them a positive feeling and/or a feeling of empowerment. Medium positive correlations were observed between arousal and dominance, and between arousal and liking. Familiarity correlated moderately positive with liking and valence. As already observed above, the scales of valence and arousal are not independent, but their positive correlation is rather low, suggesting that participants were able to differentiate between these two important concepts. In summary, the affect elicitation was in general successful, though the low valence conditions were partially biased by moderate valence responses and higher arousal.

Distribution

The described dataset is publicly available for research usage upon request. The recorded data is given in both their original format and in a pre-processed form which is the one we select for our study. In particular, the data was first downsampled from 512Hz to 128Hz. EOG artefacts were removed with a blind source separation technique and a bandpass frequency filter from 4-45Hz was applied, as to isolate theta, alpha, beta and gamma rhythms. The data was then averaged to the common reference and segmented into 60 second trials and a 3 second pre-trial baseline. For detailed information regarding the dataset, readers can access the respective paper [65] and the official description page: https://eecs.qmul.ac.uk/mmv/datasets/deap/readme.html.

List of Publications

The research presented in this Thesis has resulted in 2 article Publications:

- K. Avramidis, A. Zlatintsi, C. Garoufis, and P. Maragos, "Multiscale Fractal Analysis on EEG Signals for Music-Induced Emotion Recognition", in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO-2021)*, August 2021.

  Emotion Recognition from EEG signals has long been researched as it can assist numerous medical and rehabilitative applications. However, their complex and noisy structure has proven to be a serious barrier for traditional modeling methods. In this paper, we employ multifractal analysis to examine the behavior of EEG signals in terms of presence of fluctuations and the degree of fragmentation along their major frequency bands, for the task of emotion recognition. In order to extract emotion-related features, we utilize two novel algorithms for EEG analysis, based on Multiscale Fractal Dimension and Multifractal Detrended Fluctuation Analysis. The proposed feature extraction methods perform efficiently, surpassing some widely used baseline features on the competitive DEAP dataset, indicating that multifractal analysis could serve as basis for the development of robust models for affective state recognition.

- K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Predict or Retrieve: Insights from Cross-Modal Learning between EEG and Music Stimuli", under review for the *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*, November 2021.

  Cross-modal Learning aims to extract the semantic correspondence between different types of data and project it onto a common space so that an input from one modality can retrieve information about the other. In this paper, we focus on modeling the relationship between pairs of music tracks and corresponding EEG recordings. Brain signals inherit a highly complex and chaotic structure that makes it difficult to process and thus retrieve meaningful information. In order to alleviate these disadvantages and identify similarities between them and their music stimuli, we utilize a bi-modal metric learning framework. Specifically, we combine a 3D convolution model to process EEG signals with a pre-trained network for music tagging in order to create a common latent space. We then align the embeddings of these networks using metric learning, further constraining the whole process using emotion labels. The resulting framework can be utilized for emotion recognition both directly, by performing supervised predictions, and indirectly, by providing relevant music samples from EEG input queries. By applying this system independently to all 32 subjects of the DEAP Dataset we also extract common patterns for the brain regions that interpret music stimuli and the temporal variance of the induced emotions.

There have been also published two articles out of the thesis' scope:

- A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music", in *Proceedings of the 28th European Signal Processing Conference (EUSIPCO-2020)*, August 2020.

  Instrument classification is one of the fields in Music Information Retrieval (MIR) that has attracted a lot of research interest. However, the majority of that is dealing with monophonic music, while efforts on polyphonic material mainly focus on predominant instrument recognition. In this paper, we propose an approach for instrument classification in polyphonic music from purely monophonic data, that involves performing data augmentation by mixing different audio segments. A variety of data augmentation techniques focusing on different sonic aspects, such as overlaying audio segments of the same genre, as well as pitch and tempo-based synchronization, are explored. We utilize Convolutional Neural Networks for the classification task, comparing shallow to deep network architectures. We further investigate the usage of a combination of the above classifiers, each trained on a single augmented dataset. An ensemble of VGG-like classifiers, trained on non-augmented, pitch-synchronized, tempo-synchronized and genre-similar excerpts, respectively, yields the best results, achieving slightly above 80% in terms of label ranking average precision (LRAP) in the IRMAS test set.ruments in over 2300 testing tracks.

- K. Avramidis, A. Kratimenos, C. Garoufis, A. Zlatintsi, and P. Maragos, "Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms", in *Proceedings of the 46th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, June 2021.

  Sound Event Detection and Audio Classification tasks are traditionally addressed through time-frequency representations of audio signals such as spectrograms. However, the emergence of deep neural networks as efficient feature extractors has enabled the direct use of audio signals for classification purposes. In this paper, we attempt to recognize musical instruments in polyphonic audio by only feeding their raw waveforms into deep learning models. Various recurrent and convolutional architectures incorporating residual connections are examined and parameterized in order to build end-to-end classi-fiers with low computational cost and only minimal preprocessing. We obtain competitive classification scores and useful instrument-wise insight through the IRMAS test set, utilizing a parallel CNN-BiGRU model with multiple residual connections, while maintaining a significantly reduced number of trainable parameters.

Bibliography

[1] M. Abdul-Mageed and L. Ungar, "EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks," in *Proc. of the Association for Computational Linguistics (ACL)*, 2017.

[2] F. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, 07 2020.

[3] S. M. Alarcão and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.

[4] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *Proceedings of Machine Learning Research*, S. Dasgupta and D. McAllester, Eds., vol. 28. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 1247–1255.

[5] H. L. Attwood and W. A. MacKay, *Essentials of Neurophysiology.* Mosby Inc., 1989.

[6] K. Avramidis, A. Zlatintsi, C. Garoufis, and P. Maragos, "Multiscale Fractal Analysis on EEG Signals for Music-Induced Emotion Recognition," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO-2021)*, Aug. 2021.

[7] M. M. H. E. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, pp. 572–587, 2011.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *CoRR*, vol. abs/1409.0473, 2015.

[9] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2017, pp. 1–5.

[10] F. Bao, X. Liu, and C. Zhang, "PyEEG: An open source python module for EEG/MEG feature extraction," *Computational Intelligence and Neuroscience*, March 2011.

[11] L. Barrett, R. Adolphs, S. Marsella, A. Martinez, and S. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychological Science in the Public Interest*, pp. 1–68, 2019.

[12] D. Bashwiner, "Brain and Music. By Stefan Koelsch," *Music Theory Spectrum*, vol. 39, no. 2, pp. 269–274, 08 2017.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417.

[14] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 747–754.

[15] C. J. Beedie, P. C. Terry, A. M. Lane, and T. J. Devonport, "Differential assessment of emotions and moods: Development and validation of the Emotion and Mood Components of Anxiety Questionnaire," *Personality and Individual Differences*, vol. 50, no. 2, pp. 228–233, 2011.

[16] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proceedings of the 12th International Society on Music Information Retrieval Conference (ISMIR)*, 2011.

[17] M. Bigerelle and A. Iost, "Fractal dimension and classification of music," *Chaos, Solitons & Fractals*, vol. 11, no. 14, pp. 2179–2192, 2000.

[18] F. Bre, J. Gimenez, and V. Fachinotti, "Prediction of Wind Pressure Coefficients on Building Surfaces using Artificial Neural Networks," *Energy and Buildings*, vol. 158, 11 2017.

[19] C. Busso, M. Bulut, C.-C. Lee, E. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[20] W. B. Cannon, "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory," *The American Journal of Psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[21] R. Caton, "The Electric Currents of the Brain," *American Journal of EEG Technology*, vol. 10, no. 1, pp. 12–14, 1970.

[22] D. Chafale and A. Pimpalkar, "Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic," *International Journal of Computer Sciences and Engineering (IJCSE)*, vol. 2, pp. 14–18, 01 2014.

[23] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1552–1556.

[24] S. Chen, Z. Gao, and S. Wang, "Emotion recognition from peripheral physiological signals enhanced by EEG," in *Proc. ICASSP*, 2016, pp. 2827–2831.

[25] Y. Chen, I. Hung, M. Huang, C. Hou, and K. Cheng, "Physiological signal analysis for patients with depression," in *Proc. Int'l Conf. on Biomedical Engineering and Informatics*, vol. 2, 2011, pp. 805–808.

[26] K. Cho, B. V. Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," *ArXiv*, vol. abs/1409.1259, 2014.

[27] K. Chwalisz, E. Diener, and D. Gallagher, "Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured." *Journal of Personality and Social Psychology*, p. 820–828, 1988.

[28] Y. Cimtay and E. Ekmekcioglu, "Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition," *Sensors*, vol. 20, no. 7, 2020.

[29] J. Cohn, A. Zlochower, J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 396–401.

[30] J. Cooley and J. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, 1965.

[31] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[32] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, 1995.

[33] A. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, and G. Prasad, "Sixteen facial expressions occur in similar contexts worldwide," *Nature*, pp. 251–257, 2021.

[34] J. Davis, A. Senghas, and K. Ochsner, "How Does Facial Feedback Modulate Emotional Experience?" *Journal of research in personality*, vol. 43, pp. 822–829, 10 2009.

[35] D. Delignières, S. Ramdani, L. Lemoine, K. Torre, M. Fortes, and G. Ninot, "Fractal analyses for 'short' time series: A re-assessment of classical methods," *Journal of Mathematical Psychology*, vol. 50, pp. 525–544, 2006.

[36] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-Task Learning of Facial Landmarks and Expression," in *Proceedings of the 2014 Canadian Conference on Computer and Robot Vision*, ser. CRV '14. USA: IEEE Computer Society, 2014, p. 98–103.

[37] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 427–431, 1979.

[38] R. Duan, J. Zhu, and B. Lu, "Differential Entropy Feature for EEG-based Emotion Classification," in *Proc. Int'l IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 81–84.

[39] A. Eke, P. Herman, L. Kocsis, and L. R. Kozak, "Fractal characterization of complexity in temporal physiological signals," *Physiological Measurement*, vol. 23, no. 1, pp. R1–R38, jan 2002.

[40] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17 2, pp. 124–9, 1971.

[41] K. Falconer, *Hausdorff Measure and Dimension.* John Wiley & Sons, Ltd, 2003, ch. 2, pp. 27–38.

[42] T. Fujioka, B. Ross, and L. J. Trainor, "Beta-Band Oscillations Represent Auditory Beat and Its Metrical Hierarchy in Perception and Imagery," *Journal of Neuroscience*, vol. 35, no. 45, pp. 15 187–15 198, 2015.

[43] G. Gálvez-Coyt, A. Muñoz-Diosdado, J. Peralta, J. Balderas-López, and F. Angulo-Brown, "Parameters of Higuchi's method to characterize primary waves in some seismograms from the Mexican subduction zone," *Acta Geophysica*, vol. 60, no. 3, pp. 910–927, Jun. 2012.

[44] D. Ghosal and M. Kolekar, "Music Genre Recognition Using Deep Neural Networks and Transfer Learning," in *INTERSPEECH*, 2018.

[45] I. Good and B. Mandelbrot, "Fractals: Form, Chance, and Dimension." *Journal of the American Statistical Association*, vol. 73, p. 438, 1978.

[46] T. Greer, B. Ma, M. Sachs, A. Habibi, and S. Narayanan, "A Multimodal View into Music's Effect on Human Neural, Physiological, and Emotional Experience," in *Proc. of the 27th ACM International Conference on Multimedia.* New York, NY, USA: Association for Computing Machinery, 2019, p. 167–175.

[47] T. Greer, K. Singla, B. Ma, and S. Narayanan, "Learning Shared Vector Representations of Lyrics and Chords in Music," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Spech and Signal Processing (ICASSP)*, 05 2019, pp. 3951–3955.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[49] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[50] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D*, vol. 31(2), no. 2, pp. 277–83, 1988.

[51] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997.

[52] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[53] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics." in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[54] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 01 2006, vol. 1.

[55] E. Ihlen, "Introduction to Multifractal Detrended Fluctuation Analysis in Matlab," *Frontiers in Physiology*, vol. 3, 2012.

[56] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[57] M. R. Islam and M. Ahmad, "Wavelet Analysis Based Classification of Emotion from EEG Signal," in *Proc. Int'l Conf. on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–6.

[58] H. Jasper, "Report of the committee on methods of clinical examination in electroencephalography," *Electroencephalography and Clinical Neurophysiology*, vol. 10, pp. 370–375, 1958.

[59] C. C. Jouny and G. K. Bergey, "Characterization of early partial seizure onset: Frequency, complexity and entropy," *Clinical Neurophysiology*, vol. 123, no. 4, pp. 658–669, 2012.

[60] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2983–2991, 2015.

[61] J. Kaiser, "Some useful properties of Teager's energy operators," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 1993, pp. 149–152 vol.3.

[62] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, "Multifractal Detrended Fluctuation Analysis of Nonstationary Time Series," *Physica A: Statistical Mechanics and its Applications*, vol. 316, no. 1, 2002.

[63] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2015.

[64] S. Koelsch, *Brain and music*. Wiley Blackwell, 2012.

[65] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, 2011.

[66] A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 156–160.

[67] J. S. Kumar and P. Bhuvaneswari, "Analysis of Electroencephalography (EEG) Signals and Its Categorization-A Study," *Procedia Engineering*, vol. 38, pp. 2525 – 2536, 2012.

[68] S. Lalitha, D. Geyasruti, R. Narayanan, and S. M, "Emotion Detection Using MFCC and Cepstrum Features," *Procedia Computer Science*, vol. 70, pp. 29–35, 2015, proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems.

[69] P. Lang, "The varieties of emotional experience: a meditation on James-Lange theory," *Psychological review*, vol. 101, no. 2, p. 211—221, April 1994.

[70] C. G. Lange and W. James, *The Emotions*. Williams & Wilkins Co., 1922.

[71] C. S. Lea, R. Vidal, A. Reiter, and G. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," *ArXiv*, vol. abs/1608.08242, 2016.

[72] C. Lee, S. S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, pp. 240–243, 2001.

[73] D. J. Levitin, *This is your brain on music: The science of a human obsession.* Dutton/Penguin Books, 2006.

[74] B. Li and A. Kumar, "Query by Video: Cross-modal Music Retrieval," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, Delft, The Netherlands*, 2019.

[75] M. Li, Y. Li, S.-L. Huang, and L. Zhang, "Semantically Supervised Maximal Correlation For Cross-Modal Retrieval," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 2291–2295.

[76] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[77] S. Li, W. Zhou, Q. Yuan, S. Geng, and D. Cai, "Feature extraction and recognition of ictal EEG using EMD and SVM," *Computers in biology and medicine*, vol. 43, no. 7, pp. 807–816, 2013.

[78] X. Li, S. Leglaive, L. Girin, and R. Horaud, "Audio-Noise Power Spectral Density Estimation Using Long Short-Term Memory," *IEEE Signal Processing Letters*, vol. 26, pp. 918–922, 2019.

[79] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A Novel Bi-hemispheric Discrepancy Model for EEG Emotion Recognition," *IEEE Trans. on Cogn. and Developmental Systems*, 2020.

[80] J. Liu, H. Meng, M. Li, F. Zhang, R. Qin, and A. Nandi, "Emotion detection from EEG recordings based on supervised and unsupervised dimension reduction," *Concurrency and Computation: Practice and Experience*, 03 2018.

[81] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805–1812, 2014.

[82] T. Liu, L. Han, L. Ma, and D. Guo, "Audio-based deep music emotion recognition," *AIP Conference Proceedings*, vol. 1967, no. 1, p. 040021, 2018.

[83] Y. Liu and O. Sourina, "EEG-based subject-dependent emotion recognition algorithm using fractal dimension," in *Proc IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 3166–3171.

[84] F. H. Lopes da Silva, "The Impact of EEG/MEG Signal Processing and Modeling in the Diagnostic and Management of Epilepsy," *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 143–156, 2008.

[85] S. Losorelli, D. T. Nguyen, J. Dmochowski, and B. Kaneshiro, "NMED-T: A Tempo-Focused Dataset of Cortical and Behavioral Responses to Naturalistic Music," in *Proc. of the 18th International Society for Music Information Retrieval Conference*, 2017.

[86] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[87] Y. Luo and B.-L. Lu, "EEG Data Augmentation for Emotion Recognition Using a Conditional Wasserstein GAN," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2535–2538.

[88] B. Mandelbrot, "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," *Science*, vol. 156, no. 3775, pp. 636–638, 1967.

[89] L. Mano, B. Faiçal, V. Gonçalves, G. Pessin, P. Gomes, A. de Carvalho, and J. Ueyama, "An intelligent and generic approach for detecting human emotions: a case study with facial expressions," *Soft Computing*, vol. 24, 06 2020.

[90] P. Maragos, "Fractal Signal Analysis Using Mathematical Morphology," in *Advances in Electronics and Electron Physics*, P. W. Hawkes, Ed. Academic Press, 1994, vol. 88, pp. 199–246.

[91] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[92] ——, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[93] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1925–1932, 1999.

[94] P. D. Marrero Fernandez, F. A. Guerrero Pena, T. Ing Ren, and A. Cunha, "FER-Att: Facial Expression Recognition With Attention Net," in *Proc. Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[95] E. Mas-Herrero, A. Dagher, M. Farrés-Franch, and R. J. Zatorre, "Unraveling the Temporal Dynamics of Reward Signals in Music-Induced Pleasure with TMS," *Journal of Neuroscience*, vol. 41, no. 17, pp. 3889–3899, 2021.

[96] J. H. McDermott and A. J. Oxenham, "Music perception, pitch, and the auditory system," *Current Opinion in Neurobiology*, vol. 18, no. 4, pp. 452–463, 2008, sensory systems.

[97] V. Menon and D. Levitin, "The rewards of music listening: Response and physiological connectivity of the mesolimbic system," *NeuroImage*, vol. 28, no. 1, pp. 175–184, 2005.

[98] H. Minkowski, "Ueber die Begriffe Länge, Oberfläche und Volumen." *Jahresbericht der Deutschen Mathematiker-Vereinigung*, vol. 9, no. 1, pp. 115–121, 1901.

[99] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, 2017, pp. 2227–2231.

[100] A. Mollahosseini, D. Chan, and M. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[101] S. Nagel, "Towards a Home-use BCI: Fast Asynchronous Control and Robust non-Control State Detection," Ph.D. dissertation, Universität Tübingen, 12 2019.

[102] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," in *Proc. ICASSP*, 2019, pp. 7390–7394.

[103] S. Nozaradan, I. Peretz, M. Missal, and A. Mouraux, "Tagging the Neuronal Entrainment to Beat and Meter," *Journal of Neuroscience*, vol. 31, no. 28, pp. 10 234–10 240, 2011.

[104] A. Ofner and S. Stober, "Shared Generative Representation of Auditory Concepts and EEG to Reconstruct Perceived and Imagined Music," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 392–399.

[105] Z. Y. Ong, A. Saidatul, and Z. Ibrahim, "Power Spectral Density Analysis for Human EEG-based Biometric Identification," in *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, 2018, pp. 1–6.

[106] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. Paiva, "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis," in *10th International Symposium on Computer Music Multidisciplinary Research – CMMR'2013*, 2013.

[107] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Transactions on Affective Computing*, pp. 614–626, 2020.

[108] R. Panda, B. Rocha, and R. P. Paiva, "Music Emotion Recognition with Standard and Melodic Audio Features," *Applied Artificial Intelligence*, vol. 29, no. 4, pp. 313–334, 2015.

[109] C.-K. Peng, S. Buldyrev, S. Havlin, M. Simons, H. Stanley, and A. Goldberger, "Mosaic Organization of DNA Nucleotides," *Physical Review. E*, vol. 49, 03 1994.

[110] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," in *AAAI*, 2018.

[111] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 81–97, 2010.

[112] L. Piho and T. Tjahjadi, "A Mutual Information based Adaptive Windowing of Informative EEG for Emotion Recognition," *IEEE Transactions on Affective Computing*, 2018.

[113] V. Pitsikalis and P. Maragos, "Analysis and classification of speech signals by generalized fractal dimension features," *Speech Communication*, vol. 51, no. 12, pp. 1206–1223, 2009.

[114] R. Plutchik, "A Psychoevolutionary Theory of Emotions," *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982.

[115] B. Podobnik and H. Stanley, "Detrended Cross-Correlation Analysis: A New Method for Analyzing 2 Nonstationary Time Series." *Physical Review Letters*, 2008.

[116] H. Poikonen, V. Alluri, E. Brattico, O. Lartillot, M. Tervaniemi, and M. Huotilainen, "Event-related Brain Responses while Listening to Entire Pieces of Music," *Neuroscience*, vol. 312, pp. 58–73, 2016.

[117] J. Pons, T. Lidy, and X. Serra, "Experimenting with Musically Motivated Convolutional Neural Networks," in *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016, pp. 1–6.

[118] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-End Learning for Music Audio Tagging at Scale," in *Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[119] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[120] C. Qing, R. Qiao, X. Xu, and Y. Cheng, "Interpretable Emotion Recognition Using EEG Signals," *IEEE Access*, vol. 7, pp. 94 160–94 170, 2019.

[121] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[122] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, "Attention-based Modeling for Emotion Detection and Classification in Textual Conversations," *arXiv preprint arXiv:1906.07020*, 2019.

[123] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski, "Deep State Space Models for Time Series Forecasting," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 7785–7794.

[124] C. Raphael, "Automatic Transcription of Piano Music," in *ISMIR*, 2002.

[125] F. Raposo, D. M. de Matos, R. Ribeiro, S. Tang, and Y. Yu, "Towards Deep Modeling of Music Semantics using EEG Regularizers," *arXiv:1712.05197*, 2017.

[126] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[127] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review*, vol. 65, no. 6, p. 386, 1958.

[128] J. A. Russell, "A Circumplex Model of Affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[129] O. W. Sacks, *Musicophilia : Tales of Music and the Brain*, 1st ed. New York: Alfred A. Knopf, 2007.

[130] F. Samson, T. Zeffiro, A. Toussaint, and P. Belin, "Stimulus Complexity and Categorical Effects in Human Auditory Cortex: An Activation Likelihood Estimation Meta-Analysis," *Frontiers in Psychology*, vol. 1, p. 241, 2011.

[131] S. Sanei and J. Chambers, *Brain–Computer Interfacing*. John Wiley & Sons, Ltd, 2007, ch. 7, pp. 239–265.

[132] M. Sarprasatham, "Emotion Recognition: A Survey," *International Journal of Advanced Research in Computer Science*, vol. 3, pp. 14–19, 01 2015.

[133] S. Schachter and J. Singer, "Cognitive, Social, and Physiological Determinants of Emotional State." *Psychological Review*, p. 379–399, 1962.

[134] R. S. Schaefer, P. Desain, and P. Suppes, "Structural Decomposition of EEG Signatures of Melodic Processing," *Biological Psychology*, pp. 253–259, 2009.

[135] K. R. Scherer and H. G. Wallbott, "Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning," *Journal of Personality and Social Psychology*, vol. 66, no. 2, pp. 310–328, 1994.

[136] T. Schäfer, P. Sedlmeier, C. Städtler, and D. Huron, "The Psychological Functions of Music Listening," *Frontiers in Psychology*, vol. 4, p. 511, 2013.

[137] M. Shamsyeh Zahedi and J. Zeil, "Fractal dimension and the navigational information provided by natural scenes," *PLOS ONE*, vol. 13, no. 5, pp. 1–19, 05 2018.

[138] C. Shan, S. Gong, and P. W. McOwan, "Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study," *Image Vision Computation*, vol. 27, no. 6, p. 803–816, May 2009.

[139] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A Review of Emotion Recognition Using Physiological Signals," *Sensors*, vol. 18, no. 7, 2018.

[140] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, Sep. 2014.

[141] N. Singer, N. Jacoby, T. Lin, G. Raz, L. Shpigelman, G. Gilam, R. Y. Granot, and T. Hendler, "Common modulation of limbic network activation underlies musical emotions as they unfold," *NeuroImage*, vol. 141, pp. 517–529, 2016.

[142] F. M. Smits, C. Porcaro, C. Cottone, A. Cancelli, P. M. Rossini, and F. Tecchio, "Electroencephalographic Fractal Dimension in Healthy Ageing and Alzheimer's Disease," *PLOS ONE*, vol. 11, no. 2, pp. 1–16, 02 2016.

[143] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, "A Bayesian Framework for Video Affective Representation," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[144] Y. Song, S. Dixon, and M. Pearce, "Evaluation of Musical Features for Emotion Classification," in *Proceedings of the 13th International Society on Music Information Retrieval Conference (ISMIR)*, 2012.

[145] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[146] S. Stober, T. Prätzlich, and M. Müller, "Brain Beats: Tempo Extraction from EEG Data," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 01 2016, pp. 276–282.

[147] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Towards Music Imagery Information Retrieval: Introducing the OpenMIIR Dataset of EEG Recordings from Music Perception and Imagination." in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[148] C. Strapparava and A. Valitutti, "WordNet Affect: an Affective Extension of WordNet," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004.

[149] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning.* Springer, 10 2015, vol. 36.

[150] H. Takayasu, *Fractals in the Physical Sciences.* Manchester Univ Pr, 1990.

[151] J. Tao and T. Tan, "Affective Computing: A Review," in *Affective Computing and Intelligent Interaction.* Springer Berlin Heidelberg, 2005, pp. 981–995.

[152] A. W. Toga, K. A. Clark, P. M. Thompson, D. W. Shattuck, and J. D. Van Horn, "Mapping the Human Connectome," *Neurosurgery*, vol. 71, no. 1, pp. 1–5, 07 2012.

[153] P. Toiviainen, V. Alluri, E. Brattico, M. Wallentin, and P. Vuust, "Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data," *NeuroImage*, vol. 88, pp. 170–180, 2014.

[154] M. Trimble and D. Hesdorffer, "Music and the Brain: The Neuroscience of Music and Musical Appreciation," *BJPsych. International*, vol. 14, p. 28–31, 2017.

[155] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE ," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[156] G. K. P. Veeramallu, Y. Anupalli, S. k. Jilumudi, and A. Bhattacharyya, "EEG-based Automatic Emotion Recognition using EMD and Random Forest Classifier," in *Proceedings of the International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–6.

[157] A. Vinay, A. Lerch, and G. Leslie, "Mind the Beat: Detecting Audio Onsets from EEG Recordings of Music Listening," *ArXiv*, vol. abs/2102.06393, 2021.

[158] J. L. Walker, "Alpha EEG correlates of Performance on a Music Recognition Task," in *Psychobiology*, 1980.

[159] W. G. Walter, "The Electro-Encephalogram in Cases of Cerebral Tumour," *Proceedings of the Royal Society of Medicine*, vol. 30, no. 5, pp. 579–598, 1937.

[160] W. G. Walter and V. J. Dovey, "Electro-Engephalography in Cases of Subcortical Tumour," *Journal of Neurology, Neurosurgery & Psychiatry*, pp. 57–65, 1944.

[161] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial Cross-Modal Retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia.* New York, NY, USA: Association for Computing Machinery, 2017, p. 154–162.

[162] F. Wang, S.-h. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks," in *MultiMedia Modeling.* Cham: Springer International Publishing, 2018, pp. 82–93.

[163] J. Wang, L. Yu, K. Lai, and X. Zhang, "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model," in *ACL*, 2016.

[164] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A Comprehensive Survey on Cross-modal Retrieval," *ArXiv*, vol. abs/1607.06215, 2016.

[165] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On Deep Multi-View Representation Learning," in *International Conference on Machine Learning (ICML)*, 2015, pp. 1083–1092.

[166] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep Variational Canonical Correlation Analysis," *arXiv preprint arXiv:1610.03454*, 2016.

[167] X. Wang, D. Nie, and B. Lu, "EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines," in *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, 2011.

[168] Y. Wang, Z. Huang, B. McCane, and P. Neo, "EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2018.

[169] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[170] S. M. Wilson, I. Molnar-Szakacs, and M. Iacoboni, "Beyond Superior Temporal Cortex: Intersubject Correlations in Narrative Speech Comprehension," *Cerebral Cortex*, vol. 18, no. 1, pp. 230–242, 05 2007.

[171] J. Wolpaw, N. Birbaumer, W. Heetderks, D. McFarland, P. Peckham, G. Schalk, E. Donchin, L. Quatrano, C. Robinson, and T. Vaughan, "Brain-Computer Interface Technology: A Review of the First International Meeting," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 164–173, 2000.

[172] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal Metric Learning for Tag-based Music Retrieval," *arXiv preprint:2010.16030*, 2020.

[173] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, 12 2013.

[174] Z. R. Zald DH, *Music*. CRC Press/Taylor & Francis, 2011.

[175] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[176] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-Piloted Deep Network for Facial Expression Recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 425–442.

[177] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep Supervised Cross-Modal Retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 394–10 403.

[178] W.-L. Zheng and B.-L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[179] W. Zhou, "Multifractal Detrended Cross-Correlation Analysis for two Nonstationary Signals." *Physical Review. Statistical, Nonlinear, and Soft Matter Physics*, 2008.

[180] A. Zlatintsi and P. Maragos, "Multiscale Fractal Analysis of Musical Instrument Signals With Application to Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 737–748, 2013.