Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εργαστηριο Ψηφιακης Επεξεργασιας Εικονας και Σηματων

Explainable Artificial Intelligence for EEG Analysis

Diploma Thesis

by

Georgios Kontos

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Ψηφιακής Επεξεργασίας Εικόνας και Σημάτων

Explainable Artificial Intelligence for EEG Analysis

Diploma Thesis

by

Georgios Kontos

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2η Οκτωβρίου 2025.

.....	.....	.....
Γεώργιος Στάμου	Αθανάσιος Βουλόδημος	Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.	Ομ. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2025

.....

Γεωργιος Κοντος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Στην οικογένειά μου

Περίληψη

Οι πρόσφατες εξελίξεις στον τομέα της εξηγήσιμης τεχνητής νοημοσύνης (XAI) έχουν αναδιαμορφώσει τον τρόπο αξιολόγησης και εφαρμογής των συστημάτων τεχνητής νοημοσύνης. Το επίκεντρο έχει μετατοπιστεί από τη μεγιστοποίηση της ακρίβειας και μόνο, στην εξασφάλιση διαφάνειας, ερμηνευσιμότητας και αξιοπιστίας — ιδιότητες που είναι απαραίτητες όταν η τεχνητή νοημοσύνη χρησιμοποιείται σε κρίσιμους τομείς, όπως η υγειονομική περίθαλψη. Σε κλινικά περιβάλλοντα, η ερμηνευσιμότητα αποτελεί προϋπόθεση για την υιοθέτηση των μοντέλων, καθώς οι επαγγελματίες του ιατρικού τομέα πρέπει να κατανοούν και να επικυρώνουν τη συλλογιστική τους. Τα διαφανή μοντέλα όχι μόνο ενισχύουν την εμπιστοσύνη των χρηστών, αλλά συμβάλλουν και στην επιστημονική ανακάλυψη, αποκαλύπτοντας σημαντικές σχέσεις μέσα σε πολύπλοκα βιοϊατρικά δεδομένα.

Το ηλεκτροεγκεφαλογράφημα (EEG) έχει αναδειχθεί ως ένας σημαντικός τομέας εφαρμογής της XAI, λόγω της περίπλοκης και θορυβώδους φύσης των εγκεφαλικών σημάτων. Το EEG παρέχει ένα μη επεμβατικό και οικονομικά αποδοτικό μέσο παρακολούθησης της νευρικής δραστηριότητας. Ωστόσο, η ανάλυσή του παραμένει δύσκολη, εξαιτίας του χαμηλού λόγου σήματος προς θόρυβο, των μη στατικών ιδιοτήτων και της μεγάλης μεταβλητότητας μεταξύ ατόμων. Οι τεχνικές τεχνητής νοημοσύνης —ιδιαίτερα η μηχανική μάθηση και η βαθιά μάθηση— έχουν δείξει σημαντικές δυνατότητες στην εξαγωγή χρήσιμων πληροφοριών από τα δεδομένα EEG, για εργασίες όπως η ανίχνευση επιληπτικών κρίσεων, η αναγνώριση συναισθημάτων και η σταδιοποίηση του ύπνου. Ωστόσο, η αδιαφανής φύση αυτών των μοντέλων περιορίζει την κλινική τους χρησιμότητα, καθώς η λογική πίσω από τις προβλέψεις τους συχνά παραμένει ασαφής. Οι τεχνικές XAI συμβάλλουν στην αντιμετώπιση αυτού του ζητήματος, εντοπίζοντας σχετικά χαρακτηριστικά, οπτικοποιώντας τις διαδικασίες λήψης αποφάσεων και συνδέοντας την αλγοριθμική συλλογιστική με φυσιολογικούς μηχανισμούς.

Η παρούσα διατριβή διερευνά τον τρόπο με τον οποίο η XAI μπορεί να βελτιώσει τόσο την κατανόηση όσο και την αξιοπιστία στην ανάλυση EEG. Συνδυάζει καθιερωμένες διαδικασίες προεπεξεργασίας και εξαγωγής χαρακτηριστικών με ερμηνεύσιμες μεθόδους μάθησης. Η μελέτη εφαρμόζει αυτές τις τεχνικές στην ανίχνευση επιληπτικών κρίσεων —μια κλινικά σημαντική και υπολογιστικά απαιτητική εργασία— χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων EEG ανοιχτής πρόσβασης. Πέρα από την αξιολόγηση της απόδοσης, η εργασία διερευνά τον τρόπο με τον οποίο μέθοδοι εξήγησης, όπως το SHAP και η συλλογιστική βάσει κανόνων, μπορούν να αποκαλύψουν την υποκείμενη δομή των αποφάσεων του μοντέλου και τη συσχέτισή τους με μοτίβα εγκεφαλικής δραστηριότητας.

Τελικά, η εργασία αυτή στοχεύει να αποδείξει ότι η ερμηνευσιμότητα δεν αποτελεί απλώς ένα βοηθητικό χαρακτηριστικό, αλλά μια απαραίτητη προϋπόθεση για την εφαρμογή της τεχνητής νοημοσύνης στον ιατρικό τομέα. Ενσωματώνοντας τις αρχές της XAI τόσο στη διαδικασία μοντελοποίησης όσο και στην αξιολόγηση, η εργασία συμβάλλει στην ανάπτυξη διαφανών, ανθρωποκεντρικών συστημάτων τεχνητής νοημοσύνης, ικανών να υποστηρίξουν τη λήψη κλινικών αποφάσεων και να προωθήσουν την κατανόησή μας για τη λειτουργία του εγκεφάλου.

Λέξεις-κλειδιά — Ηλεκτροεγκεφαλογράφημα, Εξηγήσιμη Τεχνητή Νοημοσύνη, Μηχανική μάθηση, Βαθειά μάθηση, SHAP, Ανίχνευση επιληπτικών κρίσεων

Abstract

Recent advancements in eXplainable Artificial Intelligence (XAI) have reshaped how artificial intelligence systems are evaluated and applied. The focus has shifted from maximizing accuracy alone to ensuring transparency, interpretability, and trustworthiness—qualities that are essential when AI is used in sensitive domains such as healthcare. In clinical contexts, interpretability is a prerequisite for adoption, as medical professionals must understand and validate model reasoning before relying on its outcomes. Transparent models not only enhance user confidence but also contribute to scientific discovery by uncovering meaningful relationships within complex biomedical data.

Electroencephalography (EEG) has emerged as a major application area for XAI due to the intricate and noisy nature of brain signals. EEG provides a non-invasive and cost-effective means of monitoring neural activity, yet its analysis remains challenging because of the data's low signal-to-noise ratio, nonstationary properties, and variability across subjects. Artificial intelligence techniques—particularly machine learning and deep learning—have shown great promise in extracting useful information from EEG data for tasks such as seizure detection, emotion recognition, and sleep staging. However, the opaque nature of these models limits their clinical usability, as the rationale behind predictions often remains unclear. XAI techniques help address this issue by identifying relevant features, visualizing decision processes, and linking algorithmic reasoning to physiological mechanisms.

This thesis investigates how XAI can enhance both understanding and reliability in EEG analysis. It combines established preprocessing and feature extraction procedures with interpretable learning methods to balance predictive accuracy and transparency. The study specifically applies these techniques to seizure detection, a clinically important and computationally demanding task, using a large open-access EEG dataset. Beyond evaluating performance, the work explores how explainability methods such as SHAP and rule-based reasoning can reveal the underlying structure of model decisions and their correspondence to brain activity patterns.

Ultimately, the thesis aims to demonstrate that interpretability is not merely an auxiliary feature but a core requirement for trustworthy EEG analysis. By integrating XAI principles into both the modeling and evaluation process, this work contributes toward developing transparent, human-centered AI systems capable of supporting clinical decision-making and advancing our understanding of brain function.

Keywords — Electroencephalography, eXplainable Artificial Intelligence (XAI), Machine learning, Deep learning, SHAP, Seizure detection

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας δεν θα ηταν δυνατή χωρίς τη βοήθεια και τη συμπαράσταση ορισμένων ανθρώπων, στους οποίους και θα ήθελα να εκφράσω τις ειλικρινές μου ευχαριστίες.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Στάμου Γεώργιο για την δυνατότητα που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα. Επιπλέον, θέλω να ευχαριστήσω και τους υποψήφιους διδάκτορες Λυμπεράτο Βασίλη, Σπανό Νίκο και Μενή-Μαστρομιχαλάκη Ορφέα, για την καθοδήγηση, τις υποδείξεις, την κατανόηση και υπομονή για όλη την διάρκεια της στενής συνεργασίας που είχαμε.

Στη συνέχεια θα ήθελα να ευχαριστήσω τους γονείς μου Γιάννη και Δέσποινα, που χωρίς την ψυχική, ηθική και υλική υποστήριξη, την εμπιστοσύνη και αδιαπραγμάτευτη αγάπη τους τίποτα από αυτά δεν θα ήταν εφικτό. Ακόμα, θα ήθελα να ευχαριστήσω τα αδέρφια μου Παναγιώτα, Θανάση και Ιάσονα για τη συνεχή συμπαράσταση τους και την ικανότητα τους να με κάνουν πάντα να (χαμο)γελώ.

Τέλος θα ήθελα να ευχαριστήσω του φίλους μου και τους συμφοιτητές μου για την ενθάρρυνση, τις συζητήσεις, τη συνεργασία που μοιραστήκαμε σε κάθε στάδιο αυτής της πορείας αλλά και που φρόντισαν να κάνουν μοναδικά τα 5 χρόνια σπουδών.

Κοντός Γεώργιος, Οκτώβρης 2025

Contents

List of Figures

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Εισαγωγή

Το ηλεκτροεγκεφαλογράφημα (EEG) παραμένει ένα βασικό διαγνωστικό εργαλείο για εγκεφαλικές παθήσεις λόγω της μη επεμβατικής φύσης της και της ευκολίας χρήσης. Βοηθά στον εντοπισμό διαταραχών ύπνου και στην ανίχνευση ανώμαλων προτύπων που σχετίζονται με την επιληψία. Ωστόσο, τα σήματα EEG έχουν χαμηλό πλάτος και καλύπτουν πολλαπλές συχνοτικές ζώνες, καθιστώντας τα δύσκολα στην ανάλυση με συμβατικά εργαλεία λογισμικού. Η τεχνητή νοημοσύνη (AI) μπορεί να γεφυρώσει αυτό το κενό ανιχνεύοντας κρυφά πρότυπα στα δεδομένα EEG, επιτρέποντας ανάλυση σε πραγματικό χρόνο για καθήκοντα ανίχνευσης και ταξινόμησης [170].

Παρά τις δυνατότητές της, η AI στην ανάλυση EEG αντιμετωπίζει μια σημαντική πρόκληση—τη "μαύρο κουτί" φύση της [29]. Για να είναι αξιόπιστα τα μοντέλα AI στη ιατρική διάγνωση και πρόγνωση, η διαφάνεια είναι κρίσιμη. Είναι απαραίτητο οι αποφάσεις να προκύπτουν από ουσιαστικά πρότυπα EEG και όχι από τυχαίες συσχετίσεις που υπάρχουν στα δεδομένα εκπαίδευσης. Σ' αυτό το πλαίσιο, η Εξηγήσιμη Τεχνητή Νοημοσύνη (XAI) παίζει σημαντικό ρόλο, προσφέροντας πληροφορίες για τη λογική πίσω από τις αποφάσεις των μοντέλων AI [38]. Οι τεχνικές XAI επιτρέπουν την ερμηνευσιμότητα των μοντέλων, αντιμετωπίζοντας ηθικά και κλινικά ζητήματα και ενισχύοντας την εμπιστοσύνη στην AI στον χώρο της υγειονομικής περίθαλψης. Η XAI ενισχύει τη διαφάνεια στις διαγνώσεις βασισμένες σε EEG, επισημαίνοντας τα πιο σημαντικά χαρακτηριστικά που επηρεάζουν τις προβλέψεις του μοντέλου. Τεχνικές όπως η εξαγωγή χαρακτηριστικών [122] και οι χάρτες σημαντικότητας (saliency maps) [169] αποκαλύπτουν ποιες ιδιότητες του EEG καθοδηγούν τις αποφάσεις της AI, διασφαλίζοντας τη συμφωνία τους με φυσιολογικές αρχές. Με την βελτίωση της ερμηνευσιμότητας και της αξιοπιστίας των μοντέλων, η XAI διευκολύνει την υιοθέτηση της AI σε πραγματικές κλινικές εφαρμογές.

Η παρούσα εργασία κατηγοριοποιεί και αναλύει τις πρόσφατες εξελίξεις στην XAI για ανάλυση EEG, επισημαίνοντας τον τρόπο με τον οποίο αντιμετωπίζουν βασικές προκλήσεις σε καθήκοντα σχετικά με το EEG, με τον τελικό στόχο να εισάγει το πεδίο στους ερευνητές με προσιτό και κατανοητό τρόπο. Παρουσιάζουμε διάφορα benchmarks και σύνολα δεδομένων που χρησιμοποιούνται κυρίως για την αξιολόγηση της ερμηνευσιμότητας στην ανάλυση EEG. Εντοπίζουμε κενά στην τρέχουσα έρευνα και εξερευνούμε μελλοντικές κατευθύνσεις, με στόχο την καθοδήγηση περαιτέρω εξελίξεων και εφαρμογών της εξηγησιμης AI σε αυτόν τον τομέα.

Πέρα από την ανασκόπηση της βιβλιογραφίας, αυτή η διπλωματική συμβάλλει με ένα πρακτικό πλαίσιο για την προεπεξεργασία και την εξαγωγή χαρακτηριστικών από EEG, προσαρμοσμένο στο έργο της ανίχνευσης επιληπτικών κρίσεων. Το πλαίσιο εφαρμόζεται σε ένα δημόσια διαθέσιμο σύνολο δεδομένων, που αρχικά χρησιμοποιήθηκε στο Una Europa Seizure Detection Challenge. Εκπαιδεύτηκαν και αξιολογήθηκαν διάφορα μοντέλα μηχανικής μάθησης και βαθιάς μάθησης, περιλαμβάνοντας κλασικές προσεγγίσεις όπως το Random Forest και προηγμένες αρχιτεκτονικές όπως το xLSTM, επιτρέποντας μια ολοκληρωμένη αξιολόγηση διαφορετικών παραδειγμάτων μοντελοποίησης. Για την αντιμετώπιση του κρίσιμου ζητήματος της ερμηνευσιμότητας, χρησιμοποιήσαμε το SHAP για ανάλυση ανάθεσης χαρακτηριστικών και ενσωματώσαμε το te2rules, μια καινοτόμο μέθοδο για παραγωγή ανθρώπινα αναγνώσιμων κανόνων από εκπαιδευμένα μοντέλα. Αυτές οι τεχνικές εξηγησιμότητας παρέχουν πληροφορίες για τα πιο σημαντικά χαρακτηριστικά για την ανίχνευση κρίσεων, γεφυρώνοντας το χάσμα μεταξύ προβλεπτικής ακρίβειας και κλινικής ερμηνευσιμότητας. Συνδυάζοντας ισχυρή μοντελοποίηση

με μεθόδους XAI, αυτή η εργασία προωθεί όχι μόνο την τεχνική κατανόηση αλλά επιδιώκει επίσης την παραγωγή αποτελεσμάτων που είναι ουσιαστικά και αξιόπιστα σε πραγματικά ιατρικά πλαίσια.

Η δομή της διπλωματικής έχει ως εξής:

- Αρχικά παρέχουμε όλο το απαραίτητο υπόβαθρο στις βασικές μεθόδους δέντρων αποφάσεων και συνδυασμένων (ensemble) μεθόδων μάθησης.
- Δίνουμε τα θεμέλια της Εξηγήσιμης Τεχνητής Νοημοσύνης (XAI) και της ανάλυσης EEG. Έπειτα, παρέχουμε μια λεπτομερή ανασκόπηση της βιβλιογραφίας σχετικά με τις μεθόδους XAI στην ανάλυση EEG.
- Τέλος, αναλύουμε τη μεθοδολογία μας, παρέχοντας λεπτομέρειες για την προεπεξεργασία και την εξαγωγή χαρακτηριστικών. Παρουσιάζουμε τα αποτελέσματα τόσο της απόδοσης όσο και της ανάλυσης ερμηνευσιμότητας και καταλήγουμε στα συμπεράσματά μας.

1.2 Θεωρητικό υπόβαθρο

1.2.1 Μοντέλα βασισμένα σε δέντρα και Μοντέλα ενσωμάτωσης

Στην παρούσα διπλωματική εργασία εστιάζουμε σε μοντέλα μηχανικής μάθησης βασισμένα σε δέντρα και στις επεκτάσεις τους μέσω μεθόδων συνόλων (ensemble). Ένα μεμονωμένο δέντρο απόφασης είναι εύκολο να κατανοηθεί και να ερμηνευθεί, αλλά συνήθως υποφέρει από υπερεφαρμογή (overfitting) και περιορισμένη ακρίβεια. Οι μέθοδοι συνόλων, όπως τα Τυχαία Δάση (Random Forests) και το Gradient Boosting, αντιμετωπίζουν αυτά τα προβλήματα συνδυάζοντας πολλά δέντρα, κάτι που γενικά οδηγεί σε πιο ανθεκτικές και ακριβείς προβλέψεις. Με την πάροδο του χρόνου, έχουν αναπτυχθεί αρκετές αποδοτικές υλοποιήσεις του gradient boosting, όπως τα XGBoost, LightGBM, CatBoost και HistGradientBoosting, οι οποίες πλέον χρησιμοποιούνται ευρέως στην πράξη. Παρακάτω περιγράφουμε συνοπτικά αυτά τα μοντέλα.

Δέντρα Απόφασης

Τα δέντρα αποφάσεων αποτελούν απλά αλλά ισχυρά προγνωστικά μοντέλα, τα οποία χωρίζουν τον χώρο χαρακτηριστικών σε περιοχές μέσω της επαναληπτικής διάσπασης των δεδομένων με βάση τις τιμές των χαρακτηριστικών [25]. Εκτιμώνται ιδιαίτερα για την ερμηνευσιμότητά τους και την ικανότητά τους να μοντελοποιούν μη γραμμικές σχέσεις. Ωστόσο, τα μεμονωμένα δέντρα είναι επιρρεπή σε υπερπροσαρμογή και υψηλή διακύμανση, γεγονός που τα καθιστά λιγότερο ανθεκτικά στην πράξη.

Τυχαία Δάση (Random Forests)

Τα Τυχαία Δάση (Random Forests) αντιμετωπίζουν τους περιορισμούς των μεμονωμένων δέντρων αποφάσεων, συνδυάζοντας πολλά δέντρα σε ένα πλαίσιο bagging [24]. Κάθε δέντρο εκπαιδεύεται σε ένα bootstrap δείγμα των δεδομένων και σε κάθε διάσπαση λαμβάνεται υπόψη μόνο ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτή η τυχαιότητα μειώνει τη συσχέτιση μεταξύ των δέντρων και βελτιώνει τη γενίκευση. Τα Τυχαία Δάση χρησιμοποιούνται ευρέως λόγω της ανθεκτικότητάς τους, της σχετικά χαμηλής ανάγκης για ρύθμιση παραμέτρων και της ισχυρής απόδοσης σε ποικιλία πεδίων.

Μέθοδοι Gradient Boosting

Το Gradient Boosting αποτελεί μία τεχνική συνόλου (ensemble), η οποία κατασκευάζει μοντέλα με διαδοχικό τρόπο, όπου κάθε νέος μαθητής επιχειρεί να διορθώσει τα σφάλματα των προηγούμενων [58]. Συνδυάζει αδύναμους ταξινομητές, συνήθως δέντρα αποφάσεων, σε ένα ισχυρό προγνωστικό μοντέλο μέσω βελτιστοποίησης βασισμένης σε κλίση (gradient-based optimization). Κατά την τελευταία εικοσαετία, έχουν αναπτυχθεί αποδοτικές υλοποιήσεις του gradient boosting, οι οποίες αντιμετωπίζουν ζητήματα κλιμακωσιμότητας, κανονικοποίησης καθώς και διαχείρισης κατηγορικών μεταβλητών. Στην παρούσα διατριβή αξιοποιήθηκαν οι ακόλουθες βιβλιοθήκες.

XGBoost Το XGBoost (Extreme Gradient Boosting) αποτελεί ένα από τα πλέον ευρέως χρησιμοποιούμενα πλαίσια gradient boosting [28]. Εισάγει καινοτομίες όπως προσεγγίσεις δεύτερης τάξης για την κλίση, μάθηση ευαισθητοποιημένη στην αραιότητα (sparsity-aware learning), καθώς και αποδοτική διαχείριση ελλιπών τιμών. Επιπλέον, εφαρμόζει τεχνικές όπως shrinkage και υποδειγματοληψία χαρακτηριστικών, οι οποίες περιορίζουν την υπερπροσαρμογή. Ως εκ τούτου, ο αλγόριθμος επιδεικνύει υψηλή αποτελεσματικότητα σε προβλήματα πρόβλεψης με δομημένα δεδομένα.

LightGBM Το LightGBM (Light Gradient Boosting Machine) αναπτύχθηκε με σκοπό τη βελτίωση της αποδοτικότητας εκπαίδευσης και της κλιμακωσιμότητας [72]. Σε αντίθεση με την παραδοσιακή ανάπτυξη δέντρων κατά επίπεδα, το LightGBM εφαρμόζει στρατηγική ανάπτυξης ανά φύλλο με περιορισμούς βάθους, γεγονός που επιτρέπει τη δημιουργία πιο σύνθετων δέντρων και, συχνά, την επίτευξη υψηλότερης ακρίβειας. Παράλληλα, χρησιμοποιεί κατανομή χαρακτηριστικών σε ιστογράμματα (histogram-based binning), μειώνοντας σημαντικά τον χρόνο εκπαίδευσης και τη μνήμη που απαιτείται, στοιχείο που το καθιστά ιδιαιτέρως κατάλληλο για μεγάλης κλίμακας σύνολα δεδομένων.

CatBoost Το CatBoost αποτελεί αλγόριθμο gradient boosting ο οποίος έχει σχεδιαστεί για την αποδοτική διαχείριση κατηγορικών μεταβλητών χωρίς την ανάγκη εκτεταμένης προεπεξεργασίας [125]. Εισάγει την τεχνική ordered boosting, μία προσέγγιση βασισμένη σε μεταθέσεις, η οποία μετριάζει το prediction shift και περιορίζει την υπερπροσαρμογή. Επιπλέον, ο αλγόριθμος επιδεικνύει ισχυρή βασική απόδοση ακόμη και με ελάχιστη ρύθμιση παραμέτρων, γεγονός που τον καθιστά ιδιαιτέρως χρήσιμο σε εφαρμοσμένα σενάρια μηχανικής μάθησης.

HistGradientBoosting Το HistGradientBoosting αποτελεί μία αποδοτική υλοποίηση του gradient boosting, ενσωματωμένη στη βιβλιοθήκη `scikit-learn` [120]. Εμπνευσμένο από το LightGBM, χρησιμοποιεί κατανομή χαρακτηριστικών σε ιστογράμματα ώστε να επιταχύνει την εκπαίδευση και να μειώσει τις απαιτήσεις σε μνήμη. Αν και υπολείπεται σε λειτουργικό εύρος σε σύγκριση με τα XGBoost, LightGBM ή CatBoost, ενσωματώνεται απρόσκοπτα στο οικοσύστημα της scikit-learn και παρουσιάζει ανταγωνιστική απόδοση σε σύνολα δεδομένων μεσαίου μεγέθους.

Συνδυασμός Μοντέλων (Model Ensembling)

Οι μέθοδοι συνόλων (ensemble methods) συνδυάζουν πολλαπλά μοντέλα με στόχο την επίτευξη καλύτερης προγνωστικής απόδοσης σε σχέση με τους μεμονωμένους ταξινομητές [35]. Το bagging μειώνει τη διακύμανση, υπολογίζοντας τον μέσο όρο των προβλέψεων από ανεξάρτητα μοντέλα, ενώ το boosting μειώνει τη μεροληψία μέσω της διαδοχικής βελτίωσης των προηγούμενων ταξινομητών. Μια άλλη στρατηγική συνόλων, το stacking, συνδυάζει ετερογενή μοντέλα με τη χρήση ενός μετα-ταξινομητή (meta-learner), ώστε να βελτιστοποιήσει την προγνωστική ακρίβεια. Αυτές οι προσεγγίσεις έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές σε δομημένα ταμπουλαρισμένα δεδομένα, όπου τα ensembles συχνά ξεπερνούν σε απόδοση τα μεμονωμένα μοντέλα.

1.2.2 Εξηγήσιμη Τεχνητή Νοημοσύνη

Βασικές Έννοιες

Η Εξηγήσιμη Τεχνητή Νοημοσύνη (Explainable Artificial Intelligence – XAI) περιλαμβάνει μεθόδους και τεχνικές που έχουν σχεδιαστεί ώστε να καταστήσουν κατανοητές στους ανθρώπους τις αποφάσεις και τις προβλέψεις των συστημάτων Τεχνητής Νοημοσύνης (ΤΝ) [37]. Καθώς τα «μαύρα κουτιά» (black-box) μοντέλα ΤΝ γίνονται ολοένα και πιο διαδεδομένα σε κρίσιμους τομείς, η ανάγκη για ερμηνεύσιμη και διαφανή ΤΝ έχει αυξηθεί αναλόγως [61]. Οι προσεγγίσεις XAI επιδιώκουν να ισορροπήσουν ανάμεσα στην παροχή επεξηγήσεων κατανοητών από τον άνθρωπο και στη διατήρηση ισχυρής απόδοσης του μοντέλου [106]. Οι μέθοδοι αυτές μπορούν να κατηγοριοποιηθούν με βάση διάφορες διαστάσεις: την προσέγγιση ερμηνευσιμότητας (εκ των υστέρων – post-hoc έναντι εκ των προτέρων – ante-hoc), το εύρος της εξήγησης (ολική – global έναντι τοπικής – local), τον τύπο εξήγησης (αποδόσεις χαρακτηριστικών, κανόνες, παραδείγματα, οπτικοποιήσεις) και την εξάρτησή τους από την αρχιτεκτονική του μοντέλου (ανεξάρτητες από το μοντέλο – model-agnostic έναντι εξειδικευμένων για συγκεκριμένο μοντέλο – model-specific) [13].

Κατηγορίες

Προσέγγιση Ερμηνευσιμότητας Μία βασική διάκριση γίνεται μεταξύ *post-hoc* και *ante-hoc* ερμηνευσιμότητας. Οι *post-hoc* τεχνικές εφαρμόζονται μετά την εκπαίδευση του μοντέλου, προσφέροντας επεξηγήσεις για πολύπλοκα «μαύρα κουτιά» μέσω μεθόδων όπως το LIME [132], το οποίο προσεγγίζει τοπικά τα όρια απόφασης. Αντίθετα, οι *ante-hoc* προσεγγίσεις δίνουν προτεραιότητα σε εγγενώς ερμηνεύσιμα μοντέλα, όπως τα δέντρα αποφάσεων ή οι λίστες κανόνων [84], όπου η λογική του μοντέλου είναι διαφανής εκ σχεδιασμού χωρίς την ανάγκη πρόσθετων μεθόδων επεξήγησης.

Εμβέλεια της Εξήγησης Μία ακόμη θεμελιώδης διάκριση αφορά την εμβέλεια των εξηγήσεων: *παγκόσμια* έναντι *τοπικών*. Οι παγκόσμιες εξηγήσεις στοχεύουν να αποσαφηνίσουν τη συνολική λογική λήψης αποφάσεων ενός μοντέλου, παρέχοντας εικόνα για τη γενική του συμπεριφορά μέσω τεχνικών όπως η εξαγωγή κανόνων ή τα υποκατάστατα μοντέλα. Οι τοπικές εξηγήσεις, αντίθετα, επικεντρώνονται σε μεμονωμένες προβλέψεις, εντοπίζοντας τα χαρακτηριστικά ή τα δείγματα που ασκούν τη μεγαλύτερη επιρροή [132]. Έτσι παρέχουν λεπτομερή εικόνα, ιδιαίτερα χρήσιμη σε περίπλοκες ή κρίσιμες περιπτώσεις.

Τύπος Εξήγησης Οι εξηγήσεις μπορούν να λάβουν διάφορες μορφές. Οι μέθοδοι απόδοσης χαρακτηριστικών εκτιμούν τη συνεισφορά κάθε εισόδου σε μια πρόβλεψη [155], ενώ οι μεθοδολογίες βασισμένες σε κανόνες παράγουν ερμηνεύσιμους κανόνες απόφασης [171]. Οι εξηγήσεις βασισμένες σε παραδείγματα χρησιμοποιούν τυπικά ή αντιπαραθετικά δείγματα για να απεικονίσουν τη συμπεριφορά του μοντέλου [76], ενώ οι τεχνικές οπτικοποίησης προσφέρουν διαισθητικές αναπαραστάσεις των διαδικασιών λήψης αποφάσεων, ιδιαίτερα χρήσιμες για δεδομένα εικόνας ή σήματος [106]. Η επιλογή του τύπου εξήγησης εξαρτάται συχνά από το πεδίο εφαρμογής, το εκάστοτε έργο και τις ανάγκες των χρηστών.

Εξάρτηση από την Αρχιτεκτονική του Μοντέλου Οι μέθοδοι XAI μπορούν επίσης να διακριθούν βάσει της εξάρτησής τους από το υποκείμενο μοντέλο. Οι ανεξάρτητες από το μοντέλο (model-agnostic) μέθοδοι αντιμετωπίζουν το μοντέλο ως «μαύρο κουτί», αναλύοντας μόνο τις εισόδους και τις εξόδους χωρίς πρόσβαση σε εσωτερικές παραμέτρους [90]. Τέτοιες προσεγγίσεις, όπως το SHAP και άλλες μεθόδους που βασίζονται σε διαταραχές (perturbations), προσφέρουν ευρεία εφαρμοσιμότητα σε διαφορετικούς τύπους μοντέλων. Αντίθετα, οι εξειδικευμένες για συγκεκριμένα μοντέλα (model-specific) μέθοδοι αξιοποιούν την εσωτερική δομή συγκεκριμένων αλγορίθμων—όπως οι μέθοδοι που βασίζονται σε κλίσεις για νευρωνικά δίκτυα—προκειμένου να παραγάγουν πιο ακριβείς ερμηνείες [142].

Τεχνικά θεμέλια ευρέως χρησιμοποιούμενων τεχνικών XAI

Σε αυτήν την υποενότητα παρουσιάζονται τα τεχνικά θεμέλια των πλέον διαδεδομένων μεθόδων XAI που χρησιμοποιούνται στην ανάλυση EEG. Αναλύονται οι θεωρητικές διατυπώσεις και οι βασικές εξισώσεις, παρέχοντας το απαραίτητο υπόβαθρο. Οι τεχνικές που εξετάζονται είναι οι πλέον συνηθισμένες στην εφαρμογή XAI για ανάλυση EEG.

SHapley Additive exPlanations (SHAP) Οι τιμές SHAP αποτελούν ένα ενοποιημένο μέτρο σημαντικότητας χαρακτηριστικών, βασισμένο στη θεωρία παιγνίων, το οποίο εξετάζει πώς διαφορετικοί «παίκτες» (χαρακτηριστικά) συμβάλλουν στη συνολική απόδοση [90]. Για την αξιολόγηση ενός δείγματος, κάθε χαρακτηριστικό λαμβάνει μια τιμή SHAP που δείχνει τη σχετική του συνεισφορά στη λήψη της απόφασης του μοντέλου. Ο επίσημος ορισμός δίνεται από:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus \{i\}) \right]$$

όπου x είναι το προς εξήγηση δείγμα, f το μοντέλο, i το χαρακτηριστικό προς αξιολόγηση, M ο αριθμός των χαρακτηριστικών και x' το σύνολο όλων των πιθανών υποσυνόλων/παραλλαγών του x.

Local Interpretable Model-agnostic Explanations (LIME) Το LIME αποτελεί μια τοπική, ανεξάρτητη από το μοντέλο μέθοδο που προσεγγίζει τοπικά ένα σύνθετο ταξινομητή με ένα ερμηνεύσιμο πρότυπο (π.χ.

γραμμικό μοντέλο ή δέντρο) [132]. Ο στόχος είναι η εύρεση ενός ερμηνεύσιμου μοντέλου $g \in G$ το οποίο είναι τοπικά πιστό στον βασικό ταξινομητή f. Η εξήγηση προκύπτει από την επίλυση:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

όπου x είναι το προς εξήγηση δείγμα, $\mathcal{L}$ μετρά την τοπική προσέγγιση του g προς το f στην τοπικότητα ορισμένη από π_x, και $\Omega(g)$ τιμωρεί την πολυπλοκότητα του g. Το LIME παράγει εξηγήσεις στο επίπεδο των δεδομένων, χρησιμοποιώντας διαταραχές (perturbations) των χαρακτηριστικών με βάση στατιστικά χαρακτηριστικά των δεδομένων εκπαίδευσης.

Gradient-weighted Class Activation Mapping (Grad-CAM) Το Grad-CAM έχει σχεδιαστεί για συνελικτικά νευρωνικά δίκτυα (CNNs) και χρησιμοποιεί τις παραγώγους που ρέουν στο τελευταίο συνελικτικό επίπεδο για να αποδώσει σημαντικότητα σε χωρικές θέσεις του χάρτη χαρακτηριστικών [142]. Ο συντελεστής σημαντικότητας για τον k-οστό χάρτη χαρακτηριστικών και την κλάση c υπολογίζεται ως:

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial S_c(x)}{\partial f_k(x)_{i,j}}$$

όπου Z ο συνολικός αριθμός χωρικών θέσεων στον χάρτη, $S_c(x)$ η βαθμολογία (score) της κλάσης c, και $f_k(x)_{i,j}$ η ενεργοποίηση στη θέση (i, j). Ο Grad-CAM χάρτης θερμότητας για την κλάση c δίνεται από:

$$M_c^{\text{Grad}}(x) = \text{ReLU}\left(\sum_k \alpha_c^k f_k(x) \right)$$

όπου η ReLU διασφαλίζει ότι λαμβάνονται υπόψη μόνο θετικές συνεισφορές.

DeepLIFT (Deep Learning Important FeaTures) Το DeepLIFT προτάθηκε ως μία αναδρομική μέθοδος ερμηνείας προβλέψεων για δίκτυα βαθιάς μάθησης και στοχεύει στην εκτίμηση της σημαντικότητας εισόδων στις προβλέψεις [146, 147]. Το DeepLIFT χρησιμοποιεί πολλαπλασιαστές που περιγράφουν τη μεταβολή των εξόδων όταν οι είσοδοι διαφέρουν από ένα σημείο αναφοράς. Η ιδιότητα «summation-to-delta» διατυπώνεται ως:

$$\sum_{i=1}^{n} C_{\Delta x_i, \Delta o} = \Delta o$$

όπου $o = f(x)$ η έξοδος του μοντέλου, $\Delta o = f(x) - f(r)$, $\Delta x_i = x_i - r_i$ και r το αναφορικό (reference) input. Το DeepLIFT ανήκει στις προσθετικές (additive) μεθόδους απόδοσης χαρακτηριστικών.

Layer-wise Relevance Propagation (LRP) Η μέθοδος LRP είναι μια τεχνική βασισμένη στην οπισθοδιάδοση για την κατανομή της «σχετικότητας» (relevance) της πρόβλεψης στα εισερχόμενα νευρώνια [15]. Όπως επισημαίνεται στο [147], το LRP είναι ισοδύναμο με το DeepLIFT στην περίπτωση όπου οι αναφορικές ενεργοποιήσεις όλων των νευρώνων τίθενται στο μηδέν. Στην πράξη, το LRP αποδίδει τιμές σχετικότητας πίσω από τα επίπεδα του δικτύου γεγονός που επιτρέπει την τοπική ερμηνεία της συμβολής κάθε εισόδου. Το LRP επίσης ανήκει στις προσθετικές μεθόδους απόδοσης χαρακτηριστικών.

Αξιολόγηση μεθόδων XAI

Η αξιολόγηση των μεθόδων XAI παραμένει μια πολύπλοκη αλλά ταυτόχρονα ουσιώδης διαδικασία. Μια αποτελεσματική αξιολόγηση οφείλει να ισορροπεί ανάμεσα στην πιστότητα της εξήγησης ως προς την πραγματική λογική του μοντέλου και στην ερμηνευσιμότητα και χρησιμότητά της για τον ανθρώπινο χρήστη [37, 65]. Ποσοτικά μέτρα, όπως η πιστότητα, η πληρότητα και η ανθεκτικότητα, προσφέρουν αντικειμενικές ενδείξεις για το πόσο καλά οι εξηγήσεις αντικατοπτρίζουν τη συμπεριφορά του μοντέλου [173], ενώ ποιοτικές αξιολογήσεις—συχνά μέσω μελετών χρηστών—εκτιμούν την εμπιστοσύνη, τη χρηστικότητα και τη γνωστική τους λογικότητα [111, 92]. Αυτή η διττή απαίτηση είναι ιδιαίτερα κρίσιμη στις εφαρμογές EEG, όπου οι εξηγήσεις πρέπει να είναι όχι μόνο τεχνικά ορθές αλλά και νευροεπιστημονικά ουσιαστικές [137]. Επομένως, είναι αναγκαία η ύπαρξη τεκμηριωμένων πλαισίων αξιολόγησης, ώστε να επιλέγονται οι μέθοδοι XAI που ενισχύουν τόσο τη διαφάνεια όσο και την πρακτική τους χρησιμότητα στο πεδίο της νευροαπεικόνισης.

1.2.3 Εξηγήσιμη Τεχνητή Νοημοσύνη και Ανάλυση Ηλεκτροεγκεφαλογραφήματος

Εφαρμογές EEG

Το ηλεκτροεγκεφαλογράφημα (EEG) έχει αναδειχθεί σε ένα ευέλικτο εργαλείο τόσο σε κλινικά όσο και σε ερευνητικά περιβάλλοντα, προσφέροντας πολύτιμες γνώσεις για τη δραστηριότητα του εγκεφάλου σε ένα ευρύ φάσμα εφαρμογών. Οι εφαρμογές αυτές εκτείνονται από την ιατρική διάγνωση έως την αλληλεπίδραση ανθρώπου-υπολογιστή, συμπεριλαμβανομένης της παρακολούθησης ύπνου για αξιολόγηση σταδίων και ανίχνευση διαταραχών [133, 1, 164], την ανίχνευση επιληπτικών κρίσεων [133], τα διεπαφή εγκεφάλου-υπολογιστή όπως ο P300 [78], την αναγνώριση συναισθημάτων [174, 68, 124, 30], την αποκατάσταση μετά από εγκεφαλικό επεισόδιο [143], την ανάλυση της σχιζοφρένειας [31], τη διαχείριση της επιληψίας [27], και την εκτίμηση της νοητικής κόπωσης [181].

Σύνολα Δεδομένων

Η ερευνητική κοινότητα έχει καθιερώσει αρκετά σύνολα δεδομένων αναφοράς για αυτές τις εφαρμογές, διευκολύνοντας την αναπαραγώγιμη έρευνα και ουσιαστικές συγκρίσεις απόδοσης. Τα έργα που συγκεντρώνονται στην παρούσα επισκόπηση αξιοποιούν ορισμένα βασικά σύνολα δεδομένων EEG, τα οποία καλύπτουν ποικίλες εφαρμογές. Μελέτες που αφορούν την παρακολούθηση ύπνου χρησιμοποιούν συχνά τα Sleep-EDF και SHHS για την ταξινόμηση σταδίων ύπνου και την ανίχνευση άπνοιας [60, 126]. Η αναγνώριση συναισθημάτων βασίζεται στα SEED και DEAP για την ανάλυση συναισθηματικών καταστάσεων μέσω προτύπων EEG [177, 77]. Σύνολα δεδομένων για τη μείζονα καταθλιπτική διαταραχή, όπως τα HUSM και MODMA, παρέχουν EEG δεδομένα για την ανάλυση διαταραχών διάθεσης [108, 26]. Η ανίχνευση κρίσεων αξιοποιεί τα CHB-MIT, TUH corpus και Bonn datasets για τον εντοπισμό επιληπτικών επεισοδίων και την ταξινόμηση κρίσεων [112, 10]. Η έρευνα κινητικής απεικόνισης χρησιμοποιεί τα σύνολα δεδομένων BCI Competition για την αποκωδικοποίηση νευρικών σημάτων σε διεπαφές εγκεφάλου-υπολογιστή [157]. Η πρόβλεψη εγκεφαλικού επεισοδίου υποστηρίζεται από το Acute dataset για μοντελοποίηση πρόγνωσης [8], ενώ οι μελέτες σχιζοφρένειας βασίζονται στα σύνολα δεδομένων UNM και IBIB PAN, τα οποία συνδυάζουν EEG και απεικονιστικές νευροεπιστημονικές μεθόδους [153, 113]. Ο Πίνακας 3.1 συνοψίζει τα σύνολα δεδομένων και τα φυσιολογικά τους σήματα.

Προκλήσεις

Η ανάλυση των συνόλων δεδομένων EEG έχει εξελιχθεί από τις παραδοσιακές προσεγγίσεις μηχανικής μάθησης (machine learning) σε προηγμένες αρχιτεκτονικές βαθιάς μάθησης (deep learning). Ενώ οι κλασικές μέθοδοι μηχανικής μάθησης [135] προσφέρουν ερμηνεύσιμες λύσεις μέσω χειροποίητων χαρακτηριστικών, οι σύγχρονες τεχνικές βαθιάς μάθησης έχουν επιδείξει ανώτερη απόδοση στην αυτόματη εξαγωγή χαρακτηριστικών και στην αναγνώριση προτύπων. Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks – CNNs)[162, 96] διαπρέπουν στην εξαγωγή χωρικών χαρακτηριστικών, ενώ τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks – RNNs)[123] και τα Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory – LSTM)[75] αποτυπώνουν αποτελεσματικά τις χρονικές εξαρτήσεις στα σήματα EEG. Οι πρόσφατες εξελίξεις περιλαμβάνουν Υβριδικές Αρχιτεκτονικές (hybrid architectures)[180] που συνδυάζουν πολλαπλές προσεγγίσεις, καθώς και Μοντέλα Τύπου Transformer (transformer-based models)[79] που αξιοποιούν μηχανισμούς αυτοπροσοχής (self-attention mechanisms) για τη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων. Τα Βασικά Μοντέλα (Foundation Models)[32] αποτελούν την πιο πρόσφατη εξέλιξη, με στόχο την παροχή μεταφερόμενων αναπαραστάσεων EEG σε πολλαπλές εργασίες.

Ωστόσο, παρά την εντυπωσιακή τους απόδοση, αυτά τα προηγμένα μοντέλα συχνά λειτουργούν ως «μαύρα κουτιά» (black boxes), γεγονός που εγείρει ανησυχίες σχετικά με την ερμηνευσιμότητα και τη διαφάνειά τους. Ο περιορισμός αυτός είναι ιδιαίτερα κρίσιμος στις ιατρικές εφαρμογές, όπου η κατανόηση της διαδικασίας λήψης αποφάσεων είναι ουσιαστική για την κλινική υιοθέτηση. Η πρόκληση αυτή αναδεικνύει τη διαρκώς αυξανόμενη σημασία των τεχνικών Επεξηγήσιμης Τεχνητής Νοημοσύνης (Explainable Artificial Intelligence – XAI) στην ανάλυση EEG.

Εφαρμογές XAI στην ανάλυση Ηλεκτροεγκεφαλογραφήματος

Η παρούσα εργασία αναλύει διάφορες μεθόδους Επεξηγήσιμης Τεχνητής Νοημοσύνης (Explainable Artificial Intelligence – XAI) που εφαρμόζονται στην ανάλυση ηλεκτροεγκεφαλογραφήματος, κατηγοριοποιώντας τις με

Table 1.1: Datasets used in EEG analysis with XAI insights.

Task	Dataset	Physiological Signal
Sleep Monitoring	Sleep-EDF [60, 73],	EEG, EMG, EOG
	Sleep Cassette [60, 73]	EEG, EOG, EMG
	SHHS [126, 60]	EEG, EMG, ECG, EOG
	CHAT [131, 98]	n/a
Emotion Recognition	SEED [177]	EEG
	DEAP [77]	EEG, EMG, EOG, BVP
	DENS [14]	EEG
Major Depressive Disorder	HUSM [108]	EEG
	MDD [107]	EEG
	MODMA [26, 145]	EEG
Seizure Detection	HUH [150]	EEG
	CHB-MIT [160]	EEG
	TUH corpus [112]	EEG
	Bonn [10]	EEG
	Siena [34]	EEG
	UBMC [167]	EEG
	UCI-EEG [3]	EEG
Motor Imagery	BCI IV 2a [157]	EEG, EOG
	BCI IV 2b [157]	EEG, EOG
	BCI III IVa [36]	EEG
	EEGMIMID [60]	EEG, EMG, EOG
	Stieger2021 [151]	EEG
Stroke prediction	Acute [8]	EEG
Schizophrenia	UNM [153]	FMRI, SMRI, EEG
	IBIB PAN [113]	EEG

βάση την προσέγγιση επεξήγησης, το εύρος, την εξάρτηση από το μοντέλο και τον τύπο επεξήγησης. Καλύπτουμε εκ των υστέρων μεθόδους (post-hoc methods), όπως η απόσταξη μοντέλου (model distillation) και η οπισθοδιάδοση (backpropagation), καθώς και μοντέλα ερμηνεύσιμα εκ σχεδιασμού (interpretable-by-design models). Διακρίνουμε μεταξύ ολικών (global) και τοπικών (local) επεξηγήσεων, ανεξάρτητων από το μοντέλο (model-agnostic) και εξειδικευμένων για συγκεκριμένο μοντέλο (model-specific) μεθόδων, και εξετάζουμε τύπους επεξήγησης που περιλαμβάνουν αποδόσεις χαρακτηριστικών (feature attribution), προσεγγίσεις βασισμένες σε κανόνες (rule-based approaches) και τεχνικές οπτικοποίησης (visualization techniques), όλα στο πλαίσιο εφαρμογών ηλεκτροεγκεφαλογραφήματος. Η ανάλυση μπορεί να διαβαστεί στο υποκεφάλαιο 3.3.4

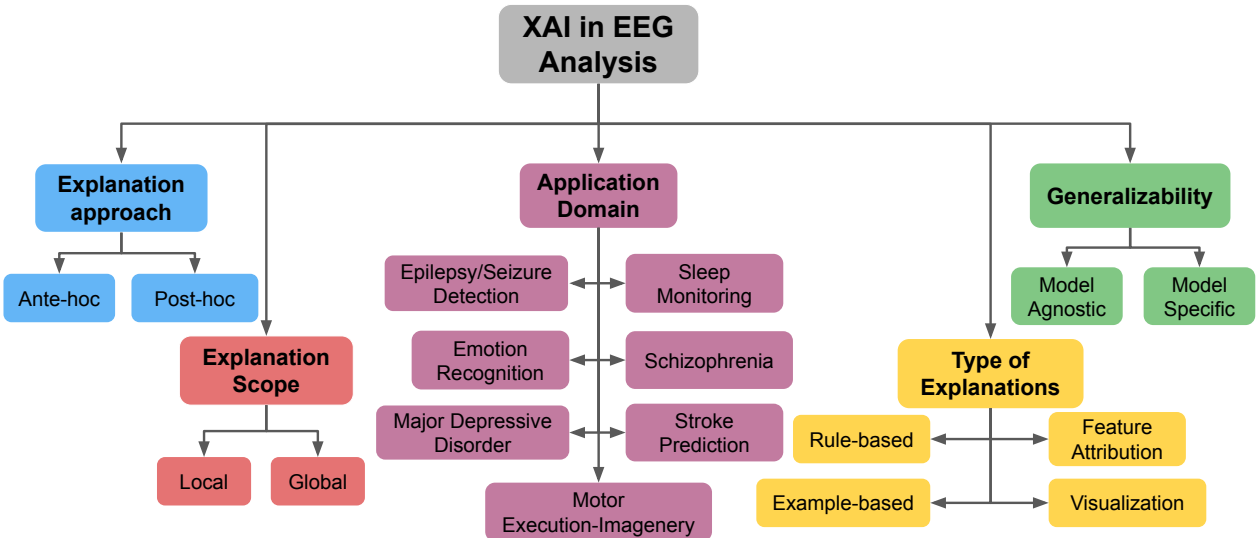

Figure 1.2.1: Ταξινόμηση προσεγγίσεων XAI στην ανάλυση EEG.

1.3 Προτεινόμενη Μεθοδολογία

Σε αυτήν την ενότητα, παρουσιάζουμε τις μεθοδολογίες που ακολουθήθηκαν τόσο για την επισκόπηση των μεθόδων XAI στο EEG όσο και για το πρόβλημα της ανίχνευσης κρίσεων. Τα αποτελέσματα της επισκόπησης παρουσιάστηκαν στην 3.3.4. Για το πρόβλημα της ανίχνευσης κρίσεων, τα μοντέλα εκπαιδεύονται με χαρακτηριστικά που εξάγονται από σήματα EEG, ECG και EMG. Στη συνέχεια, εφαρμόζουμε δύο διαφορετικές *post-hoc* μεθόδους επεξηγησιμότητας για να κατανοήσουμε τις αποφάσεις των μοντέλων.

Αρχικά, επισημαίνουμε τις κύριες συνεισφορές της παρούσας διατριβής και κατόπιν εξηγούμε αναλυτικά τις μεθοδολογίες που ακολουθήθηκαν.

1.3.1 Συνεισφορά

Οι συνεισφορές αυτής της διατριβής είναι πολλαπλές και μπορούν να συνοψιστούν ως εξής:

- Εξετάζουμε όλες τις σχετικές τεχνικές XAI στην ανάλυση EEG· έτσι, η μελέτη μας προσφέρει στους ερευνητές μια σαφή εικόνα της τρέχουσας κατάστασης του πεδίου και εντοπίζει πιθανά ερευνητικά κενά.
- Παρουσιάζεται μια ολοκληρωμένη ανάλυση των πρόσφατων τάσεων και εξελίξεων στην XAI για την ανάλυση EEG. Η μελέτη παρέχει μια επισκόπηση των θεμελιωδών εργασιών EEG, καταγράφει τα διαθέσιμα σύνολα δεδομένων και προτείνει μια δομημένη ταξινόμηση των μεθόδων XAI.
- Παρουσιάζουμε μια ερμηνεύσιμη μεθοδολογία για πολυτροπική ανίχνευση κρίσεων χρησιμοποιώντας σήματα EEG, ECG και EMG. Η προτεινόμενη προσέγγιση αξιοποιεί τεχνικές προεπεξεργασίας και εξαγωγής χαρακτηριστικών και επιδεικνύει ανώτερη απόδοση σε σύγκριση με τις συμβατικές μεθόδους βαθιάς μάθησης.
- Με βάση την ανάλυση επεξηγησιμότητας, αναδεικνύουμε τα πιο σημαντικά χαρακτηριστικά που συμβάλλουν στην ανίχνευση κρίσεων.

1.3.2 Ακολουθούμενες Μεθοδολογίες

Επισκόπηση μεθόδων XAI στο EEG

Σε αυτήν την ενότητα παρουσιάζουμε τη μεθοδολογία που χρησιμοποιήθηκε για την εκπόνηση μιας ολοκληρωμένης και συστηματικής επισκόπησης της βιβλιογραφίας σχετικά με την εφαρμογή του XAI στην ανάλυση EEG. Ξεκινάμε με τα ερευνητικά ερωτήματα που καθοδήγησαν την έρευνά μας, έπειτα περιγράφουμε τη στρατηγική επιλογής των εργασιών και ολοκληρώνουμε με το πλαίσιο που χρησιμοποιήθηκε για την ανάλυση των συγκεντρωμένων μελετών.

Ερευνητικά Ερωτήματα Η παρούσα μελέτη επικεντρώνεται στη διερεύνηση του ρόλου του XAI στην ανάλυση EEG, απαντώντας σε βασικά ερευνητικά ερωτήματα που εξετάζουν τις μεθόδους, τις εφαρμογές και τις ευρύτερες επιπτώσεις του XAI σε αυτόν τον τομέα. Τα παρακάτω ερευνητικά ερωτήματα καθοδηγούν την επισκόπηση:

EE1: Ποιες είναι οι βασικές μέθοδοι XAI που χρησιμοποιούνται στην ανάλυση EEG; Το ερώτημα αυτό αποσκοπεί στον εντοπισμό και την κατηγοριοποίηση των τεχνικών ερμηνευσιμότητας που εφαρμόζονται σε μελέτες με βάση το EEG. Μέσα από την εξέταση αυτών των μεθόδων, η μελέτη επιδιώκει να προσφέρει μια ολοκληρωμένη επισκόπηση του τρόπου με τον οποίο ενσωματώνεται η ερμηνευσιμότητα στην έρευνα EEG και ποιες προσεγγίσεις είναι οι πιο διαδεδομένες.

EE2: Σε ποιες συγκεκριμένες εργασίες ανάλυσης EEG έχουν εφαρμοστεί μέθοδοι XAI; Η ανάλυση EEG περιλαμβάνει ένα ευρύ φάσμα εφαρμογών, όπως ανίχνευση επιληπτικών κρίσεων, ταξινόμηση σταδίων ύπνου, εκτίμηση γνωστικού φόρτου και αναγνώριση συναισθημάτων. Το ερώτημα αυτό διερευνά τις διάφορες εργασίες στις οποίες έχουν εφαρμοστεί τεχνικές XAI, αναδεικνύοντας τάσεις, προκλήσεις και κενά στη διαθέσιμη βιβλιογραφία.

Μέσα από τις απαντήσεις στα παραπάνω ερευνητικά ερωτήματα και τη συνολική μας ανάλυση, επιδιώκουμε επίσης να παρουσιάσουμε τι είδους περιορισμοί υπάρχουν στην εφαρμογή τεχνικών XAI στον τομέα του EEG και ποιες περαιτέρω προσεγγίσεις θα μπορούσαν να προσφέρουν οφέλη για μελλοντική έρευνα.

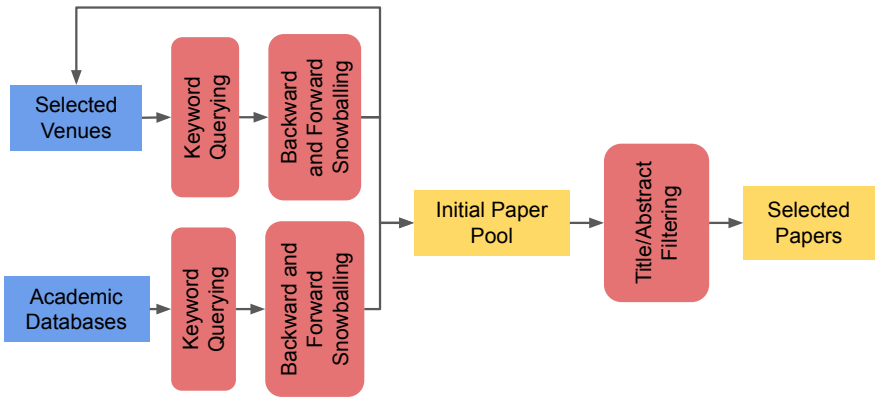

Figure 1.3.1: Μεθοδολογία για τον εντοπισμό άρθρων.

Στρατηγική Αναζήτησης Βιβλιογραφίας Η στρατηγική αναζήτησής μας ακολούθησε δύο παράλληλες προσεγγίσεις για τη δημιουργία μιας αρχικής συλλογής άρθρων, η οποία στη συνέχεια φιλτραρίστηκε με βάση τη συνάφεια τίτλου και περίληψης ώστε να προκύψει το τελικό σύνολο που μελετήθηκε σε αυτή την επισκόπηση, όπως απεικονίζεται στο Σχήμα 1.3.1. Η συλλογή της αρχικής ομάδας άρθρων περιλάμβανε (1) στοχευμένες αναζητήσεις σε επιλεγμένα συνέδρια/περιοδικά και (2) γενικές αναζητήσεις με λέξεις-κλειδιά σε μεγάλες ακαδημαϊκές βάσεις δεδομένων, σε συνδυασμό με backward και forward snowballing για τον εντοπισμό σχετικών εργασιών. Η αρχική συλλογή των 151 άρθρων υπέστη διαδικασία φιλτραρίσματος με βάση τίτλους και περιλήψεις, αποδίδοντας τελικά 66 σχετικά άρθρα.

Ξεκινήσαμε με αναζήτηση στα πρακτικά του World Conference on Explainable Artificial Intelligence με τους όρους "EEG" και "electroencephalography" και χρησιμοποιήσαμε backward και forward snowballing για να επεκτείνουμε το αρχικό μας σύνολο άρθρων. Αυτή η διαδικασία αποκάλυψε επιπλέον βασικούς επιστημονικούς χώρους —όπως IEEE ISBI, IEEE EMBC, IEEE BIBM και IEEE NER— οι οποίοι δημοσιεύουν εργασίες στη διασταύρωση XAI και EEG. Για κάθε χώρο, πραγματοποιήσαμε συστηματική αναζήτηση δημοσιεύσεων χρησιμοποιώντας λέξεις-κλειδιά σχετικές με XAI και EEG και εφαρμόσαμε εκ νέου snowballing για τη διεύρυνση του συνόλου μας.

Η δεύτερη προσέγγιση περιλάμβανε ερωτήματα σε μεγάλες ακαδημαϊκές βάσεις δεδομένων, όπως PubMed, ScienceDirect, IEEE Xplore, Springer και Google Scholar. Χρησιμοποιήσαμε το ίδιο σύνολο λέξεων-κλειδιών σχετικών με XAI και EEG για την ανάκτηση σχετικών δημοσιεύσεων. Όπως και στην πρώτη προσέγγιση, εφαρμόσαμε snowballing για τον εντοπισμό πρόσθετων άρθρων από τις βιβλιογραφικές αναφορές και τις παραπομπές των ανακτημένων εργασιών.

Μετά τη συγκέντρωση της αρχικής συλλογής των 151 άρθρων, πραγματοποιήσαμε διαδικασία φιλτραρίσματος βάσει τίτλου και περίληψης για να βελτιστοποιήσουμε την επιλογή. Δώσαμε προτεραιότητα σε μελέτες που ευθυγραμμίζονταν άμεσα με το ερευνητικό μας αντικείμενο, εξασφαλίζοντας ευρεία αλλά σχετική κάλυψη της διασταύρωσης XAI και EEG. Επιπλέον, προτιμήθηκαν άρθρα που παρείχαν σαφή πειραματική τεκμηρίωση ή καινοτόμες μεθοδολογικές συνεισφορές, βοηθώντας μας να εστιάσουμε σε υψηλής ποιότητας και ουσιαστική έρευνα.

Ανίχνευση κρίσεων

Η μεθοδολογία ξεκινά με το σύνολο δεδομένων SeizeIT2 [20] ως είσοδο. Το συγκεκριμένο σύνολο δεδομένων επιλέχθηκε, καθώς αποτέλεσε τη βάση για το Una Europa Epilepsy Data Challenge, προσφέροντας ένα καθιερωμένο σημείο αναφοράς για την έρευνα στον τομέα της ανίχνευσης κρίσεων. Επιπλέον, πρόκειται για ένα μεγάλης κλίμακας σύνολο δεδομένων που περιλαμβάνει χιλιάδες ώρες πολυτροπικών καταγραφών, όχι μόνο EEG αλλά και EMG και ECG σημάτων. Η κλίμακα και η πολυτροπική του φύση το καθιστούν ιδιαίτερα κατάλληλο για την ανάπτυξη ανθεκτικών και γενικεύσιμων μοντέλων.

Το επόμενο στάδιο της διαδικασίας αφορά την προεπεξεργασία και την εξαγωγή χαρακτηριστικών, τα οποία προετοιμάζουν τα δεδομένα για την εκπαίδευση των μοντέλων. Το στάδιο αυτό είναι κρίσιμο, καθώς τα ακατέργαστα σήματα είναι συχνά θορυβώδη και, χωρίς επαρκή επεξεργασία, μπορεί να αποκρύψουν αντί να αναδείξουν σημαντικά πρότυπα. Κατά την προεπεξεργασία, οι συνεχείς καταγραφές τμηματοποιούνται σε παράθυρα στα-

θερού μήκους και διέρχονται από φιλτράρισμα για την απομάκρυνση θορύβου και τεχνητών παραμορφώσεων, εξασφαλίζοντας ότι η επακόλουθη ανάλυση επικεντρώνεται σε φυσιολογικά σημαντικά συστατικά του σήματος. Στη συνέχεια, εφαρμόζεται εξαγωγή χαρακτηριστικών και από τις τρεις κατηγορίες σημάτων (EEG, ECG και EMG), οδηγώντας σε έναν πλούσιο πολυδιάστατο πίνακα δεδομένων. Τα εξαγόμενα χαρακτηριστικά περιλαμβάνουν στατιστικά μέτρα, χρονικές ιδιότητες, μετρήσεις πολυπλοκότητας και φασματικά χαρακτηριστικά που προκύπτουν από διαφορετικές συχνοτικές ζώνες.

Αφού προετοιμαστούν τα δεδομένα, τα εξαγόμενα χαρακτηριστικά χρησιμοποιούνται για την εκπαίδευση και αξιολόγηση μοντέλων μηχανικής μάθησης. Διερευνάται μια ποικιλία ταξινομητών, με σκοπό τον εντοπισμό εκείνων που επιτυγχάνουν την καλύτερη ισορροπία μεταξύ προγνωστικής απόδοσης και ερμηνευσιμότητας. Ειδικότερα, δίνεται έμφαση σε μεθόδους βασισμένες σε δέντρα και σε μεθόδους συνόλων (ensemble), όπως Random Forest, XGBoost, LightGBM, CatBoost και HistGradientBoosting. Τα μοντέλα αυτά είναι ιδιαίτερα κατάλληλα για πολυτροπικά δεδομένα σε μορφή πινάκων και προσφέρουν το πλεονέκτημα της διαχείρισης μη γραμμικών αλληλεπιδράσεων και ετερογενών χαρακτηριστικών. Επιπλέον, σε σύγκριση με «μαύρα κουτιά», προσφέρονται περισσότερο για την εφαρμογή τεχνικών ερμηνευσιμότητας, γεγονός που τα καθιστά ιδανική επιλογή για κλινικά σενάρια όπου αδιαφανείς προβλέψεις δεν θα ήταν αποδεκτές. Η έμφαση δίνεται όχι μόνο στην επίτευξη υψηλής ακρίβειας ανίχνευσης, αλλά και στην ανάπτυξη μοντέλων που θα μπορούσαν να εμπιστευθούν και να ελέγξουν οι κλινικοί γιατροί στην πράξη.

Τέλος, η ροή εργασίας ολοκληρώνεται με την ανάλυση ερμηνευσιμότητας, όπου εφαρμόζονται δύο συμπληρωματικές εκ των υστέρων (post-hoc) τεχνικές για την ερμηνεία των εκπαιδευμένων μοντέλων. Αρχικά, χρησιμοποιούμε το SHAP σε όλα τα μοντέλα, δημιουργώντας SHAP beeswarm plots που παρέχουν μια συνολική εικόνα της σημασίας των χαρακτηριστικών, καθώς και την κατεύθυνση και το μέγεθος της συμβολής κάθε χαρακτηριστικού στις προβλέψεις. Με αυτόν τον τρόπο, εντοπίζουμε όχι μόνο ποια χαρακτηριστικά είναι τα πιο καθοριστικά, αλλά και πώς οι τιμές τους επηρεάζουν την πιθανότητα ανίχνευσης κρίσης. Επιπλέον, εφαρμόζουμε το TE2Rules στα μοντέλα XGBoost και Random Forest, εξάγοντας κανόνες ευανάγνωστους από τον άνθρωπο από αυτούς τους ταξινομητές.

Συνοψίζοντας, η μεθοδολογία έχει σχεδιαστεί ώστε όχι μόνο να κατασκευάζει ακριβή μοντέλα ανίχνευσης κρίσεων, αλλά και να δίνει έμφαση στη διαφάνεια και την ερμηνευσιμότητα σε κάθε στάδιο της ροής εργασίας. Με τον συνδυασμό δεδομένων αναφοράς, πολυτροπικών χαρακτηριστικών, ταξινομητών βασισμένων σε σύνολα και εργαλείων εκ των υστέρων ερμηνευσιμότητας, η προσέγγιση εξισορροπεί την προγνωστική ισχύ με την κλινική σημασία. Μια λεπτομερής περιγραφή κάθε σταδίου της μεθοδολογίας, μαζί με τις πειραματικές ρυθμίσεις και τα αποτελέσματα, παρέχεται στο Κεφάλαιο 1.4.

1.4 Πειράματικό Μέρος

1.4.1 Σύνολο Δεδομένων

Σετ Δεδομένων SeizeIT2

Τα μοντέλα που παρουσιάζονται σε αυτή τη διατριβή εκπαιδεύτηκαν και αξιολογήθηκαν χρησιμοποιώντας το σετ δεδομένων SeizeIT2 [20]. Το σετ δεδομένων περιλαμβάνει καταγραφές από 125 ασθενείς, συνολικής διάρκειας περίπου 11.640 ωρών φορητών δεδομένων, που συλλέχθηκαν σε πέντε διαφορετικές Ευρωπαϊκές Μονάδες Παρακολούθησης Επιληψίας. Για τους περισσότερους συμμετέχοντες καταγράφηκαν τέσσερις διαφορετικές μορφές σήματος: bte-EEG, ECG, EMG και δεδομένα κίνησης. Όλα τα δεδομένα των συμμετεχόντων περιλαμβάνουν φορητό bte-EEG. Σε ποσοστό 3% του συνόλου, τα δεδομένα ECG, EMG και κίνησης δεν καταγράφηκαν λόγω τεχνικών σφαλμάτων ή προβλημάτων στη ρύθμιση.

Το σετ δεδομένων είναι ανοικτό και χρησιμοποιήθηκε στον διαγωνισμό Seizure Detection Challenge, ο οποίος οργανώθηκε από το KU Leuven σε συνεργασία με την Una Europa, με στόχο την ανάπτυξη καινοτόμων και ανθεκτικών πλαισίων μηχανικής μάθησης (ML) για την επεξεργασία δεδομένων EEG, με τελικό σκοπό την ανίχνευση επιληπτικών κρίσεων. Σύμφωνα με τις οδηγίες των διοργανωτών, οι καταγραφές των πρώτων 96 συμμετεχόντων χρησιμοποιήθηκαν για εκπαίδευση, ενώ οι υπόλοιποι για αξιολόγηση. Η τελική αξιολόγηση πραγματοποιήθηκε σε ένα κρυφό σύνολο δοκιμών.

Για την εφαρμογή αυτή, δεν ήταν εφικτό να πραγματοποιηθεί εκπαίδευση σε όλες τις χιλιάδες ώρες καταγραφών· συνεπώς, χρησιμοποιήθηκε μόνο ένα υποσύνολο για εκπαίδευση. Επιπλέον, το σετ δεδομένων είναι έντονα μη

ισορροπημένο, καθώς περιλαμβάνει λίγες ώρες επιληπτικών κρίσεων σε σύγκριση με μεγάλες περιόδους χωρίς κρίσεις. Τα κατασκευασμένα υποσύνολα δεδομένων περιλάμβαναν όλες τις περιόδους που είχαν επισημανθεί ως επεισόδια κρίσεων, καθώς και τυχαία επιλεγμένες περιόδους χωρίς κρίσεις. Εξετάστηκαν τρεις λόγοι κρίσεων προς μη-κρίσεις: 1:2, 1:10 και 1:100. Για την αξιολόγηση χρησιμοποιήθηκαν όλες οι καταγραφές που περιλάμβαναν κάθε μορφή σήματος.

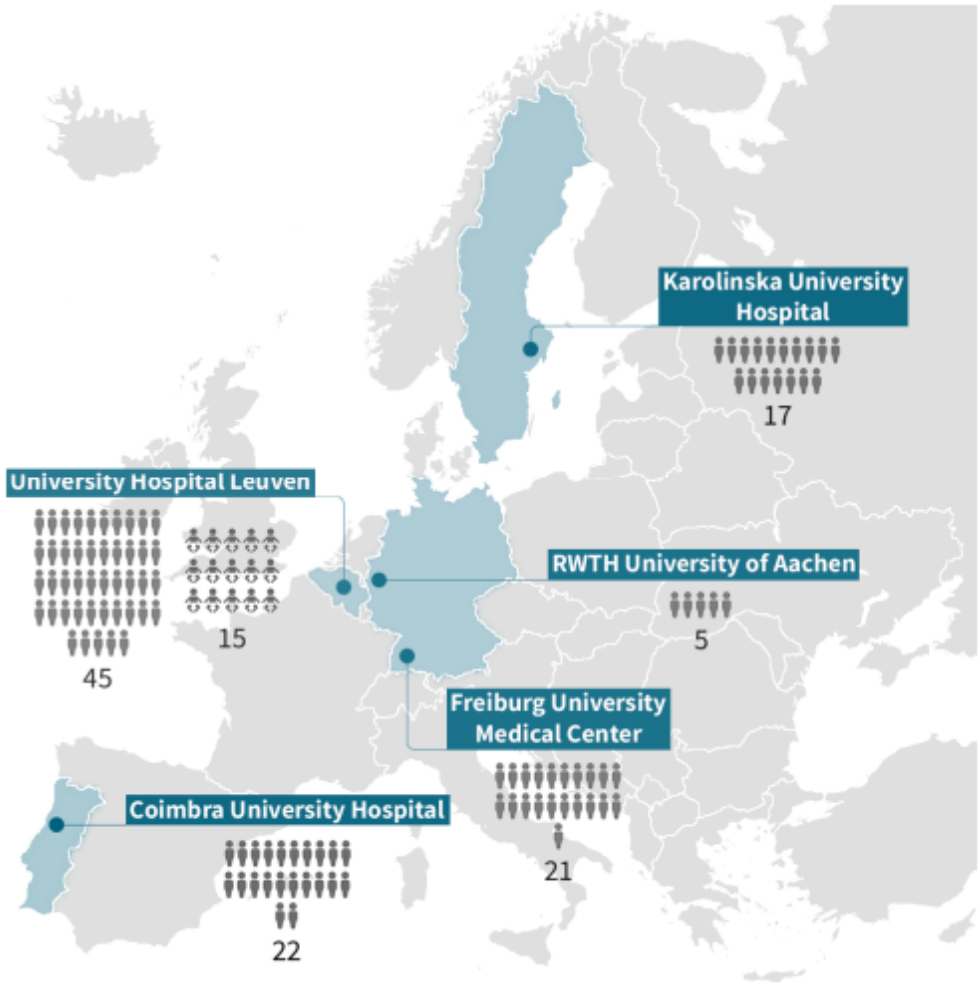

Figure 1.4.1: Αριθμός συμμετεχόντων ανά EMU που συμπεριλαμβάνονται στο SeizeIT2

Εμπλουτισμός Δεδομένων (Data Augmentation)

Μία συχνή πρόκληση στην ανάλυση βιοϊατρικών σημάτων, όπως τα EEG για την ανίχνευση κρίσεων, είναι η περιορισμένη ποιότητα και ποσότητα των διαθέσιμων δεδομένων [67]. Στο σετ δεδομένων μας, οι καταγραφές προέρχονται από φορητές συσκευές, οι οποίες παρέχουν χαμηλότερης ποιότητας δεδομένα, ενώ η συνολική διάρκεια των επεισοδίων κρίσεων είναι σχετικά περιορισμένη. Στη διατριβή αυτή, πειραματίστηκα με σύνολα δεδομένων τόσο με όσο και χωρίς εμπλουτισμένα δεδομένα κρίσεων.

Η τεχνική εμπλουτισμού που χρησιμοποιήθηκε ήταν η ίδια με εκείνη που παρουσιάζεται στο [67]. Συγκεκριμένα, εφαρμόστηκαν υποκατάστατα μέσω Μετασχηματισμού Fourier (FT Surrogates), μία μαθηματική μέθοδος που μετασχηματίζει μια χρονική συνάρτηση, όπως ένα σήμα EEG, σε συνάρτηση συχνότητας. Η μετασχηματισμένη αναπαράσταση (ή φάσμα) προσφέρει μια εναλλακτική θεώρηση των δεδομένων και αναδεικνύει διαφορετικά χαρακτηριστικά του υποκείμενου σήματος.

Τα FT surrogates αποτελούν μια ιδιαίτερη μορφή εμπλουτισμού δεδομένων, όπου παράγονται νέα υποκατάστατα σήματα μέσω τυχαίας αναδιάταξης των φάσεων του Μετασχηματισμού Fourier του αρχικού σήματος EEG. Σημαντικό είναι ότι η διαδικασία αυτή διατηρεί το φάσμα ισχύος, δηλαδή την κατανομή της ενέργειας του

σήματος στις διάφορες συχνότητες, διαφυλάσσοντας τις συνολικές δομικές ιδιότητες του αρχικού σήματος, ενώ μεταβάλλει την χρονική οργάνωση (π.χ. ακολουθία και χρονισμός γεγονότων) [67].

Η χρήση των FT surrogates εξυπηρέτησε δύο βασικούς σκοπούς. Αρχικά, αύξησε το μέγεθος των συνόλων εκπαίδευσης, δημιουργώντας περισσότερα δείγματα κρίσεων. Ταυτόχρονα, προσέθεσε ποικιλία στα σύνολα δεδομένων, επιτρέποντας στα μοντέλα να μάθουν να αναγνωρίζουν κρίσεις σε ευρύτερο φάσμα συνθηκών.

Προεπεξεργασία και Εξαγωγή Χαρακτηριστικών

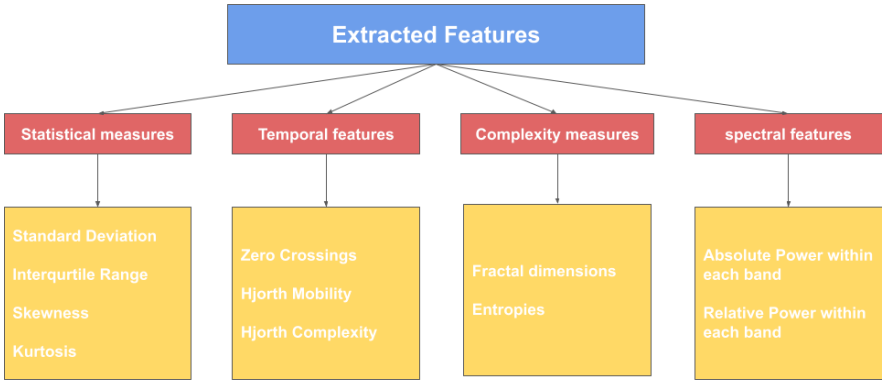

Figure 1.4.2: Χαρακτηριστικά που εξήχθησαν από το EEG σήμα

Οι προσεγγίσεις προεπεξεργασίας και εξαγωγής χαρακτηριστικών που χρησιμοποιήθηκαν στη διατριβή αυτή βασίστηκαν σε μελέτες για την ανίχνευση κρίσεων από φορητά δεδομένα [67] και για την πρόβλεψη υποτύπων μετατραυματικής διαταραχής άγχους από EEG σε κατάσταση ηρεμίας [85].

Αρχικά, τα δεδομένα EEG υποβλήθηκαν σε ζωνοπερατό φιλτράρισμα (band-pass) μεταξύ 1–100 Hz, ώστε να διατηρηθούν οι συχνοτικές συνιστώσες που σχετίζονται περισσότερο με τη δραστηριότητα των κρίσεων, ενώ ταυτόχρονα αφαιρέθηκαν αργές παρεκκλίσεις και υψηλής συχνότητας θόρυβος. Εφαρμόστηκε επίσης notch φίλτρο στα 50 Hz για την εξάλειψη των παρεμβολών από το ηλεκτρικό δίκτυο, οι οποίες είναι συχνές σε κλινικές εγγραφές. Μετά το φιλτράρισμα, τα σήματα χωρίστηκαν σε μη επικαλυπτόμενα χρονικά παράθυρα διάρκειας 1 δευτερολέπτου (epochs), παρέχοντας συνεπή χρονικά διαστήματα για τον υπολογισμό χαρακτηριστικών. Λήφθηκαν υπόψη τρεις αναλογίες κρίσεων προς μη-κρίσεις (1:2, 1:10 και 1:100) για την αντιμετώπιση της εγγενούς ανισορροπίας των κλάσεων, και για κάθε αναλογία, διεξήχθησαν πειράματα τόσο με όσο και χωρίς αυξημένα δείγματα κρίσεων, ώστε να αξιολογηθεί η επίδραση της επέκτασης δεδομένων (data augmentation) στην απόδοση των μοντέλων.

Μετά την προεπεξεργασία, εξαχθήκαν χαρακτηριστικά από όλες τις μορφές δεδομένων (EEG, ECG και EMG), δημιουργώντας μια πλούσια αναπαράσταση των δεδομένων για τα μοντέλα μηχανικής μάθησης. Τα χαρακτηριστικά κατηγοριοποιήθηκαν σε τέσσερις βασικές ομάδες: στατιστικά μέτρα, χρονικά χαρακτηριστικά, μέτρα πολυπλοκότητας και φασματικά χαρακτηριστικά. Από κοινού, αυτά τα χαρακτηριστικά καταγράφουν συμπληρωματικές πληροφορίες σχετικά με τη διανομή των σημάτων, τη δυναμική τους, τη μη γραμμικότητα και το φασματικό περιεχόμενο. Ένα σύνοψη των εξαγόμενων χαρακτηριστικών παρουσιάζεται στο Σχήμα 1.4.2, ενώ λεπτομερείς περιγραφές δίνονται στις επόμενες παραγράφους.

Τα στατιστικά μέτρα που χρησιμοποιήθηκαν ήταν η τυπική απόκλιση (STD), το ενδοτεταρτημοριακό εύρος (IQR), η ασυμμετρία (Skewness) και η κύρτωση (Kurtosis). Η STD και το IQR χρησιμοποιούνται για την καταγραφή της μεταβλητότητας μέσα στα σήματα EEG. Η STD αντικατοπτρίζει την μέση απόκλιση των δεδομένων από τον μέσο όρο, ενώ το IQR δείχνει την εξάπλωση του μεσαίου τμήματος των δεδομένων. Η ασυμμετρία και η κύρτωση περιγράφουν τα χαρακτηριστικά της κατανομής πιθανότητας του σήματος: η ασυμμετρία μετράει την ασυμμετρία της κατανομής, ενώ η κύρτωση αξιολογεί το βάρος των ουρών της κατανομής σε σχέση με την κανονική κατανομή.

Τα χρονικά χαρακτηριστικά που αναλύθηκαν περιλαμβάνουν τον αριθμό διασχίσεων μηδενός (zero crossings), την κινητικότητα Hjorth (Hjorth mobility) και την πολυπλοκότητα Hjorth (Hjorth complexity). Οι διασχίσεις μηδενός παρέχουν εκτίμηση του φασματικού περιεχομένου του σήματος, ενώ οι παράμετροι Hjorth λειτουργούν ως περιγραφικά μέτρα των χαρακτηριστικών του σήματος. Ειδικότερα, η κινητικότητα αντικατοπτρίζει τη μέση συχνότητα ή τον ρυθμό μεταβολής του σήματος, ενώ η πολυπλοκότητα δείχνει πόσο κοντά μοιάζει το σήμα με μια καθαρή ημιτονοειδή κυματομορφή.

Τα μέτρα πολυπλοκότητας περιλάμβαναν φρακταλικές διαστάσεις (fractal dimensions) και εντροπίες (entropies). Οι φρακταλικές διαστάσεις καταγράφουν πώς το επίπεδο λεπτομέρειας των δεδομένων μεταβάλλεται σε διαφορετικές κλίμακες, ενώ η εντροπία ποσοτικοποιεί τον βαθμό τυχαιότητας ή απρόβλεπτης συμπεριφοράς του σήματος.

Τα φασματικά χαρακτηριστικά, όπως η ισχύς (power) σε διάφορες ενεργειακές ζώνες (Delta, Theta, Alpha, Beta και Gamma) εξήχθησαν επίσης. Το σήμα EEG χωρίστηκε σε διαφορετικές συχνοτικές ζώνες και μετρήθηκε η απόλυτη και σχετική ισχύς σε κάθε ζώνη. Οι πέντε κανονικές συχνοτικές ζώνες ορίζονται ως εξής: delta (1.25–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz) και gamma (30–49 Hz). Επίσης, υπολογίστηκε και η αναλογία theta/beta.

Τα σήματα EMG φιλτραρίστηκαν με ζωνοπερατό φίλτρο 20–450 Hz για την απομάκρυνση τεχνητών κινήσεων και θορύβου υψηλής συχνότητας, και εφαρμόστηκε notch φίλτρο στα 50 Hz. Στη συνέχεια, χωρίστηκαν σε εποχές διάρκειας 1 δευτερολέπτο χωρίς επικάλυψη για εξαγωγή χαρακτηριστικών. Υπολογίστηκαν δείκτες όπως τετράγωνο μέσης ρίζας (RMS), Μέση απόλυτη τιμή (MAV), Ρυθμός Διασχίσεων του Μηδενός (ZC), μήκος κυματομορφής (WL), καθώς και στατιστικά μέτρα (STD, Διακύμανση).

Τα σήματα ECG προεπεξεργάστηκαν με ζωνοπερατό φίλτρο 0.5–50 Hz για απομάκρυνση βασικής μετατόπισης και θορύβου υψηλής συχνότητας, και notch φίλτρο στα 50 Hz. Διαχωρίστηκαν σε εποχές 1 δευτ., από τις οποίες εξήχθησαν στατιστικά και φυσιολογικά χαρακτηριστικά. Εκτός από βασικά στατιστικά (μέση τιμή, STD, RMS, Διακύμανση, εύρος κορυφής-κορυφής), εφαρμόστηκε ανίχνευση R-κορυφών, με την οποία εκτιμήθηκαν τα διαστήματα RR και ο καρδιακός ρυθμός. Υπολογίστηκαν επίσης μέσο RR, τυπική απόκλιση RR και στιγμιαίος καρδιακός ρυθμός, παρέχοντας πληροφορίες για τη βραχυπρόθεσμη δυναμική του καρδιακού ρυθμού.

1.4.2 Μετρικές

Η αξιολόγηση των μοντέλων μας βασίστηκε στη συνάρτηση βαθμολόγησης που χρησιμοποιήθηκε επίσης στο Seizure Detection Challenge. Η συνάρτηση αυτή βασίζεται στις μετρικές της ευαισθησίας (sensitivity) και του Λόγου Ψευδών Συναγερμών (False Alarm Ratio - FAR). Οι μετρικές αυτές είναι ιδιαίτερα κατάλληλες για εργασίες ανίχνευσης επιληπτικών κρίσεων, οι οποίες αποτελούν έντονα μη ισορροπημένα προβλήματα ταξινόμησης λόγω της σπανιότητας των επεισοδίων κρίσεων σε σύγκριση με τις μεγάλες χρονικές περιόδους χωρίς κρίσεις.

Ευαισθησία

Η ευαισθησία μετράται σε επίπεδο γεγονότων, πράγμα που σημαίνει ότι η απόδοση αξιολογείται στο επίπεδο ολόκληρων επεισοδίων κρίσεων και όχι σε επιμέρους παράθυρα πρόβλεψης. Με αυτόν τον τρόπο διασφαλίζεται ότι η κλινική σημασία της ανίχνευσης κρίσεων αντικατοπτρίζεται στην αξιολόγηση, καθώς η απώλεια ενός επεισοδίου είναι πιο κρίσιμη από την απώλεια ενός μόνο παραθύρου. Για τον υπολογισμό της ευαισθησίας εφαρμόζεται η μέθοδος any-overlap (OVLP) [179]. Σύμφωνα με την OVLP, ένα True Positive (TP) καταγράφεται όταν η προβλεπόμενη υπόθεση έχει οποιαδήποτε χρονική επικάλυψη με το αντίστοιχο γεγονός κρίσης στην επισημείωση αναφοράς. Ένα False Negative (FN) συμβαίνει όταν δεν υπάρχει καμία επικάλυψη. Η μέθοδος αυτή είναι επιεικής, καθώς ακόμη και μερική επικάλυψη υπολογίζεται ως επιτυχής ανίχνευση, γεγονός που οδηγεί συνήθως σε υψηλότερες τιμές ευαισθησίας αλλά υποεκτιμά τον αριθμό των ψευδών ανιχνεύσεων.

Λόγος Ψευδών Συναγερμών (FAR)

Οι ψευδείς συναγερμοί (FAs) αντιστοιχούν σε εσφαλμένες ανιχνεύσεις όπου το μοντέλο προβλέπει ένα επεισόδιο κρίσης το οποίο δεν επικαλύπτεται με καμία επισημασμένη κρίση στην αναφορά. Αντί να αναφέρεται η ειδικότητα (specificity), η οποία είναι λιγότερο ενημερωτική σε εξαιρετικά μη ισορροπημένα σύνολα δεδομένων, χρησιμοποιείται ο Λόγος Ψευδών Συναγερμών (FAR). Το FAR υπολογίζεται ως ο αριθμός των ψευδώς θετικών ανιχνεύσεων κανονικοποιημένος ως προς τη διάρκεια της καταγραφής και εκφράζεται ως αριθμός ψευδών συναγερμών ανά ώρα (FA/h). Στην αξιολόγησή μας, οι FAs υπολογίστηκαν με τη μέθοδο βαθμολόγησης που

βασίζεται σε εποχές (epoch-based, EPOCH) [179]. Σε αυτήν τη μέθοδο, τόσο η αναφορά όσο και η υπόθεση διακριτοποιούνται σε μη επικαλυπτόμενες εποχές σταθερής διάρκειας. Κάθε εποχή χαρακτηρίζεται ως κρίση/μη κρίση, και τα σφάλματα καταγράφονται ως εισαγωγές, διαγραφές ή αντικαταστάσεις, με όλα τα σφάλματα να έχουν ίσο βάρος. Αυτό παρέχει μια πιο συντηρητική εκτίμηση των ψευδών συναγερμών σε σύγκριση με την OVLP, μειώνοντας τον κίνδυνο υποεκτίμησης των εσφαλμένων ανιχνεύσεων.

Συνάρτηση Βαθμολόγησης

Για να επιτευχθεί ισορροπία ανάμεσα στην υψηλή ευαισθησία και στον χαμηλό ρυθμό ψευδών συναγερμών, υιοθετήσαμε τη συνδυαστική συνάρτηση βαθμολόγησης που ορίστηκε στον διαγωνισμό. Η ευαισθησία υπολογίζεται με τη μέθοδο OVLP, ενώ το FAR εκτιμάται με την εποχική προσέγγιση (EPOCH). Ο τελικός βαθμός υπολογίζεται ως ένας σταθμισμένος συνδυασμός των δύο μετρικών, με έναν συντελεστή βαρύτητας 0.4 να εφαρμόζεται στο FAR ώστε να εξισορροπηθεί η επιρροή του σε σχέση με την ευαισθησία. Το αποτέλεσμα είναι ένας ενιαίος δείκτης απόδοσης που επιβραβεύει τα μοντέλα τα οποία ανιχνεύουν με αξιοπιστία τα επεισόδια κρίσεων χωρίς να παράγουν υπερβολικό αριθμό ψευδών συναγερμών.

$$Score = \underbrace{Sensitivity(\%)}_{\text{OVLP}} - 0.4 * \underbrace{\frac{FA}{h}}_{\text{EPOCH}}$$

1.4.3 Περιγραφή Πειραμάτων

Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε ένα AWS EC2 instance εξοπλισμένο με GPU. Όλες οι αξιολογήσεις πραγματοποιήθηκαν στο υποσύνολο των καταγραφών που περιείχαν το πλήρες σύνολο των μορφών σήματος (EEG, EMG και ECG), ώστε τα μοντέλα να μπορούν να αξιοποιήσουν την πολυτροπική πληροφορία.

Μοντέλα Μηχανικής Μάθησης

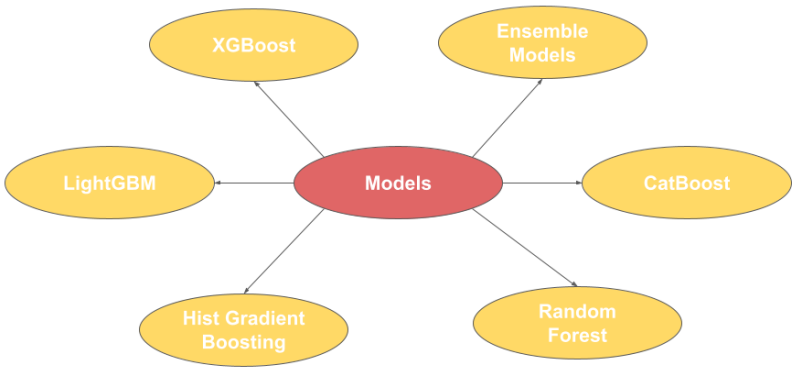

Figure 1.4.3: Μοντέλα Μηχανικής μάθησης που χρησιμοποιήθηκαν

Τα μοντέλα μηχανικής μάθησης εκπαιδεύτηκαν σε σύνολα δεδομένων με τα χειροποίητα χαρακτηριστικά που περιγράφονται στην 5.1. Έξι εκδόσεις κάθε μοντέλου εκπαιδεύτηκαν: για κάθε λόγο κρίσεων προς μη-κρίσεις χρησιμοποιήθηκε ένα σύνολο δεδομένων με ενισχυμένα δείγματα και ένα χωρίς. Κάθε έκδοση αξιολογήθηκε σε πολλαπλά κατώφλια για την ταξινόμηση ενός δείγματος ως κρίση, και επιλέχθηκε το κατώφλι με την καλύτερη απόδοση.

Τα πειράματα διεξήχθησαν με τη χρήση διαφόρων μοντέλων μηχανικής μάθησης, συγκεκριμένα XGBoost, LightGBM, CatBoost, Random Forest και HistGradientBoosting, καθώς και συνδυασμών αυτών με μεθόδους ensemble. Πιο συγκεκριμένα, αξιολογήθηκαν δύο προσεγγίσεις ensemble. Η πρώτη περιλάμβανε όλα τα προαναφερθέντα μοντέλα, ενώ η δεύτερη περιορίστηκε στα τρία με την καλύτερη απόδοση: XGBoost, CatBoost και LightGBM. Και στις δύο περιπτώσεις, η υλοποίηση βασίστηκε σε στρατηγική πλειοψηφικής ψήφου, όπου

κάθε μοντέλο προβλέπει την κλάση για ένα δεδομένο δείγμα και η κλάση που λαμβάνει τις περισσότερες ψήφους επιλέγεται ως τελική έξοδος.

Μοντέλα Βαθιάς Μάθησης

Σε αντίθεση με τα μοντέλα μηχανικής μάθησης, τα μοντέλα βαθιάς μάθησης εκπαιδεύτηκαν απευθείας στα προεπεξεργασμένα χρονοσειριακά δεδομένα, χωρίς να απαιτούνται χειροποίητα χαρακτηριστικά. Ωστόσο, τα μοντέλα βαθιάς μάθησης απέτυχαν να επιτύχουν την αναμενόμενη απόδοση, υποαποδίδοντας σε σχέση με τις προσεγγίσεις μηχανικής μάθησης.

GDN Το μοντέλο GDN προτάθηκε στο [33] και αποτελεί ένα Graph Neural Network κατάλληλο για ανίχνευση ανωμαλιών σε πολυμεταβλητές χρονοσειρές. Η υλοποίηση του μοντέλου είναι διαθέσιμη ανοιχτά στο GitHub. Κάθε κανάλι EEG, ECG, EMG και MOV θεωρήθηκε κόμβος στο γράφημα του νευρωνικού δικτύου. Το μοντέλο είχε σχεδιαστεί για δεδομένα διαφορετικά από βιοσήματα. Επιπλέον, τα δεδομένα EEG περιλάμβαναν μόνο δύο κανάλια, με αποτέλεσμα να δημιουργούνται ελάχιστοι κόμβοι στο GNN. Αυτοί οι λόγοι κατέστησαν το GDN ακατάλληλο για το συγκεκριμένο έργο.

xLSTM Το μοντέλο xLSTM αποτελεί επέκταση της τυπικής αρχιτεκτονικής LSTM και προτάθηκε πρόσφατα από τους [18]. Η υλοποίησή του είναι επίσης διαθέσιμη ανοιχτά στο GitHub. Για τους σκοπούς της παρούσας εργασίας χρησιμοποιήσαμε την εκδοχή xLSTMBlockStack, η οποία έχει σχεδιαστεί ειδικά για εφαρμογές πέραν της γλωσσικής επεξεργασίας. Παρά τον ελπιδοφόρο σχεδιασμό του, το μοντέλο δεν πέτυχε ικανοποιητικά αποτελέσματα στα πειράματά μας. Συγκεκριμένα, παρουσίασε έντονο υπερπροσαρμογή (overfitting), γεγονός που περιόρισε σημαντικά την ικανότητά του να γενικεύει σε μη ορατά δεδομένα και τελικά το καθιστά αναποτελεσματικό για το έργο ανίχνευσης κρίσεων επιληψίας.

Εξηγησιμότητα

Για να αποκτήσουμε καλύτερη κατανόηση της διαδικασίας λήψης αποφάσεων των μοντέλων, χρησιμοποιήσαμε τη βιβλιοθήκη SHAP, η οποία παρέχει επεξηγήσεις για την απόδοση χαρακτηριστικών. Επιπλέον, το TE2Rules [81] εφαρμόστηκε στους ταξινομητές XGBoost και Random Forest, επιτρέποντας την εξαγωγή κανόνων απόφασης κατανοητών από τον άνθρωπο. Αυτά τα εργαλεία μας επέτρεψαν να ερμηνεύσουμε καλύτερα τους υποκείμενους μηχανισμούς των μοντέλων και να εντοπίσουμε τα πιο σχετιζόμενα χαρακτηριστικά για την ανίχνευση κρίσεων επιληψίας.

Από τη βιβλιοθήκη SHAP αξιοποιήσαμε το beeswarm plot, μια σύνθετη και πλούσια σε πληροφορίες απεικόνιση των τιμών SHAP που αποκαλύπτει όχι μόνο τη σχετική σημασία των χαρακτηριστικών, αλλά και τις πραγματικές τους σχέσεις με το προβλεπόμενο αποτέλεσμα. Ένα παράδειγμα beeswarm plot φαίνεται στο 5.3.7. Σε ένα τέτοιο διάγραμμα, κάθε σημείο αντιστοιχεί σε μία καταγραφή του συνόλου δεδομένων, με τη θέση του στον οριζόντιο άξονα να υποδηλώνει την τιμή SHAP και το χρώμα του να αντανακλά την αρχική τιμή του χαρακτηριστικού. Η απεικόνιση αυτή μας επιτρέπει να παρατηρούμε ταυτόχρονα πόσο ισχυρά επηρεάζει ένα χαρακτηριστικό την απόφαση του μοντέλου, την κατεύθυνση της επίδρασής του και την κατανομή αυτών των επιδράσεων σε όλα τα δείγματα. Ως εκ τούτου, τα beeswarm plots παρέχουν μια συνοπτική αλλά ολοκληρωμένη εικόνα των συνεισφορών των χαρακτηριστικών, καθιστώντας τα ιδιαίτερα χρήσιμα για τον εντοπισμό κυρίαρχων παραγόντων πρόβλεψης και για την κατανόηση των αλληλεπιδράσεων χαρακτηριστικών–αποτελεσμάτων.

1.4.4 Αποτελέσματα

1.4.5 Απόδοση

Αυτή η υποενότητα παρουσιάζει τις καλύτερες επιδόσεις που επιτεύχθηκαν για κάθε αναλογία εκπαίδευσης, διαχωρίζοντας τα μοντέλα που εκπαιδεύτηκαν με αυξημένα δείγματα (augmented samples) από εκείνα που εκπαιδεύτηκαν χωρίς επέκταση δεδομένων. Συγκρίνοντας αυτά τα αποτελέσματα, μπορούμε να αξιολογήσουμε την επίδραση της ανισορροπίας των κλάσεων στην απόδοση των μοντέλων, καθώς και την αποτελεσματικότητα της τεχνικής επέκτασης δεδομένων. Επιπλέον, τα αποτελέσματα αυτά επιτρέπουν τη διατύπωση γενικότερων συμπερασμάτων σχετικά με τα σχετικά πλεονεκτήματα και αδυναμίες των αξιολογημένων μοντέλων.

Το μοντέλο με την καλύτερη απόδοση ήταν το σύνολο (ensemble) των XGBoost, CatBoost και LightGBM. Τα επόμενα ισχυρότερα μοντέλα ήταν τα XGBoost και CatBoost, ενώ το HistGradientBoosting κατέλαβε τη χαμηλότερη θέση μεταξύ των ταξινομητών. Επιπλέον, παρατηρούμε ότι στις περισσότερες περιπτώσεις, τα μοντέλα που εκπαιδεύτηκαν με αναλογία 1:2 πέτυχαν την καλύτερη απόδοση. Αυτό αποδίδεται στο γεγονός ότι η αναλογία αυτή βοηθά στην αντιμετώπιση της εγγενούς ανισορροπίας των κλάσεων στο σύνολο δεδομένων, επιτρέποντας στους ταξινομητές να διακρίνουν καλύτερα τα δείγματα κρίσεων από τα μη-κρίσεων. Ωστόσο, η επίδραση της προσθήκης αυξημένων δειγμάτων διαφέρει από μοντέλο σε μοντέλο.

XGBoost

Οι καλύτερες επιδόσεις των διαφόρων μοντέλων XGBoost, ανάλογα με την αναλογία εκπαίδευσης και τη χρήση επαυξημένων δεδομένων (augmented data). Η υψηλότερη απόδοση επιτεύχθηκε με το μοντέλο που εκπαιδεύτηκε στο σύνολο δεδομένων με αναλογία 1:2 και περιλάμβανε επαυξημένα δείγματα. Συνολικά, ο ταξινομητής XGBoost παρουσίασε μικτά αποτελέσματα στις διάφορες αναλογίες και επαύξησης δεδομένων.

LightGBM

Η καλύτερη απόδοση επιτεύχθηκε από το μοντέλο που εκπαιδεύτηκε με αναλογία 1:2 χωρίς επαυξημένα δείγματα. Γενικά, η απόδοση βελτιωνόταν καθώς η αναλογία αυξανόταν, ενώ τα μοντέλα που εκπαιδεύτηκαν με αναλογία 1:100 δεν κατάφεραν να εκτελέσουν επιτυχώς το έργο. Επιπλέον, τα μοντέλα που εκπαιδεύτηκαν με επαυξημένα δείγματα παρουσίασαν σταθερά χαμηλότερη απόδοση σε σύγκριση με εκείνα που εκπαιδεύτηκαν μόνο με τα αρχικά δεδομένα.

CatBoost

Το CatBoost παρουσιάζει αντίθετη συμπεριφορά σε σχέση με το LightGBM. Σε αυτή την περίπτωση, η απόδοση βελτιώνεται καθώς μειώνεται η αναλογία, και στις περισσότερες περιπτώσεις η εκπαίδευση με επαυξημένα δείγματα αποδεικνύεται ωφέλιμη. Το μοντέλο με την καλύτερη απόδοση ήταν αυτό που εκπαιδεύτηκε στο σύνολο δεδομένων με αναλογία 1:100 και περιλάμβανε επαυξημένα δείγματα.

Random Forest

Στη περίπτωση Random Forest, τόσο η αύξηση της αναλογίας όσο και η ενσωμάτωση επαυξημένων δειγμάτων αποδείχθηκαν ωφέλιμες. Το μοντέλο με την καλύτερη απόδοση ήταν αυτό που εκπαιδεύτηκε στο σύνολο δεδομένων με αναλογία 1:2 και περιλάμβανε επαυξημένα δείγματα.

HistGradientBoosting

Το HistGradientBoosting παρουσίασε συνολικά τα χαμηλότερα αποτελέσματα, δείχνοντας συμπεριφορά παρόμοια με το LightGBM. Συγκεκριμένα, η μείωση της αναλογίας και η προσθήκη επαυξημένων δειγμάτων οδήγησαν σε μειωμένη απόδοση, ενώ τα μοντέλα που εκπαιδεύτηκαν με αναλογία 1:100 δεν κατάφεραν να εκτελέσουν επιτυχώς το έργο. Τα καλύτερα αποτελέσματα σε αυτή την περίπτωση επιτεύχθηκαν με το μοντέλο που εκπαιδεύτηκε με αναλογία 1:2 χωρίς επαυξημένα δείγματα.

1.4.6 Ερμηνευσιμότητα - SHAP

Σε αυτήν την υποενότητα παρουσιάζονται τα αποτελέσματα των γραφημάτων SHAP beeswarm που χρησιμοποιήθηκαν για την ερμηνεία των εκπαιδευμένων μοντέλων. Η ανάλυση δείχνει ότι το σημαντικότερο χαρακτηριστικό για την ανίχνευση κρίσεων είναι το ενδοτεταρτημοριακό εύρος (IQR), ακολουθούμενο από το εύρος αιχμής-προς-αιχμή του σήματος ECG και την απόλυτη ισχύ της ζώνης θήτα. Παρατηρούμε ότι υψηλότερες τιμές του IQR, του εύρους αιχμής-προς-αιχμή του σήματος ECG και της απόλυτης ισχύος της ζώνης θήτα οδηγούν τα μοντέλα στο να χαρακτηρίζουν το δείγμα ως κρίση.

Τα ευρήματά μας συμφωνούν με προηγούμενη έρευνα, γεγονός που υποδηλώνει ότι τα αναγνωρισμένα χαρακτηριστικά αποτελούν πράγματι αξιόπιστους δείκτες επιληπτικής δραστηριότητας. Για το ενδοτεταρτημοριακό εύρος (IQR), οι [19] έδειξαν ότι το IQR είναι ένα ιδιαίτερα διακριτικό χαρακτηριστικό που διαχωρίζει αποτελεσματικά τα φυσιολογικά, μεσοκριτικά και κριτικά τμήματα EEG, επιτυγχάνοντας σχεδόν 100% ακρίβεια ταξ-

ινόμησης. Όσον αφορά τα χαρακτηριστικά πλάτους του ECG, οι [165] ποσοτικοποίησαν τις αλλαγές στη μορφολογία του συμπλέγματος QRS—συμπεριλαμβανομένου του εύρους αιχμής-προς-αιχμή—ως χρήσιμους δείκτες για την ανίχνευση κρίσεων επιληψίας. Τέλος, στο πλαίσιο της φασματικής ισχύος, οι [56] ανέφεραν ότι η αύξηση της απόλυτης ισχύος στη ζώνη θήτα αποτελεί μία από τις πιο συνεπείς υπογραφές της επιληπτικής δραστηριότητας σε πολλές μελέτες.

Άλλα σημαντικά χαρακτηριστικά περιλαμβάνουν το διάστημα RR (προερχόμενο από το σήμα ECG), τη λοξότητα του κύματος ECG και τον καρδιακό ρυθμό. Όπως φαίνεται στα διαγράμματα SHAP beeswarm, χαμηλότερες τιμές του διαστήματος RR (που αντιστοιχούν σε υψηλότερο καρδιακό ρυθμό) συσχετίζονται με γεγονότα κρίσης. Το εύρημα αυτό είναι συνεπές με κλινικές αποδείξεις, καθώς πολλές μελέτες έχουν δείξει ότι οι κρίσεις συχνά προκαλούν έντονες αλλαγές στον καρδιακό ρυθμό. Η μελέτη [70] ανέφερε ότι οι μεταβολές στα διαστήματα RR είναι συχνές στην επιληψία του κροταφικού λοβού, ενώ η μελέτη [114] διαπίστωσε ότι τα συντομευμένα διαστήματα RR και ο αυξημένος καρδιακός ρυθμός συχνά συνοδεύουν κρίσεις.

Υψηλότερες τιμές καρδιακού ρυθμού συνδέονται με γεγονότα κρίσης. Το εύρημα αυτό υποστηρίζεται από προηγούμενες μελέτες, οι οποίες κατέδειξαν ότι πολλές κρίσεις συνοδεύονται από έντονες αυξήσεις στον καρδιακό ρυθμό, ένα φαινόμενο γνωστό ως κριτική ταχυκαρδία. Για παράδειγμα, η μελέτη [40] ανέφερε ότι η κριτική ταχυκαρδία εμφανίζεται στην πλειονότητα των εστιακών κρίσεων, ενώ η μελέτη [17] έδειξε ότι οι μεταβολές στον καρδιακό ρυθμό είναι από τα πιο συνεπή αυτόνομα σημάδια κατά τη διάρκεια των κρίσεων.

XGBoost

Το XGBoost πέτυχε την καλύτερη απόδοση ανάμεσα στα δοκιμασμένα μοντέλα. Η ανάλυση SHAP δείχνει ότι το πιο σημαντικό χαρακτηριστικό για το XGBoost – συμφωνώντας με τους περισσότερους αξιολογημένους ταξινομητές – είναι το ενδοτεταρτημοριακό εύρος (IQR). Ωστόσο, εμφανίζονται ορισμένες διαφορές στην κατάταξη των επόμενων χαρακτηριστικών: το δεύτερο σημαντικότερο χαρακτηριστικό για το XGBoost είναι η απόλυτη ισχύς της ζώνης theta, ακολουθούμενη από το πλάτος αιχμής-προς-αιχμή (peak-to-peak amplitude) του σήματος ECG. Επιπλέον, ο καρδιακός ρυθμός φαίνεται να αποτελεί λιγότερο σημαντικό χαρακτηριστικό για το XGBoost σε σύγκριση με τους υπόλοιπους ταξινομητές.

LightGBM

Τα πιο σημαντικά χαρακτηριστικά που αναγνωρίστηκαν από το LightGBM είναι σε μεγάλο βαθμό συνεπή με εκείνα των άλλων αξιολογημένων ταξινομητών. Ωστόσο, μια αξιοσημείωτη διαφορά είναι ότι η τυπική απόκλιση (STD) του σήματος EEG φαίνεται να είναι λιγότερο σημαντική στην κατάταξη χαρακτηριστικών του LightGBM.

CatBoost

Σε αντίθεση με τους άλλους αξιολογημένους ταξινομητές, το CatBoost παρουσιάζει αρκετές σημαντικές διαφορές στην κατάταξη της σημασίας των χαρακτηριστικών. Το πιο σημαντικό χαρακτηριστικό για το CatBoost είναι ο καρδιακός ρυθμός, ενώ το ενδοτεταρτημοριακό εύρος (IQR) εμφανίζεται μόνο στην τέταρτη θέση και η τυπική απόκλιση (STD) κατατάσσεται πέμπτη. Παρά αυτές τις διαφορές στη σημασία των χαρακτηριστικών, τα αποτελέσματα που παρουσιάζονται στην Ενότητα 1.4.5 δείχνουν ότι το CatBoost ήταν το δεύτερο καλύτερο μοντέλο συνολικά.

Random Forest

Το Random Forest παρουσιάζει επίσης αρκετές σημαντικές διαφορές από τους άλλους αξιολογημένους ταξινομητές όσον αφορά την κατάταξη της σημασίας των χαρακτηριστικών. Το πιο σημαντικό χαρακτηριστικό είναι η τυπική απόκλιση (STD), ακολουθούμενη από το ενδοτεταρτημοριακό εύρος (IQR). Ενδιαφέρον είναι ότι το πλάτος αιχμής-προς-αιχμή (peak-to-peak amplitude) του σήματος ECG κατατάσσεται μόνο στην έβδομη θέση. Επιπλέον, η απόλυτη ισχύς των δέλτα και άλφα ζωνών έχει μεγαλύτερη επιρροή στις αποφάσεις του Random Forest σε σύγκριση με τους άλλους ταξινομητές.

HistGradientBoosting

Τα πιο σημαντικά χαρακτηριστικά που αναγνωρίστηκαν για το HistGradientBoosting είναι σε μεγάλο βαθμό συνεπή με τις μέσες τάσεις που παρατηρούνται στους υπόλοιπους ταξινομητές. Μια μικρή διαφορά είναι ότι η

τυπική απόκλιση φαίνεται να έχει μικρότερη επιρροή σε αυτό το μοντέλο. Παρ' όλα αυτά, όπως φαίνεται στην Ενότητα 1.4.5, παρά το γεγονός ότι το προφίλ σημασίας χαρακτηριστικών του είναι ευρέως ευθυγραμμισμένο με των άλλων, το HistGradientBoosting ήταν ο χειρότερος ταξινομητής συνολικά.

Σύνολα Μοντέλων (Ensemble Models)

Τα SHAP beeswarm plots δημιουργήθηκαν επίσης για τα δύο σύνολα μοντέλων, υπολογίζοντας τον μέσο όρο των τιμών SHAP από τα επιμέρους μοντέλα τους. Συνολικά, τα σύνολα εμφανίζουν παρόμοια μοτίβα σημασίας χαρακτηριστικών, με μόνο μικρές διαφοροποιήσεις στα λιγότερο σημαντικά χαρακτηριστικά. Όπως φαίνεται στην Ενότητα 1.4.5, το σύνολο που αποτελείται από τα τρία καλύτερα μοντέλα (XGBoost, CatBoost και LightGBM) ξεπέρασε σε απόδοση το σύνολο που περιλάμβανε όλους τους ταξινομητές.

1.4.7 Ερμηνευσιμότητα - TE2Rules

Η παρούσα υποενότητα παρουσιάζει τα αποτελέσματα της βιβλιοθήκης TE2Rules, η οποία προτάθηκε στο [81]. Η τεχνική αυτή εξάγει έναν κατάλογο κανόνων που αποτυπώνει τις αναγκαίες και ικανές συνθήκες για την ταξινόμηση μέσω των Tree Ensemble. Ο αλγόριθμος που χρησιμοποιείται βασίζεται στο Apriori Rule Mining. Η βιβλιοθήκη TE2Rules είναι συμβατή με τα μοντέλα XGBoost και Random Forest από τα μοντέλα που χρησιμοποιήσαμε.

Εξετάζοντας τους παραγόμενους κανόνες για το Random Forest, παρατηρούμε ότι τα χαρακτηριστικά που χρησιμοποιούνται συχνότερα είναι η τυπική απόκλιση (std) και το εύρος τεταρτημορίων (IQR), γεγονός που συνάδει με τα ευρήματα της ανάλυσης SHAP. Ωστόσο, αρκετοί κανόνες ενσωματώνουν επίσης χαρακτηριστικά όπως η ασυμμετρία (skewness), ο λόγος θήτα προς βήτα (theta-to-beta ratio) και τα χαρακτηριστικά πολυπλοκότητας Hjorth, τα οποία δεν εμφανίστηκαν ως ιδιαίτερα σημαντικά στα διαγράμματα SHAP. Αυτό υποδηλώνει ότι οι επεξηγήσεις που βασίζονται σε κανόνες ενδέχεται να αποτυπώνουν διαφορετικές σχέσεις από αυτές που αποκαλύπτει η ανάλυση SHAP.

Παρόμοια μοτίβα μπορούν να παρατηρηθούν και στους κανόνες που εξάγονται από το μοντέλο XGBoost. Και πάλι, η τυπική απόκλιση (std) και το εύρος τεταρτημορίων (IQR) εμφανίζονται με συνέπεια στους περισσότερους κανόνες, ενώ άλλα χαρακτηριστικά όπως η ασυμμετρία (skewness), ο λόγος θήτα προς βήτα και η πολυπλοκότητα Hjorth αναδεικνύονται επίσης, παρά το γεγονός ότι είχαν μικρότερη βαρύτητα στην ανάλυση SHAP. Αυτή η ποικιλία υπογραμμίζει μια κεντρική πρόκληση στην ερμηνευσιμότητα μοντέλων, ιδίως με μεθόδους εκ των υστέρων (post-hoc), δηλαδή τη δυσκολία καθιέρωσης εμπιστοσύνης στις παραγόμενες επεξηγήσεις και την απουσία ενός οριστικού σημείου αναφοράς έναντι του οποίου να μπορούν να επικυρωθούν.

1.5 Συμπεράσματα

1.5.1 Συζήτηση

Στην παρούσα διπλωματική εργασία, δοκιμάσαμε μία μεθοδολογία βασισμένη σε χαρακτηριστικά που προκύπτουν από μη αυτόματη εξαγωγή (handcrafted features) από σήματα EEG. Τα στάδια της προεπεξεργασίας και της εξαγωγής χαρακτηριστικών παρουσίασαν σημαντικά καλύτερη απόδοση σε σχέση με ορισμένα μοντέλα βαθιάς μάθησης. Αξιοσημείωτο είναι ότι στον Una Europa Seizure Detection Challenge του 2023, η νικήτρια προσέγγιση ακολούθησε παρόμοιο πλαίσιο και επίσης ανέφερε υπεροχή έναντι λύσεων βασισμένων στη βαθιά μάθηση [67]. Ένα βασικό χαρακτηριστικό των συνόλων δεδομένων του διαγωνισμού είναι ότι αποτελούνται από καταγραφές EEG με περιορισμένο αριθμό καναλιών, τα οποία αποκτήθηκαν από φορητές συσκευές. Ως εκ τούτου, τα σήματα είναι χαμηλότερης ποιότητας και περιέχουν υψηλότερα επίπεδα θορύβου. Αυτά τα χαρακτηριστικά υποδηλώνουν ότι η προεπεξεργασία και η εξαγωγή χαρακτηριστικών είναι καταλληλότερες για το συγκεκριμένο πρόβλημα από τις προσεγγίσεις βαθιάς μάθησης.

Η Εξηγήσιμη Τεχνητή Νοημοσύνη (XAI) στην ανάλυση EEG αποτελεί ένα πεδίο που βρίσκεται σε εξέλιξη. Η συντριπτική πλειονότητα των μελετών που εξετάστηκαν χρησιμοποιούν *post-hoc* προσεγγίσεις, όπως μεθόδους βασισμένες σε οπτικοποίηση (π.χ. saliency maps, Grad-CAM), απόδοση χαρακτηριστικών (π.χ. SHAP, LIME) και τεχνικές διαταραχής/απόκρυψης. Ωστόσο, παρά τη μεγάλη τους διάδοση, οι *post-hoc* μέθοδοι επεξήγησης έχουν δεχθεί σημαντική κριτική σχετικά με την αξιοπιστία, τη σταθερότητα και τη συμφωνία τους με τον τρόπο λήψης αποφάσεων του μοντέλου. Ένα κεντρικό ζήτημα είναι η πιστότητα: κατά πόσο η εξήγηση αντικατοπτρίζει

με ακρίβεια τον εσωτερικό συλλογισμό του μοντέλου και όχι μια πειστική αλλά ενδεχομένως παραπλανητική αφήγηση [37, 134]. Οι *post-hoc* μέθοδοι όπως οι LIME και SHAP βασίζονται συχνά σε προσεγγίσεις (π.χ. τοπικά υποκατάστατα μοντέλα), οι οποίες μπορούν να εισάγουν τεχνητά στοιχεία που παρερμηνεύουν την πραγματική συμπεριφορά του μοντέλου. Ένα ακόμη κρίσιμο ζήτημα είναι η σταθερότητα: οι εξηγήσεις μπορεί να είναι ιδιαίτερα ευαίσθητες σε μικρές μεταβολές των δεδομένων εισόδου, γεγονός που δημιουργεί ερωτήματα για την αξιοπιστία τους [59], ενώ πρόσφατες εργασίες υπογραμμίζουν ότι η αποτελεσματικότητα των *post-hoc* μεθόδων εξαρτάται από το μοντέλο και είναι ευάλωτη σε συσχετίσεις μεταξύ χαρακτηριστικών [136]. Ακόμη χειρότερα, οι ίδιες οι μέθοδοι επεξήγησης μπορούν να χειραγωγηθούν, επιτρέποντας το λεγόμενο fairwashing, όπου προκατειλημμένα μοντέλα συνοδεύονται από φαινομενικά δίκαιες εξηγήσεις [5, 148]. Αυτές οι αδυναμίες αναδεικνύουν τον κίνδυνο επιθέσεων όχι μόνο στα μοντέλα, αλλά και στις ίδιες τις εξηγήσεις, υπονομεύοντας την εμπιστοσύνη των χρηστών. Ως εκ τούτου, ορισμένοι ερευνητές υποστηρίζουν τη χρήση εγγενώς ερμηνεύσιμων μοντέλων, γνωστών και ως interpretable by design, ως μια πιο αξιόπιστη εναλλακτική έναντι των *post-hoc* μεθόδων [134].

Διαπιστώνουμε ότι οι μέθοδοι XAI έχουν εφαρμοστεί σε ένα ευρύ φάσμα εργασιών που σχετίζονται με EEG. Αυτές περιλαμβάνουν σημαντικές ιατρικές και ψυχολογικές εφαρμογές όπως η *ανίχνευση επιληψίας και κρίσεων*, η *παρακολούθηση ύπνου*, η *διάγνωση σχιζοφρένειας*, η *νοητική απεικόνιση και εκτέλεση κινήσεων*, η *αναγνώριση συναισθημάτων*, η *πρόβλεψη εγκεφαλικού επεισοδίου* και η *μεγάλη καταθλιπτική διαταραχή*. Παρόλο που αυτό δείχνει την ευελιξία της XAI σε διαφορετικούς τομείς, οι περισσότερες μέθοδοι σχεδιάζονται για μια συγκεκριμένη εργασία και σπάνια δοκιμάζονται σε άλλες εφαρμογές. Αυτό περιορίζει τη γενικότητά τους και δυσχεραίνει τη σύγκριση μεταξύ διαφορετικών μελετών EEG. Επιπλέον, τα περισσότερα έργα εστιάζουν αποκλειστικά σε δεδομένα EEG, χωρίς να τα συνδυάζουν με άλλους τύπους δεδομένων, όπως εγκεφαλικές εικόνες (π.χ. MRI ή fMRI), που χρησιμοποιούνται ευρέως στη νευροεπιστήμη. Υπάρχει επίσης έλλειψη μελετών ερμηνευσιμότητας σε πολυτροπικές προσεγγίσεις—περιπτώσεις όπου το EEG συνδυάζεται με οπτικά ή άλλα αισθητηριακά δεδομένα—παρά το γεγονός ότι τέτοιοι συνδυασμοί θα μπορούσαν να προσφέρουν πιο ολοκληρωμένες και αξιόπιστες γνώσεις για τη λειτουργία του εγκεφάλου.

Ένας βασικός περιορισμός των υφιστάμενων προσεγγίσεων είναι η απουσία σαφώς καθορισμένων στόχων ερμηνευσιμότητας και αληθών επεξηγήσεων αναφοράς. Όπως σημειώνουν οι [63], οι δημοφιλείς μέθοδοι XAI μπορούν να αποδίδουν εσφαλμένα σημασία σε άσχετα χαρακτηριστικά εισόδου, γεγονός που αποτελεί σημαντικό κίνδυνο σε ιατρικές εφαρμογές όπου η ερμηνευσιμότητα είναι κρίσιμη. Η έλλειψη αντικειμενικών κριτηρίων αξιολόγησης και σαφών ορισμών του προβλήματος εμποδίζει την επικύρωση της ορθότητας των εξηγήσεων, περιορίζοντας τη χρησιμότητά τους για τη βελτίωση των μοντέλων και την πρόοδο της επιστημονικής γνώσης. Επιπλέον, οι τρέχουσες προσεγγίσεις XAI στην ανάλυση EEG στερούνται ολοκληρωμένης επικύρωσης μέσω αξιολόγησης από ανθρώπους, ιδίως από ειδικούς στον χώρο της υγείας. Μια τέτοια επικύρωση είναι απαραίτητη για την καλλιέργεια εμπιστοσύνης και την ενίσχυση της υιοθέτησης σε ιατρικά περιβάλλοντα. Η εφαρμογή πλαισίων αξιολόγησης με επίκεντρο τον άνθρωπο, όπως αυτό που προτείνεται από τους [116], θα μπορούσε να βοηθήσει στην αξιολόγηση των εξηγήσεων μέσω ανατροφοδότησης από ειδικούς, γεφυρώνοντας το χάσμα μεταξύ τεχνικών λύσεων XAI και πρακτικών αναγκών στον τομέα της υγειονομικής περίθαλψης.

1.5.2 Μελλοντικές Κατευθύνσεις

Παρόλο που η βαθιά μάθηση δεν πέτυχε τα αναμενόμενα αποτελέσματα στα πειράματά μας, παραμένει ένα ισχυρό εργαλείο με μεγάλες δυνατότητες στην αυτόματη εξαγωγή χαρακτηριστικών και στην αναγνώριση προτύπων. Πρόσφατες μελέτες έχουν προτείνει υβριδικές αρχιτεκτονικές που ενσωματώνουν πολλαπλές μεθοδολογίες [180], παρουσιάζοντας ενθαρρυντικά αποτελέσματα. Στο πλαίσιο μελλοντικής έρευνας, η εφαρμογή τέτοιων υβριδικών προσεγγίσεων σε αυτό το σύνολο δεδομένων θα μπορούσε να βελτιώσει την απόδοση στο πρόβλημα της ανίχνευσης κρίσεων επιληψίας.

Οι τεχνικές XAI μπορούν να προσφέρουν πολύτιμες γνώσεις σε ένα εύρος εργασιών ανάλυσης EEG. Παρέχοντας διαφάνεια και ερμηνευσιμότητα στις αποφάσεις των μοντέλων, συμβάλλουν στην ενίσχυση της εμπιστοσύνης και της αξιοπιστίας, καθιστώντας τις τεχνικές αυτές πιο πρακτικές για εφαρμογή στον πραγματικό κόσμο. Παρά το αυξανόμενο ενδιαφέρον για την XAI στην ανάλυση EEG, το πεδίο εξακολουθεί να αντιμετωπίζει προκλήσεις, όπως ο περιορισμένος αριθμός διαφορετικών μοντέλων, η έλλειψη ολοκληρωμένης επικύρωσης από ειδικούς, καθώς και οι ανεπαρκώς καθορισμένοι στόχοι επεξήγησης. Η μελλοντική έρευνα θα πρέπει να επικεντρωθεί στην ενσωμάτωση αξιολόγησης με γνώμονα τους ειδικούς, ώστε τα συστήματα ΤΝ που βασίζονται σε EEG να είναι ταυτόχρονα αποτελεσματικά και αξιόπιστα σε ιατρικές εφαρμογές.

Chapter 2

Introduction

Electroencephalography (EEG) remains a primary diagnostic tool for brain-related conditions due to its non-invasive nature, high temporal resolution, and relative affordability. It provides a direct measure of neural activity, making it indispensable for understanding the dynamics of brain function. Clinically, EEG aids in identifying sleep disorders, detecting abnormal patterns associated with epilepsy, and monitoring brain states during anesthesia and coma. In research contexts, it supports investigations into cognition, attention, and emotion. However, EEG signals are inherently noisy and complex. They are low in amplitude, easily contaminated by artifacts such as eye blinks or muscle movements, and span multiple overlapping frequency bands. This variability, coupled with the multichannel and temporal nature of the data, makes analysis challenging when relying solely on conventional software tools or manual inspection.

In this landscape, Artificial Intelligence (AI) has emerged as a transformative approach to EEG analysis. Through automated feature extraction and pattern recognition, AI systems can detect hidden relationships within high-dimensional EEG data that are not easily visible to human experts. Deep learning models, in particular, can learn abstract representations directly from raw signals, allowing for robust classification and prediction in tasks such as seizure detection, mental workload estimation, and cognitive state monitoring [170]. These capabilities open new possibilities for real-time and scalable EEG-based applications, ranging from clinical diagnostics to brain–computer interface systems.

Despite its potential, AI in EEG analysis faces a major challenge—its "black-box" nature [29]. Most AI models, especially deep neural networks, provide little to no transparency regarding how specific predictions are made. In medical contexts, this opacity poses significant risks. A lack of interpretability may lead to decisions based on spurious correlations rather than physiologically meaningful patterns, which can erode trust among clinicians and hinder adoption in healthcare practice. For AI systems to be integrated into clinical workflows, transparency and accountability are crucial. Decisions must be traceable to meaningful EEG features and justifiable in terms of established neuroscientific understanding.

Explainable Artificial Intelligence (XAI) plays a vital role in addressing this issue by providing insights into the reasoning behind AI model decisions [38]. Through post-hoc or intrinsic interpretability techniques, XAI enables practitioners to understand which aspects of the EEG data influence model outputs. This interpretability not only enhances ethical and legal accountability but also contributes to clinical confidence in automated decision support tools. In the context of EEG-based diagnostics, XAI methods can highlight the most relevant spatial, temporal, or spectral features that guide classification. For instance, feature extraction methods [122] and saliency-based visualization techniques [169] make it possible to map model attention to specific EEG channels or frequency bands, aligning computational outcomes with neuroscientific principles. Such interpretability ensures that AI systems operate as collaborative tools for clinicians rather than opaque replacements.

Building on these motivations, this work provides a comprehensive overview of recent advancements in XAI for EEG analysis. It categorizes and examines methods designed to enhance transparency in EEG-related tasks, such as seizure detection, mental state recognition, and cognitive workload estimation. The goal is not only to summarize progress but also to clarify the conceptual and methodological landscape of XAI in this

field. We present key datasets and benchmarking protocols that are widely used to assess explainability and performance. Furthermore, by identifying current research gaps—such as limited standardization, lack of clinical validation, and insufficient integration of neurophysiological priors—we highlight directions for future exploration. This thesis aims to serve as both an entry point for new researchers and a reference for ongoing developments in explainable EEG analysis.

Beyond reviewing the literature, this work contributes a practical, end-to-end framework for EEG preprocessing and feature extraction tailored to the task of seizure detection. The framework is applied to a publicly available dataset originally introduced in the Una Europa Seizure Detection Challenge. It integrates multiple signal processing steps, including noise removal, channel selection, and feature computation, designed to retain the physiological relevance of EEG patterns. A diverse set of machine learning and deep learning models were trained and evaluated, encompassing both classical algorithms such as Random Forest and advanced architectures such as xLSTM. This comparative approach allows for a balanced assessment of different modeling paradigms, evaluating not only their predictive accuracy but also their interpretability and clinical plausibility.

To address the critical issue of transparency, two complementary XAI techniques were employed. SHAP (SHapley Additive exPlanations) was used to provide feature attribution analysis, quantifying the contribution of each feature to model outputs. In parallel, the te2rules algorithm was applied to extract human-readable rules from trained models, transforming opaque model behavior into interpretable symbolic representations. Together, these explainability methods offer a window into how EEG features drive classification decisions, bridging the gap between technical performance and medical insight. By combining robust data-driven modeling with interpretable reasoning, this work aims to deliver results that are both scientifically rigorous and clinically meaningful.

Ultimately, the broader ambition of this thesis is to contribute to the development of trustworthy AI systems for EEG analysis—systems that are not only accurate but also understandable, reliable, and aligned with clinical expertise. The integration of XAI into EEG-based diagnostics represents a critical step toward achieving human–AI collaboration in neuroscience and medicine, where explainability is not merely a desirable property but an ethical necessity.

The outline of this thesis is as follows:

- We first provide all the necessary background on tree-based and ensemble learning methods to establish the foundation for later discussions.
- We then introduce the fundamentals of Explainable Artificial Intelligence (XAI) and EEG analysis, followed by a detailed review of recent XAI methodologies applied to EEG data.
- Finally, we describe our experimental framework in detail, covering data preprocessing, feature extraction, model training, and interpretability analysis. We present both the performance outcomes and the explainability results, leading to a discussion of their implications and concluding insights.

Chapter 3

Background and Literature Review

Contents

3.1 Machine and Deep Learning Methods

In this thesis, we focus on tree-based machine learning models and their ensemble extensions. A single decision tree is easy to understand and interpret, but it usually suffers from overfitting and limited accuracy. Ensemble methods such as Random Forests and Gradient Boosting address these issues by combining many trees together, which generally leads to more robust and accurate predictions. Over time, several efficient implementations of gradient boosting, including XGBoost, LightGBM, CatBoost, and HistGradientBoosting, have been developed and are now widely used in practice.

In addition to the tree-based machine learning models, in this thesis we experimented with two novel deep learning architectures. A Graph Neural Network (GNN) designed for anomaly detection in multivariate time series was the first deep learning model chosen to be tested in the context of seizure detection. Furthermore, the Extended Long Short-Term Memory (xLSTM) model was tested, which is designed to overcome the LSTM limitations. The following subsections briefly describe and give the necessary theoretical background of the machine and deep learning models used.

3.1.1 Decision Trees

Decision trees are simple yet powerful predictive models that partition the feature space into regions by recursively splitting the data based on feature values. They are valued for their interpretability and ability to model non-linear relationships. However, individual trees are prone to overfitting and high variance, making them less robust in practice[25].

A decision tree operates by selecting, at each node, the feature and corresponding threshold that best separates the data according to a chosen impurity criterion, such as Gini impurity, information gain, or mean squared error. This hierarchical structure allows the model to capture complex feature interactions without the need for explicit feature engineering. Each leaf node corresponds to a decision outcome or predicted value, while the path from the root to a leaf represents a set of easily interpretable rules. This transparency has made decision trees widely used in domains where explainability and traceability are essential, such as healthcare, finance, and neuroscience.

Despite these advantages, individual decision trees often exhibit instability: small changes in the training data can produce substantially different tree structures. This instability results from the greedy nature of the splitting process, which optimizes decisions locally at each node without considering their global impact. As a result, while a single tree can achieve a good fit on training data, it often generalizes poorly to unseen data, particularly in high-dimensional or noisy environments. Ensemble techniques such as Random Forests and Gradient Boosted Trees were developed to address these issues by combining multiple trees to reduce variance, improve robustness, and enhance predictive accuracy.

Modern decision tree frameworks also include several enhancements that make them more suitable for real-world applications. Many implementations now support handling of missing values, mixed data types, and large-scale datasets through efficient data binning and parallelized training. These improvements retain the interpretability of decision trees while significantly improving their scalability and reliability in complex predictive modeling tasks.

In neuroimaging and affective computing, decision trees and their ensemble variants are frequently used to identify discriminative patterns in EEG, fMRI, and multimodal physiological data. Their rule-based structure enables researchers to trace which features or brain regions contribute most strongly to specific classifications or predictions. Although deep learning models often surpass them in accuracy, decision trees remain valuable as interpretable baselines and as components in hybrid frameworks that balance performance with transparency.

3.1.2 Random Forests

Random Forests address the limitations of single decision trees by combining many trees in a bagging framework [24]. Each tree is trained on a bootstrap sample of the data and at each split, only a random subset of features is considered. This randomness reduces correlation between trees and improves generalization.

Random Forests are widely used due to their robustness, relatively low need for parameter tuning, and strong performance across a variety of domains.

3.1.3 Gradient Boosting Methods

Gradient Boosting is an ensemble technique that builds models sequentially, where each new learner attempts to correct the errors of the previous ones [58]. It combines weak learners, typically decision trees, into a strong predictive model through gradient-based optimization. At each iteration, a new tree is trained to fit the negative gradient of the loss function with respect to the current model's predictions, thereby iteratively reducing the residual errors. This additive model framework enables the ensemble to approximate complex, nonlinear mappings while maintaining interpretability through the underlying tree structures.

Over the years, several efficient and scalable implementations of gradient boosting have been developed, improving upon the original algorithm's computational and regularization limitations. These implementations introduce techniques such as second-order gradient optimization, histogram-based feature binning, parallelized training, and advanced regularization mechanisms to prevent overfitting. Moreover, they offer robust support for handling missing values, categorical variables, and imbalanced datasets, which are common challenges in real-world applications. In this work, the following libraries were employed.

XGBoost

XGBoost (Extreme Gradient Boosting) is one of the most widely adopted gradient boosting frameworks due to its efficiency, scalability, and predictive accuracy [28]. It improves upon the original boosting formulation by incorporating second-order Taylor approximation of the loss function, allowing it to leverage both gradient and curvature information for more accurate split optimization. This second-order approach accelerates convergence and enhances model stability.

XGBoost also introduces several key innovations to improve generalization and computational performance. It employs shrinkage (learning rate reduction) to scale the contribution of each tree, thereby reducing the risk of overfitting, and column subsampling to introduce additional randomness, similar to Random Forests. The algorithm's sparsity-aware design allows it to efficiently handle missing or sparse feature values by learning optimal default split directions during training. Additionally, XGBoost supports parallel and distributed computation, enabling rapid training on large datasets.

Due to its strong regularization framework and robustness, XGBoost has become a standard baseline for structured data modeling. It performs exceptionally well in applications where feature interactions are complex but structured, such as tabular biomedical data, clinical risk prediction, and multimodal sensor analysis.

LightGBM

LightGBM (Light Gradient Boosting Machine) was developed to further improve the training efficiency and scalability of gradient boosting methods [72]. Unlike traditional level-wise growth used in algorithms like XGBoost, LightGBM employs a leaf-wise tree growth strategy with depth constraints. This approach allows the algorithm to focus on regions of the data with the greatest potential for loss reduction, often resulting in deeper trees with higher predictive power.

To reduce computational overhead, LightGBM implements histogram-based feature binning, which discretizes continuous features into a limited number of bins. This reduces both memory consumption and training time without significant loss of accuracy. Additionally, LightGBM supports Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which further accelerate training by focusing on informative samples and combining mutually exclusive features, respectively.

The combination of these techniques makes LightGBM highly efficient for large-scale and high-dimensional datasets. In practice, it often achieves comparable or superior performance to XGBoost while requiring less computational time, making it particularly attractive for scenarios involving high-frequency or real-time data streams such as physiological monitoring and large-scale neuroimaging feature sets.

CatBoost

CatBoost is a gradient boosting framework specifically designed to handle categorical variables effectively without extensive preprocessing [125]. Traditional boosting algorithms require categorical features to be transformed via one-hot encoding or ordinal mappings, which can lead to high dimensionality and target leakage. CatBoost overcomes these issues using ordered boosting and efficient encoding schemes based on target statistics.

The ordered boosting mechanism uses a permutation-driven approach that processes samples in a predefined order, ensuring that the encoding of each observation depends only on previously seen data. This design prevents the model from inadvertently learning from its own predictions—a common issue known as prediction shift—thus improving generalization. CatBoost also incorporates symmetry in its decision trees, enforcing consistent tree structures across all nodes, which enhances both efficiency and stability.

Another advantage of CatBoost is its strong out-of-the-box performance with minimal hyperparameter tuning, making it highly practical for applied machine learning. Its robust handling of categorical, numerical, and missing data makes it particularly useful in multimodal applications where data heterogeneity is prevalent, such as emotion recognition, EEG-based classification, or patient state modeling.

HistGradientBoosting

HistGradientBoosting is an efficient implementation of gradient boosting integrated into the `scikit-learn` library [120]. Inspired by LightGBM, it uses histogram-based feature binning to accelerate computation and reduce memory usage. By discretizing continuous features into a fixed number of bins, the algorithm computes gradient statistics over these bins rather than individual samples, significantly improving efficiency on medium-sized datasets.

Although it lacks some of the advanced features found in XGBoost, LightGBM, or CatBoost—such as built-in categorical handling or distributed training—HistGradientBoosting offers strong performance within the scikit-learn ecosystem. It supports early stopping, monotonic constraints, and native management of missing values, providing a well-balanced combination of speed, interpretability, and ease of integration into established machine learning workflows.

Its design makes it an ideal choice for medium-scale experimental studies where reproducibility, compatibility, and interpretability are prioritized. In the context of neuroimaging and physiological data modeling, HistGradientBoosting can serve as a reliable baseline model for structured feature sets, offering competitive accuracy while maintaining transparency and computational efficiency.

3.1.4 Model Ensembling

Ensemble methods combine multiple models to achieve better predictive performance than individual learners [35]. Bagging reduces variance by averaging predictions across independent models, while boosting reduces bias by sequentially improving upon previous learners. Stacking, another ensemble strategy, combines heterogeneous models using a meta-learner to optimize predictive accuracy. These approaches have proven particularly effective for structured tabular data, where ensembles often outperform single models.

Beyond traditional implementations, ensemble learning has been extensively adopted in neural and multimodal architectures, enabling improved robustness and generalization across heterogeneous data sources. For instance, in multimodal neuroimaging, ensemble frameworks can integrate modality-specific learners—such as separate deep networks for EEG and fMRI—whose outputs are fused via a meta-classifier or weighted averaging mechanism . This design mitigates modality imbalance and leverages complementary information across data types.

In deep learning contexts, ensembles of neural networks are also employed to stabilize training outcomes and estimate epistemic uncertainty. Techniques such as Monte Carlo dropout, deep ensembles, and snapshot ensembles approximate Bayesian inference by aggregating predictions from multiple stochastic realizations of the same architecture. These methods enhance calibration and improve performance in safety-critical applications, including brain-computer interface (BCI) decoding and clinical prediction tasks.

Recent research further explores hybrid ensemble paradigms, where symbolic reasoning or attention mechanisms are incorporated as higher-level aggregation layers. Such neuro-symbolic ensembles can interpret and weigh the reliability of base learners based on semantic context or learned uncertainty estimates. Moreover, ensemble distillation strategies aim to transfer the collective knowledge of an ensemble into a single compact model, thereby maintaining performance gains while reducing computational overhead.

Overall, ensemble learning represents a powerful design principle that transcends simple model aggregation. By integrating complementary learners, balancing variance and bias, and facilitating uncertainty quantification, ensemble methods provide a robust foundation for predictive modeling—particularly in complex, multimodal, or data-scarce domains.

3.1.5 Deep Learning Models

Deep learning has emerged as a dominant paradigm in modern data-driven modeling, offering exceptional capability for capturing complex, nonlinear relationships in high-dimensional datasets. Unlike traditional machine learning algorithms that rely on handcrafted features, deep neural networks (DNNs) automatically learn hierarchical representations of data through layered transformations, enabling the extraction of both low-level and abstract semantic patterns. Architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their numerous variants have been successfully applied across a wide range of domains—including computer vision, natural language processing, and biomedical signal analysis—demonstrating remarkable generalization and scalability.

In the context of neuroimaging and multimodal biosignal processing, deep learning methods have proven especially valuable for modeling complex spatial–temporal dependencies that are difficult to capture through conventional approaches. EEG and fMRI data, for instance, exhibit intricate spatial correlations and dynamic temporal fluctuations, reflecting latent neural processes that unfold over multiple scales. Deep learning frameworks can capture such dependencies through convolutional, recurrent, or attention-based mechanisms, often outperforming shallow or linear models in predictive accuracy and representation quality. Moreover, by combining modality-specific encoders within a unified architecture, deep models can integrate heterogeneous data sources—such as electrophysiological, hemodynamic, and behavioral features—into coherent multimodal embeddings.

Recent advances in graph-based and memory-augmented architectures have further expanded the expressive capacity of deep learning for structured and sequential data. Graph Neural Networks (GNNs) extend deep learning to relational domains, where data elements interact through irregular or dynamically evolving connections. This property is particularly advantageous for modeling interdependencies among EEG channels, brain regions, or sensors, where spatial relations are non-Euclidean and context-dependent. On the other hand, enhanced recurrent architectures such as Extended Long Short-Term Memory (xLSTM) networks introduce refined mechanisms for long-term information retention and efficient gradient propagation, enabling robust modeling of long-range temporal dynamics in non-stationary sequences.

The following subsections present two representative architectures that illustrate these trends: the Graph Deviation Network (GDN), a GNN-based model designed for relational anomaly detection in multivariate time series, and the Extended LSTM (xLSTM), an improved recurrent architecture tailored for capturing complex temporal dependencies. Both models exemplify the evolving landscape of deep learning research, emphasizing interpretability, adaptability, and domain relevance for applications in EEG and multimodal biosignal analysis.

GDN

The Graph Deviation Network (GDN), proposed by Deng and colleagues [33], is a specialized Graph Neural Network (GNN) architecture developed for anomaly detection in multivariate time series data. Unlike traditional approaches that treat multivariate signals as independent or loosely correlated sequences, GDN explicitly models the interdependencies among variables by representing them as nodes within a dynamically learned graph. Figure 3.1.1 provides an overview of the GDN framework and its primary components.

The central motivation behind GDN is that multivariate time series—such as those encountered in EEG, physiological monitoring, or sensor networks—often exhibit complex temporal and spatial correlations. Cap-

turing these relationships is crucial for understanding the normal dynamics of the system and for identifying deviations that signal abnormal or unexpected behavior. Rather than relying on a predefined graph structure, GDN learns the graph topology adaptively from data, enabling it to model both explicit and implicit dependencies between signals.

The GDN framework consists of four main stages, each contributing to the extraction of relational and temporal information. In the first stage, each variable in the multivariate time series is mapped into a latent embedding vector that encodes its unique statistical and dynamic characteristics. These embeddings serve as compact representations that facilitate learning inter-variable dependencies.

In the second stage, GDN learns a graph structure that captures the dependency relationships among the different time series variables. This graph can be viewed as an adjacency matrix whose edges represent the strength of relationships between variables. Importantly, the learned graph is not static—it evolves as the model observes new data, allowing for dynamic adaptation to changing correlation patterns.

The third stage involves forecasting future values of each time series variable based on its learned neighborhood relationships. GDN employs a graph attention mechanism to weigh the influence of neighboring nodes when predicting the next time step. This attention mechanism allows the model to focus on the most relevant nodes, thereby improving predictive accuracy and interpretability. The attention scores inherently provide insights into which variables most strongly influence a given signal's behavior, making this step particularly valuable for explainable modeling.

Finally, in the fourth stage, GDN performs anomaly detection by measuring deviations between the predicted and observed values. Deviations that significantly exceed the expected range are identified as anomalies, indicating potential faults, abnormal events, or irregular behaviors within the system. These deviations can also be interpreted through the learned graph structure, allowing users to trace which relationships contributed to an anomaly—thus offering an interpretable explanation for the detected event.

One of the key strengths of GDN lies in its ability to combine temporal prediction with relational reasoning. Unlike purely temporal models such as LSTM or Transformer-based architectures, GDN's graph-centric design provides a structured means to model dependencies that are neither strictly sequential nor uniformly distributed across variables. This makes GDN particularly effective in high-dimensional, interdependent systems where anomalies often arise from disruptions in relational patterns rather than individual signal fluctuations.

In the context of EEG or multimodal biosignal analysis, GDN holds significant potential. For instance, brain regions or sensor channels can be modeled as nodes in a graph, with edges representing functional or statistical relationships. Deviations in these learned connectivity patterns can reveal meaningful neural anomalies, cognitive state changes, or sensor artifacts. Moreover, the explainable nature of the graph and attention mechanisms aligns well with the objectives of XAI, as it allows for transparent reasoning about which relationships and features contributed to a detected anomaly.

xLSTM

The Extended Long Short-Term Memory (xLSTM) architecture, proposed by Beck et al. [18], represents a significant advancement over the classical Long Short-Term Memory (LSTM) model introduced by Hochreiter and Schmidhuber [64]. The original LSTM architecture was designed to address the vanishing gradient problem encountered in standard recurrent neural networks (RNNs), enabling the capture of long-term dependencies through its gated cell structure. However, despite its success, conventional LSTM models exhibit several inherent limitations. These include a restricted capacity for revising storage decisions once information is written into the memory cell, limited memory utilization efficiency, and the lack of parallelizability due to sequential memory mixing during computation.

The xLSTM framework addresses these issues through two principal innovations: the introduction of extended memory structures and the incorporation of exponential gating mechanisms. The new memory design enhances the model's ability to store and manage information over extended temporal contexts by decoupling memory allocation from the conventional gating dynamics. This modification enables the model to revisit and revise stored information, providing a more flexible and adaptive form of temporal reasoning. In contrast to

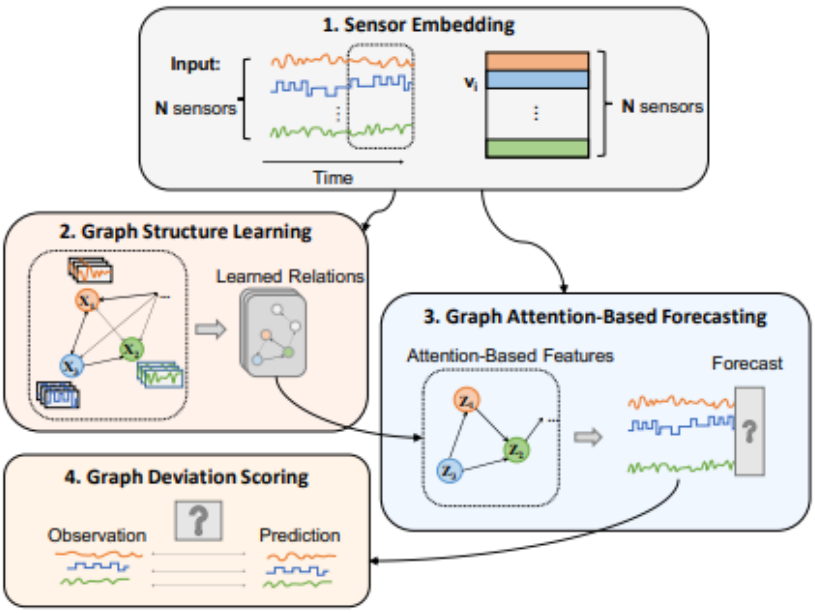

Figure 3.1.1: Overview of the GDN framework

traditional LSTMs, where the information flow is tightly coupled within a single recurrent pathway, xLSTM's design facilitates improved gradient propagation and more efficient information retention across time steps.

The addition of exponential gating further refines the model's control over information flow. Instead of relying solely on sigmoid or tanh activations for gating functions, exponential gating introduces a multiplicative scaling mechanism that allows smoother gradient transitions and more stable long-range learning. This modification enhances both training efficiency and the expressive capacity of the network, allowing it to model complex temporal dependencies without the instability often associated with deep or long recurrent architectures.

From a computational perspective, xLSTM also improves parallelizability by reducing interdependence between memory operations. This architectural refinement makes it more suitable for modern hardware accelerators such as GPUs and TPUs, where parallel execution of matrix operations significantly impacts performance. As a result, xLSTM achieves both higher scalability and lower training latency compared to standard LSTMs, making it a strong candidate for large-scale sequence modeling tasks.

The improved representational capacity and efficiency of xLSTM have positioned it as a promising model for various time-series applications, including speech recognition, biomedical signal analysis, and multimodal learning. In particular, for EEG-based modeling, xLSTM offers potential advantages in handling long-term temporal dependencies across brain signals while maintaining computational tractability. Its ability to revise stored information dynamically can be especially beneficial in non-stationary signal environments where neural dynamics evolve over time.

While ensemble methods such as Random Forests and Gradient Boosting have demonstrated strong predictive performance in many domains, their increasing structural complexity often diminishes interpretability. Similarly, deep learning architectures like xLSTM, despite their superior modeling capabilities, may also behave as "black-box" systems, making it difficult to trace the reasoning behind predictions. This lack of transparency presents challenges for deployment in sensitive or high-stakes applications, such as medical diagnosis or neurophysiological analysis, where accountability and trust are essential. To address these issues, the growing field of XAI seeks to bridge the gap between model performance and interpretability. The following section introduces the key concepts and representative methodologies in XAI, focusing on their application to EEG analysis.

3.2 Explainable AI

3.2.1 Core Concepts

Explainable Artificial Intelligence (XAI) encompasses methods and techniques designed to make AI systems' decisions and predictions understandable to humans [37]. As black-box AI models such as deep learning models, ensemble methods, and large language models become increasingly prevalent in high-stakes domains, the need for interpretable and transparent AI has grown accordingly [61]. High stakes domains such as healthcare, finance and criminal justice are in demand of interpretability and without adequate explanation mechanisms users may find it difficult to justify usage of automated decisions, without accountability and fairness, ethical concerns arise.

XAI approaches strive to balance providing human-understandable explanations with maintaining strong model performance [106]. Simpler and more transparent models often underperform compared to complex black-box systems. The central challenge is to maintain the strong performance of the opaque models that are used and get explanations in order to foster trust and adoption of AI technologies. Explainable Artificial Intelligence is a field that is currently evolving, with many researchers proposing a wide variety of approaches that aim to extract human-understandable explanations without significantly sacrificing model accuracy.

Broadly, XAI methods can be categorized along several dimensions: the interpretability approach (post-hoc versus ante-hoc), the explanation scope (global versus local), the explanation type (feature attributions, rules, examples, visualizations), and their dependency on the model architecture (model-agnostic versus model-specific) [13]. These dimensions help researchers understand the field of Explainable Artificial Intelligence, select the appropriate technique for their problem, and discover vacancies in the literature. Explainable Artificial Intelligence contributes to technical transparency, and adoption of automated systems that are ethically aligned with human values.

3.2.2 Categories

Explainable Artificial Intelligence (XAI) encompasses a diverse set of methods that can be categorized according to different conceptual dimensions. These dimensions determine how explanations are generated, what they reveal about the model, and how they relate to the underlying learning architecture. The following subsections present a detailed categorization along four principal axes: interpretability approach, explanation scope, explanation type, and dependency on model architecture. Understanding these categories is crucial for selecting appropriate XAI strategies depending on the nature of the model, the characteristics of the data, and the end-user requirements—particularly in domains such as EEG analysis, where interpretability is of paramount importance.

Interpretability Approach

A fundamental distinction in XAI lies between post-hoc and ante-hoc interpretability approaches. Post-hoc interpretability refers to techniques that are applied after a model has been trained, aiming to explain the behavior of otherwise opaque, complex, or black-box models. These methods attempt to rationalize a model's decision-making process by approximating or visualizing its internal logic without modifying the model itself. Popular post-hoc techniques include LIME [132], which approximates local decision boundaries around a given instance to produce human-understandable explanations. The versatility of post-hoc methods allows them to be integrated with a wide range of architectures, including deep neural networks, ensemble models, and multimodal systems. For example, a multimodal adaptation of LIME has been proposed in [149], demonstrating its flexibility in complex, multi-source data scenarios.

In contrast, ante-hoc (or intrinsic) interpretability focuses on designing models that are transparent by construction. In these models, interpretability is not an auxiliary component but an inherent characteristic. Examples include decision trees, rule-based systems, and generalized additive models, where the reasoning process is explicitly encoded in the model's structure. Ante-hoc methods, therefore, provide explanations that are exact rather than approximate, ensuring that every decision can be traced to a clear and logical rationale. This transparency is especially valuable in safety-critical or regulated environments, where decision accountability is required. Recent works such as [84, 93, 91, 94, 118] exemplify inherently interpretable approaches,

including symbolic and rule-based reasoning systems that prioritize explainability without compromising predictive capability.

While ante-hoc methods provide transparency by design, they often face trade-offs in flexibility and scalability compared to post-hoc techniques. As a result, contemporary research increasingly explores hybrid strategies that integrate the interpretability of ante-hoc frameworks with the expressive power of post-hoc analysis.

Explanation Scope

The scope of an explanation defines whether it targets the overall model behavior or specific individual predictions. Global explanations aim to provide insight into the general logic and structure governing model decisions across the entire dataset. These explanations reveal how the model processes information on average, identifying dominant features, recurring decision patterns, and relationships learned from the data. Techniques such as rule extraction, surrogate modeling, and hierarchical clustering are often used to generate global insights, helping to evaluate whether a model's decision boundaries align with known domain principles. Global explanations are particularly valuable for model validation, debugging, and regulatory assessment.

In contrast, local explanations concentrate on individual predictions or small subsets of data. Their purpose is to answer questions such as "Why did the model make this particular decision?" or "Which features were most influential for this instance?" Local interpretability techniques—like LIME [132] and SHAP—estimate the contribution of input features for specific outcomes, offering detailed and personalized insights. This granularity is especially crucial in medical applications, where understanding a single patient's prediction may be more important than comprehending the model's overall tendencies.

In practice, both global and local explanations are complementary. Global methods help ensure the model behaves sensibly across populations, while local methods provide transparency in critical or exceptional cases. Combining the two perspectives allows for a more holistic understanding of model reliability and ethical soundness.

Explanation Type

Explanations can also be classified according to the form or representation they take. Different explanation types communicate information at different cognitive levels, from quantitative attributions to qualitative reasoning structures.

Feature attribution methods quantify the importance or contribution of each input feature to the model's prediction [155]. These methods often employ gradient-based or perturbation-based computations to estimate how variations in individual features affect the output. They are particularly effective for high-dimensional data, such as EEG signals, where identifying relevant frequency bands or channels can provide valuable physiological insights.

Rule-based methods, on the other hand, generate interpretable logical or symbolic rules that summarize decision processes in human-readable form [171, 101, 86, 87]. These rules typically follow "if–then" structures, enabling clinicians or domain experts to trace decisions through explicit conditions. Rule extraction can be used both as an ante-hoc modeling strategy and as a post-hoc interpretation layer for more complex systems.

Example-based explanations use specific instances—either prototypical examples that represent typical model behavior or counterfactual examples that show how minimal changes to inputs could alter predictions—to help users understand the decision boundaries [76, 104]. Such methods are intuitive, allowing users to reason through comparisons rather than abstract metrics.

Lastly, visualization-based explanations present model reasoning through graphical or spatial representations. These can include saliency maps, activation maps, or dimensionality-reduced embeddings that intuitively display how features or components influence model decisions [106]. Visual explanations are especially powerful for image, signal, or spatially structured data, making them valuable tools for EEG analysis, where patterns across channels and time windows must be interpreted in context.

The selection of an explanation type depends heavily on the application, user expertise, and the nature of the data. In clinical settings, rule-based or feature-level explanations are often preferred for traceability, while visual or example-based approaches are useful for exploratory or diagnostic interpretation.

Dependency on the Model Architecture

A final important categorization concerns the dependency of XAI methods on the underlying model architecture.

Model-agnostic methods treat the model as a black box, focusing solely on its inputs and outputs without requiring access to internal parameters or gradients. This approach allows for broad applicability across diverse models and learning paradigms [90]. Techniques such as SHAP and other perturbation-based analyses fall into this category. They enable consistent comparison of interpretability across models, making them particularly useful for benchmarking or ensemble scenarios where multiple architectures are evaluated simultaneously.

In contrast, model-specific methods leverage the internal structure or mathematical properties of a given model to generate more accurate and efficient explanations. For instance, gradient-based techniques designed for neural networks exploit differentiability to trace the influence of input features through the model's layers, resulting in precise and fine-grained attributions [142]. While model-specific methods often provide higher fidelity in their explanations, they lack generality, as their design is tightly coupled to the internal mechanics of a particular architecture.

In practical applications, the choice between model-agnostic and model-specific methods involves balancing generality and precision. Model-agnostic tools facilitate transparency across heterogeneous systems, while model-specific techniques allow for deeper insights into the functioning of specialized architectures. For EEG-based AI, where interpretability and physiological validity are equally crucial, hybrid strategies that combine both paradigms are increasingly explored, offering a flexible yet rigorous framework for understanding model decisions.

3.2.3 Technical foundations of commonly used XAI techniques

In this subsection, we provide the technical foundations of widely used XAI methods in EEG analysis. Theoretical formulations and key equations are analyzed giving the necessary background information. The analyzed techniques are the most commonly used techniques when it comes to XAI in EEG analysis.

SHapley Additive exPlanations

SHAP values are a unified measure of feature importance and are a game-theoretic method that studies how different players cooperate or compete with each other [90]. To evaluate an instance, each attribute is assigned a SHAP value, which indicates the relative importance of the attribute to the model's decision-making process. The formal definition is as follows:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|(M-|z'|-1)!}{M!} \left[f_x(z') - f_x(z' \setminus \{i\}) \right]$$

where x is the instance to be explained, f is the model, i is the feature to be evaluated, and M is the number of features. Additionally, x' contains all possible perturbations of x.

Local Interpretable Model-agnostic Explanations

LIME is a local model-agnostic technique that approximates locally any classification model using an interpretable model proposed here [132]. The overall goal of LIME is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier. The explanation produced by LIME is obtained by the following:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where x is the instance to be explained, ξ is the instance explanation, g is a potentially interpretable model such as a linear model or decision tree, and f is the classification model. The function L measures the approximation of g to f in the locality defined by πx. The complexity of g is measured by $\Omega(g)$; this parameter is related to the complexity of the model g. LIME ignores the process within the model and makes explanations absolutely on the data level. Therefore, the explainer explains predictions on tabular data by perturbing features based on the statistical properties of the training data.

Gradient-weighted Class Activation Mapping

Grad-CAM is designed for CNNs and uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest [142]. The Grad-CAM method computes the importance weight α_c^k for the k-th feature map with class c in the last convolutional layer as follows:

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial S_c(x)}{\partial f_k(x)_{i,j}}$$

where Z is the total number of spatial locations in the feature map, and $\frac{\partial S_c(x)}{\partial f_k(x)_{i,j}}$ denotes the gradient of the class score $S_c(x)$ with respect to the activation $f_k(x)_{i,j}$ at the spatial location (i, j). The Grad-CAM heatmap for class c can then be computed as

$$M_c^{\text{Grad}}(x) = \text{ReLU}\left(\sum_k \alpha_c^k \cdot f_k(x)\right)$$

where ReLU is the rectified linear unit function, ensuring that only positive contributions are considered.

Deep Learning Important FeaTures

DeepLIFT was proposed as a recursive prediction explanation method for deep learning [146, 147]. DeepLIFT is designed to obtain the importance of input in the prediction of CNNs models. DeepLIFT uses multipliers that represent a slope describing how the outputs are changed when the inputs are different from reference data.
DeepLIFT uses a "summation-to-delta" property that states:

$$\sum_{i=1}^{n} C_{\Delta x_i, \Delta o} = \Delta o$$

where $o = f(x)$ is the model output, $\Delta o = f(x) - f(r)$, $\Delta x_i = x_i - r_i$, and r is the reference input. DeepLIFT is another additive feature attribution method.

Layer-wise Relevance Propagation

The layer-wise relevance propagation is a backpropagation-based method for interpreting the predictions of deep metworks [15]. As shown in [147] LRP is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero. Thus, $x = h_x(x')$ converts binary values into the original input space, where 1 means that an input takes its original value, and 0 means an input takes the 0 value. LRP like DeepLIFT is an additive feature attribution method.

3.2.4 Evaluation of XAI methods

In explainability, there is no one explanation to rule them all [100]. We need to define the objectives before selecting the proper desired explanation characteristics [99], and we need to have a consistent way to assess the quality and fit of XAI methods and explanations to our objectives and tasks. Evaluating XAI methods remains a complex yet essential endeavor. Effective evaluation must balance the explanation's fidelity to the model's true reasoning with its interpretability and usefulness to human users [37, 65, 95]. The lack of a gold standard for explanations poses a challenge for their evaluation, as there is no reference to be used as ground truth, with recent works trying to address this limitation [57]. Quantitative metrics such as faithfulness, completeness, and robustness provide objective measures of how well explanations reflect model behavior [173], while qualitative assessments—often involving user studies—assess trust, usability, and cognitive plausibility [111, 92]. This dual requirement is especially critical in EEG applications, where explanations must be not only technically sound but also neuroscientifically meaningful [137]. Therefore, principled evaluation frameworks are necessary to select XAI techniques that enhance both transparency and practical utility in neuroimaging contexts.

Table 3.1: Datasets used in EEG analysis with XAI insights.

Task	Dataset	Physiological Signal
Sleep Monitoring	Sleep-EDF [60, 73],	EEG, EMG, EOG
	Sleep Cassette [60, 73]	EEG, EOG, EMG
	SHHS [126, 60]	EEG, EMG, ECG, EOG
	CHAT [131, 98]	n/a
Emotion Recognition	SEED [177]	EEG
	DEAP [77]	EEG, EMG, EOG, BVP
	DENS [14]	EEG
Major Depressive Disorder	HUSM [108]	EEG
	MDD [107]	EEG
	MODMA [26, 145]	EEG
Seizure Detection	HUH [150]	EEG
	CHB-MIT [160]	EEG
	TUH corpus [112]	EEG
	Bonn [10]	EEG
	Siena [34]	EEG
	UBMC [167]	EEG
	UCI-EEG [3]	EEG
Motor Imagery	BCI IV 2a [157]	EEG, EOG
	BCI IV 2b [157]	EEG, EOG
	BCI III IVa [36]	EEG
	EEGMIMID [60]	EEG, EMG, EOG
	Stieger2021 [151]	EEG
Stroke prediction	Acute [8]	EEG
Schizophrenia	UNM [153]	FMRI, SMRI, EEG
	IBIB PAN [113]	EEG

3.3 EEG Analysis and XAI

3.3.1 Applications

Electroencephalography (EEG) has emerged as a versatile tool in both clinical and research settings, offering valuable insights into brain activity across diverse applications. Electroencephalography (EEG) remains a primary diagnostic tool for brain-related conditions due to its non-invasive nature and ease of use. The applications of EEG span from medical diagnosis to human-computer interaction, including sleep monitoring for stage assessment and disorder detection [133, 1, 164], seizure identification [133], brain-computer interfaces like the P300 speller [78], emotion recognition [174, 68, 124, 30], stroke rehabilitation [143], schizophrenia analysis [31], epilepsy management [27], and mental fatigue assessment [181].

3.3.2 Datasets

The research community has established several benchmark datasets for these applications, facilitating reproducible research and meaningful performance comparisons. The works gathered in this survey employ several key EEG datasets spanning diverse applications. Sleep monitoring studies frequently use Sleep-EDF and SHHS for sleep stage classification and apnea detection [60, 126]. Emotion recognition relies on SEED and DEAP to analyze emotional states through EEG patterns [177, 77]. Major depressive disorder datasets such as HUSM and MODMA provide EEG data for mood disorder analysis [108, 26]. Seizure detection utilizes CHB-MIT, TUH corpus, and Bonn datasets for epileptic event identification and seizure classification [112, 10]. Motor imagery research uses BCI Competition datasets for neural signal decoding in brain-computer interfaces [157]. Stroke prediction is supported by the Acute dataset for prognosis modeling [8], while schizophrenia investigations rely on the UNM and IBIB PAN datasets combining EEG and neuroimaging modalities [153, 113]. Table 3.1 summarizes these datasets and their physiological signals.

3.3.3 Challenges

The analysis of these EEG datasets has evolved from traditional machine learning approaches to sophisticated deep learning architectures. While classical machine learning methods [135] offer interpretable solutions through handcrafted features, modern deep learning techniques have demonstrated superior performance in automatic feature extraction and pattern recognition. Convolutional Neural Networks (CNNs) [162, 96] excel at spatial feature extraction, while Recurrent Neural Networks (RNNs) [123] and Long Short-Term Memory (LSTM) networks [75] effectively capture temporal dependencies in EEG signals. Recent advances include hybrid architectures [180] that combine multiple approaches, and transformer-based models [79] that leverage self-attention mechanisms for modeling long-range dependencies. Foundation models [32] represent the latest development, aiming to provide transferable EEG representations across multiple tasks.

However, despite their impressive performance, these advanced models often operate as black boxes, raising concerns about their interpretability and transparency. This limitation is particularly critical in medical applications where understanding the decision-making process is crucial for clinical adoption. This challenge underscores the growing importance of explainable AI techniques in EEG analysis.

3.3.4 XAI Methods in EEG

Explainable Artificial Intelligence in the analysis of EEG signal is a currently evolving field. As we can see in 3.3.1 the number of works that use XAI techniques in EEG analysis is evolving exponentially. Healthcare is a critical section and the usage of automated systems without explanations raises both clinical and ethical concerns. Therefore, it is reasonable to witness this exponential growth in research focusing on the explainability of EEG analysis. There is a great variety of methods applied in the field and a survey and taxonomy of these methods would help researchers mainly in two ways. Firstly, future researchers could discover the appropriate XAI technique for their problem. Secondly, vacancies in the field could be easily discovered.

This section analyzes various XAI methods applied to EEG analysis, categorizing them by explanation approach, scope, model dependency, and explanation type. We cover post-hoc methods such as model distillation and backpropagation, as well as interpretable-by-design models. We distinguish between global and local explanations, model-agnostic and model-specific methods, and explore explanation types including feature attribution, rule-based approaches, and visualization techniques, all within the context of EEG applications. The taxonomy of the XAI approaches in EEG analysis can be seen in 3.3.2.

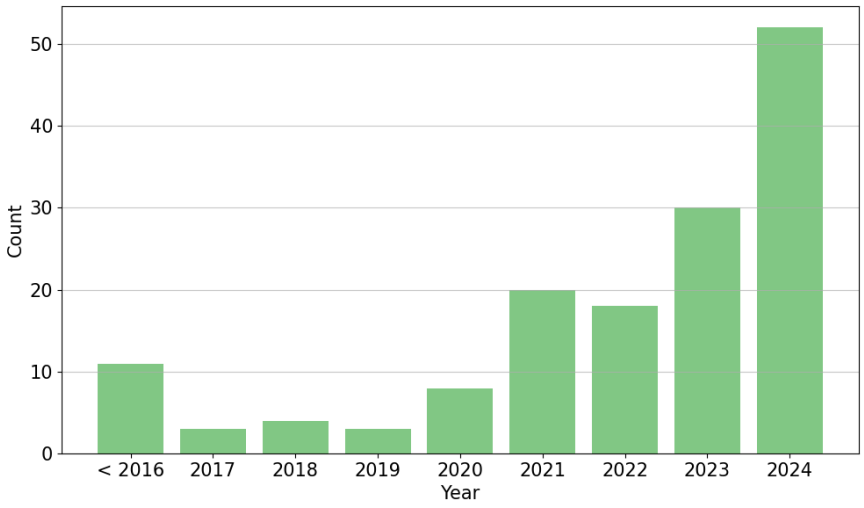

Figure 3.3.1: Research paper distribution by year.

By Explanation Approach

Machine learning interpretability methods can be broadly split into two categories: interpretable models and post-hoc interpretation techniques. The former focuses on building simpler, inherently understandable

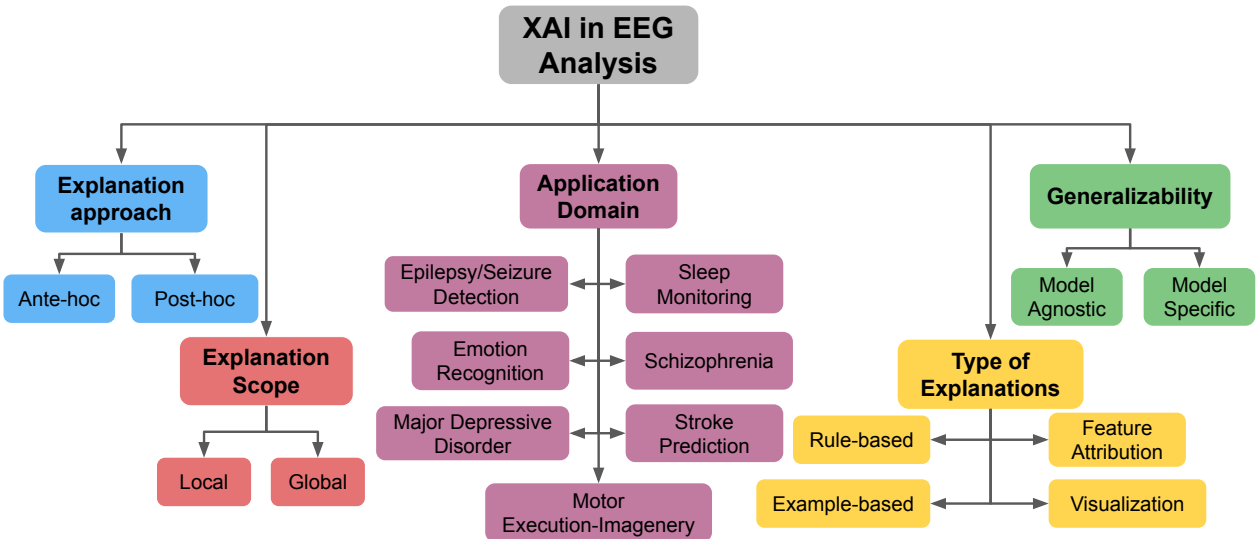

Figure 3.3.2: Taxonomy of XAI Approaches in EEG Analysis.

models, while the latter derives explanations by analyzing a model's behavior after it has been trained. Of the analyzed papers, only five used interpretable models and are discussed in the ante-hoc section. The remaining works used Post-hoc methods and some representative ones are analyzed.

Post-hoc Methods Post hoc interpretability applies interpretation methods after a model has been trained [105]. This approach can be further categorized into three main techniques: model distillation, backpropagation-based methods, and perturbation-based methods.

Model Distillation This method aims to approximate deep learning (DL) models using simpler models that replicate the input-output behavior of the original DL model. By interpreting these simplified models, insights can be gained into the functioning of the more complex model [154]. LIME is a very common distillation method that approximates locally the model using an interpretable model. Study [66] uses LIME in the context of human activity recognition, in order to understand the prediction performance and and the individual role of EEG features in detecting human activities. Study [62] also employs LIME at the end of the proposed pipeline providing insights into the individual contributions of the features in the predictions made by the model.

Backpropagation In backpropagation-based methods the gradient/relevancy score for a particular class or neuron is back-propagated in some form. Commonly used backpropagation techniques in EEG analysis are LRP and Grad-CAM. Studies [152, 46, 139, 50] employ LRP for explainability. More specifically, [152] was the first to propose the application of DNNs with LRP for EEG data analysis. With LRP transforms the single-trial DNN decisions into heatmaps indicating each data point's relevance for the outcome of the decision. Study [46] used LRP to explain the importance of modalities both locally and globally. Study [139] computed LRP for spectral and spatial importance. Study [50] proposed the use of LRP for explainable multiclass classification of neural states. Studies [88, 163] employ Grad-CAM for explainability. More specifically, in [88] Grad-CAM is employed for spatial explainability in the context of ADHD and CD characterization. In study [163] Grad-CAM is used to highlight EEG features associated with each sleep stage. Study [11] investigates five different backpropagation-based methods. The investigated methods were: 1) Saliency, 2) Guided Backpropagation, 3) LRP, 4) Integrated Gradients, 5) DeepLIFT. Furthermore study [159] evaluates four backpropagation-based methods in the context of autism. The evaluated methods were: Smooth-Grad, Smooth-Grad Squared, PatternNet, LRP.

Perturbation-based Perturbation-based methods aim at explaining the model by modifying the input of a model and observing the changes in the output. Commonly used perturbation-based methods are occlusions,

ablations and techniques like SHAP and LIME. EasyPEASI [110] proposes a perturbation EEG algorithm for spectral importance, that requires only perturbations to input data. Study [51] perturbation-based explainability is applied to gain insight into the spectral and spatial features learned by two distinct deep learning models trained on raw EEG for schizophrenia diagnosis. In the first approach, they ablated individual EEG channels, and in the second they ablated specific frequency bands. Study [140] also used two approaches perturbation-based. In the first they ablated individual EEG channels for spatial explainability, and in the second they ablated individual frequency bands within each channel for spatial-spectral explainability. The second approach resembled the EasyPEASI spectral explainability method from the aforementioned paper [110]. Study [44] also achieves spectral explainability by combining a version of EasyPEASI [41] with a metric from [48].

Ante-hoc Interpretability by design refers to machine learning models that are considered interpretable due to their structure, such as short decision trees or sparse linear models [105]. Study [166], employs an interpretable model and more specifically a Random Forest model for schizophrenia diagnosis. Model features are generated from both Generalized Partial Directed Coherence (GPDC) and direct Directed Transfer Function (dDTF) connectivity measures. Study [144] employs a combination of 1D-CNN with LSTM to extract interpretable features and a Graph Convolutional Network (GCN) for comprehensive graph modeling of multi-channel EEG signals. Additionally, an EEG subgraph construction module is introduced, aimed to identify the most pertinent EEG subgraphs relevant to the recognition task. This approach enhances the model's performance and interpretability. Study [102] uses an interpretable SincNet-based neural network for emotion recognition. Study [67] presents a novel seizure detection framework. The framework leverages a variety of robust features extracted from the EEG data providing comprehensive information about the underlying characteristics of the EEG signals that enable the model to make more accurate and interpretable predictions. In the specific study the post-hoc method of SHAP values is also employed reinforcing interpretability. Lastly, study [161] proposes an explainable feature engineering (EFE) model.

By Explanation Scope

Studies that have both global and local explanations are the majority of the analyzed works. Almost the half analyzed works provide both type of explanations. The rest are almost equally divided between local and global explanations. In the next paragraphs are presented some representatives for each category.

Global Explanations Global interpretation methods explain the entire model behavior. This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures [105]. Study [49] applies an ablation-based approach for global explainability. The method gives insight into the importance of each modality to the identification of each sleep stage. Study [115] employs SHAP to understand the significance of brain regions during prediction. To obtain the global learning of the model correct predictions of a trained model were selected and fed into the SHAP.

Local Explanations Local interpretation methods explain an individual prediction. You can zoom in on a single instance and examine what the model predicts for this input, and explain why [105]. Study [41] proposes a novel local approach for spectral explainability. In addition, it uses the local approach to form a global estimate of spectral importance and compares the results to the existing global method that was proposed here [110]. Study [176] conducts leave-one-out cross-validation experiments to investigate the leaned attention topography of each subject. Study [156] presents SHERPA, a novel SHAP-based explainability technique, that finds relevant latency ranges and electrodes.

By Model Dependency

Studies that use model-specific techniques are slightly more than the studies with model-agnostic methods. In addition there are few that use more than one XAI techniques and some of them are model-specific and the others model-agnostic. In general, perturbation-based methods are more likely to be model-agnostic and backpropagation-based methods model-specific. In the next paragraphs are presented some representative studies for each category.

Model-agnostic Methods Model-agnostic interpretation tools can be used on any machine learning model and do not have access to model internals such as weights or structural information [105]. Commonly used model-agnostic methods are SHAP and LIME. Study [121] employs the model-agnostic technique SHAP in the context of motor-imagery. SHAP is applied to an EEG signal, the different parts of the EEG signal are the players, and the prediction is the outcome of their cooperation. Study [138] combines the two model-agnostic techniques, LIME and SHAP values in seizure detection to estimate the importance value of each EEG channel. Study [109] employs SHAP in the context of seizure detection, and a variety of charts depicting the SHAP values were used to see how input features and model output relate to one another.

Model-specific Methods Model-specific methods are confined to particular model classes and work only for interpreting specific models [105]. Commonly used techniques in EEG analysis include LRP, Grad-CAM, DeepLIFT, and Integrated Gradients [158]. Study [45] presents a post hoc statistical analysis of filter activations, while [21] uses a gradient-based technique plus temporal and spatial kernel visualizations to reveal class-specific features. Study [54] combines LRP with perturbation for insights into channel importance, frequency-band contributions, and channel interactions, and [128] applies saliency maps and integrated gradient to highlight functional connectivity. In [178], Interpretability-guided Channel Selection (ICS) leverages CAM to identify high-contributing EEG channels. Finally, [53] combines LRP with perturbations to explore spatial, spatio-spectral, and temporal importance.

By Explanation Type

Studies with visualization techniques and feature attribution are the majority. The few that are Rule-based and Example-based are presented in following subsections. In the subsections of feature attribution and visualization are presented representative works of each category.

Feature Attribution Study [167] uses SHAP values in the context of seizure detection, quantitatively attributing importance to individual features regarding their contribution to a model's predictions. Study [4] also employs SHAP on Epilepsy Diagnosis. With SHAP values they interpret the model's decision-making process and identify the best feature contribution or feature importance. Study [22] employs both SHAP and LIME to identify key features that significantly contribute to the prediction. This identification reduces the feature space, saving this way time during model training and improving accuracy. Study [74] also employs both SHAP and LIME for feature importance. Study [97] performs a feature study on the output predictions which is based in input-based explanation drivers methods. Study [12] applies SHAP as an explanatory mechanism to identify the most relevant characteristics to predict schizophrenia. Study [103] uses the GNNEXPLAINER framework that assesses feature importance. Study [89] employs SHAP in epileptic seizure recognition getting for each feature the importance and also their range of effects over the data set. Study [16] also uses SHAP to get feature importance in the classification of emotional states. Study [80] relies on a human-centric approach for explainability. After the analysis, they can validate that the model actually looks at the important features for severity scoring.

Rule-based Explanations Study [168] employs four different rule-based classifiers in the task of epilepsy detection. More specifically, the classifiers are a Decision tree, a Random forest, an SVM combined with C4.5 and an SVM combined with a Random forest. Random forest was the model with the best performance and exhibiting also higher interpretabilty.

Example-based Methods Example-based explanation methods select particular instances of the dataset to explain the behavior of machine learning models or to explain the underlying data distribution [105]. Study [47] presents a novel Frequency-based Activation Maximization Explainability (FAME) approach that falls into this category of explanations. In greater detail, this new approach is suited for long-time series and can produce a sample that is representative of the features learned by the classifier for a particular class.

Visualization Techniques Visualization methods highlight the most influential features in the input that drive a model's decision [130]. Several studies [129, 71, 39, 175] employ Grad-CAM: [129] visualizes spatial EEG-channel relevance and temporal data, [71] identifies EEG features for outcome classification and network failures, [39] highlights signal parts critical for sleep-stage prediction, and [174] uncovers which temporal

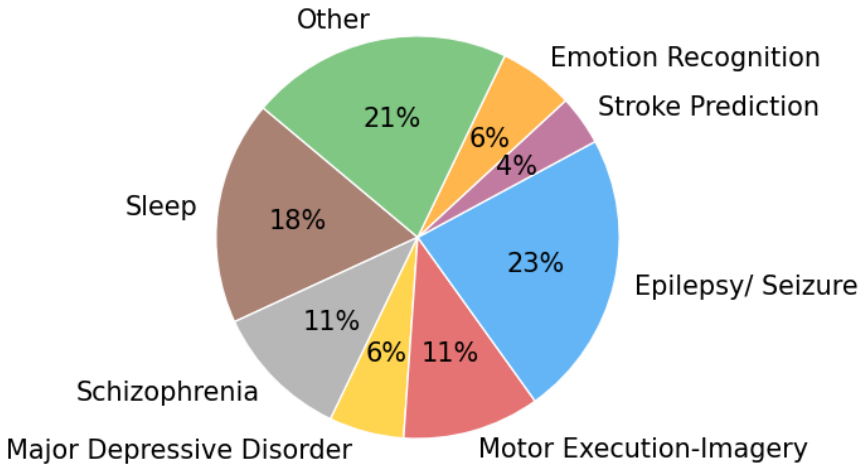

Figure 3.3.3: Distribution of EEG study applications from our analyzed papers.

segments and channels are most relevant to a CNN model. Study [42] proposes a novel visualization-based approach revealing global insights into learned waveforms, spectral features, and filter importance, leveraging LRP and filter perturbation. Study [141] introduces methods for visualizing ConvNet-learned features, while [82] demonstrates three approaches for interpreting a trained EEGNet (hidden-unit activation summaries, convolutional kernel weights, and single-trial feature relevance). In [127], SHAP Deep and SHAP Gradient visualize electrode, spectral, and temporal contributions. Study [43] presents a novel model visualization-based approach that adapts a CNN architecture to boost interpretability, and [2] employs a SHAP-based framework to generate visual explanations and identify critical EEG features for epileptic seizure detection.

By Application Domain

Application Domains of studies with XAI in EEG analysis vary, with epilepsy / seizure detection being the most common one. In 3.3.3 is presented a pie chart with the percentage of each domain. For each domain is presented one representative work.

Epilepsy/Seizure Detection Epilepsy detection through EEG traditionally requires time-consuming manual analysis. Although automated methods exist [138], their black-box nature limits interpretability. Study [172] addresses this using LIME and SHAP for feature importance analysis.

Sleep Monitoring Sleep specialists typically perform visual sleep stage scoring by analyzing patients' neurophysiological EEG signals recorded in sleep laboratories. This process is often challenging, labor-intensive, and demands significant time and human resources [9]. However, for models to be employed as an assistive solution by sleep specialists, it is essential for the models to be explainable [39]. A study that employs XAI techniques to sleep monitoring is [119]. More specifically, it employs three different Post-hoc explainability methods in the context of automated sleep scoring. The different techniques include frequency-domain occlusion, time-domain occlusion, and pattern visualization of temporal filters in the CNN.

Schizophrenia Schizophrenia is a mental disorder diagnosed based on variable symptoms, making it difficult to diagnose accurately [52]. Recent studies have applied deep learning to EEG for automated schizophrenia diagnosis, but the use of raw EEG data complicates model explainability. This lack of explainability is a challenge in healthcare, where understanding model decisions is essential. Study [52] examines the reproducibility of schizophrenia biomarkers across models with the goal of identifying those that have potential clinical implications. For explainability, they used a permutation feature approach. The first analysis was with permutation of the features associated with each canonical frequency band for spectral importance, and in the second the features associated with each channel for spatial importance.

Motor Execution-Imagery XAI methods have been applied to the domain of cognitive neuroscience and in greater detail with motor imagery and motor execution. For instance, study [117] at first applied a permutation-based method for reliable explanations. Next, a novel technique was designed for selecting the best among a few saliency-based methods due to the need for a faster method of XAI for spatio-temporal explanation. The tested methods were saliency maps, DeepLIFT and DeepSHAP, with DeepLIFT being selected.

Major Depressive Disorder Depression is a prevalent mental disorder that poses significant risks to human health and social stability [144]. Current approaches to diagnosing depression primarily depend on patient self-reports and psychiatric evaluations, making them vulnerable to subjective influences and increasing the risk of misdiagnosis or overlooked cases. Deep learning techniques are being adopted, though explainability is poor in these techniques. Study [55] identifies biomarkers of Major Depressive Disorder (MDD), by two novel convolutional neural network-based architectures. For explainability employs a variety different of approaches. The two first are model-agnostic, ablation-based for spatial and spectral explainability. The other are visualization techniques, model-specific and are uniquely enabled by the two novel model architectures.

Stroke Prediction Acute ischemic stroke is one of the leading causes of neurological disease in the elderly, exposing millions of individuals to neurological abnormalities and physical impairments [69]. Study [23] employed the XAI tools LIME and Eli5. The use of these tools made possible the investigation of how the model makes the predictions and how each input attribute influences the prediction. By identifying the key features that significantly contribute to the prediction, the feature space was reduced and time was saved during model training while accuracy was improved.

Emotion Recognition XAI methods have also been applied to the domain of emotion recognition. For instance, study [7] integrates the SHAP Deep Explainer in the emotion classification process. A spectrogram is passed into the SHAP Deep Explainer and at the end, the discrete-time frames corresponding to the output of a particular emotion are obtained.

Other Fourteen of the reviewed studies do not belong to any of the aforementioned domains. For instance, study [83] proposes a novel method that identifies channel importance regardless of the type of EEG application. Study [6] employs XAI methods in the context of industrial internal security. More specifically, it employs permutation to the AdaBoostClassifier, permutation, and SHAP to Random Forest model, and permutation to K-Nearest Neighbors.In Figure 3.3.3, the distribution of different applications is shown.

Chapter 4

Proposal

In this section, we present the followed methodologies for both the survey of XAI methods in EEG and the Seizure detection problem. The results of the survey were presented in 3.3.4. For the Seizure detection problem the models are trained with features extracted from EEG, ECG and EMG signals. Following, we employ two different post-hoc explainability methods to understand model decisions.

We first highlight the main contributions of this thesis and then explain the followed methodologies in detail.

4.1 Contributions

The contributions of this dissertation are multifaceted, spanning both theoretical and methodological advancements in the field of explainable artificial intelligence (XAI) applied to electroencephalography (EEG) analysis.

First, this work undertakes a comprehensive exploration of existing XAI techniques used in EEG research. By systematically examining the current state of explainability frameworks, visualization tools, and algorithmic strategies, the study offers a clear and structured understanding of how interpretability is currently being approached in EEG-based machine learning. Through this analysis, the dissertation not only consolidates dispersed knowledge across studies but also identifies key methodological limitations and open research gaps. This contribution serves as a valuable reference point for researchers aiming to develop more transparent, trustworthy, and clinically applicable EEG models.

Second, the dissertation provides an in-depth review and taxonomy of recent trends and developments in explainable EEG analysis. It presents a detailed overview of fundamental EEG tasks—such as classification, event detection, and brain state decoding—while also surveying the most relevant publicly available EEG datasets used in XAI studies. To bring conceptual clarity to this growing research area, a structured classification of XAI methods is proposed, organizing techniques according to factors such as their model dependence, level of interpretability, and granularity of explanation. This taxonomy helps contextualize current research efforts and supports systematic comparison across studies.

Third, the dissertation introduces a novel and interpretable multimodal framework for seizure detection that integrates EEG, electrocardiography (ECG), and electromyography (EMG) signals. The proposed methodology combines advanced preprocessing pipelines and feature extraction techniques with explainable machine learning models capable of providing transparent decision-making. Experimental results demonstrate that the multimodal approach achieves superior performance compared to conventional deep learning methods, highlighting its potential to improve both diagnostic accuracy and model interpretability in clinical contexts.

Finally, the study leverages explainability analysis to identify the most salient features contributing to seizure detection. By employing feature-importance and attribution techniques, it isolates the neural and physiological characteristics most influential in classification outcomes. This interpretability-driven insight not only enhances understanding of the mechanisms underlying seizure generation but also offers valuable guidance for clinicians and researchers seeking to refine diagnostic tools or design future multimodal studies.

4.2 Followed Methodologies

4.2.1 Survey of XAI methods in EEG

In this section, we present the methodology used to conduct a thorough and systematic review of the literature on applying XAI to EEG analysis. We begin with the research questions that guided our investigation, then describe the paper selection strategy, and conclude with the framework used for analyzing the collected studies.

Research Questions

The focus of this study is to investigate the role of XAI in EEG analysis by addressing key research questions that explore the methods, applications, and broader implications of XAI in this domain. The following research questions guide the survey:

RQ1: What are the primary XAI methods utilized in EEG analysis? This question aims to identify and categorize the explainability techniques applied to EEG-based studies. By examining these methods, the study seeks to provide a comprehensive overview of how explainability is incorporated into EEG research and which approaches are most prevalent.

RQ2: In which specific EEG analysis tasks have XAI methods been applied? EEG analysis encompasses a diverse range of applications, including seizure detection, sleep stage classification, cognitive workload assessment, and emotion recognition. This research question investigates the various tasks where XAI techniques have been implemented, highlighting trends, challenges, and gaps in the existing literature.

Through the answers to these research questions and our overall analysis, we also want to present what types of limitations exist in the employment of XAI techniques in the EEG domain and what further approaches could provide benefits for future research.

Search Strategy

Our search strategy followed two parallel approaches to compile an initial pool of papers, which were then filtered on the title and abstract relevance to produce the final set studied in this survey, as illustrated in Figure 4.2.1. The collection of the initial pool of papers involved (1) targeted searches within selected venues and (2) broad keyword-based queries in major academic databases, along with backward and forward snowballing to detect related papers. The initial collection of 151 papers underwent a filtering process based on titles and abstracts, ultimately yielding a final selection of 66 relevant papers.

We began by searching the proceedings of the World Conference on Explainable Artificial Intelligence with "EEG" and "electroencephalography" and used backward and forward snowballing to expand our initial set of papers. This process revealed additional key venues—such as IEEE ISBI, IEEE EMBC, IEEE BIBM, and IEEE NER—publishing work at the intersection of XAI and EEG. For each venue, we systematically searched publications using XAI- and EEG-related keywords and again applied snowballing to enlarge our dataset.

The second approach involved querying major academic databases, including PubMed, ScienceDirect, IEEE Xplore, Springer, and Google Scholar. We used the same set of XAI- and EEG-related keywords to retrieve relevant publications. As with the first approach, we performed snowballing to identify additional papers from the reference lists and citations of retrieved articles.

After compiling an initial pool of 151 papers, we conducted a filtering process based on titles and abstracts to refine the selection. We prioritized studies that closely aligned with our research focus, ensuring a broad yet relevant coverage of the intersection between XAI and EEG. Additionally, we favored papers that provided clear experimental validation or novel methodological contributions, which helped us focus on high-quality and impactful research.

4.2.2 Seizure detection

The methodology adopted begins with the SeizeIT2 dataset [20] as input. This dataset was chosen as it served as the basis for the Una Europa Epilepsy Data Challenge, providing a well-established benchmark for

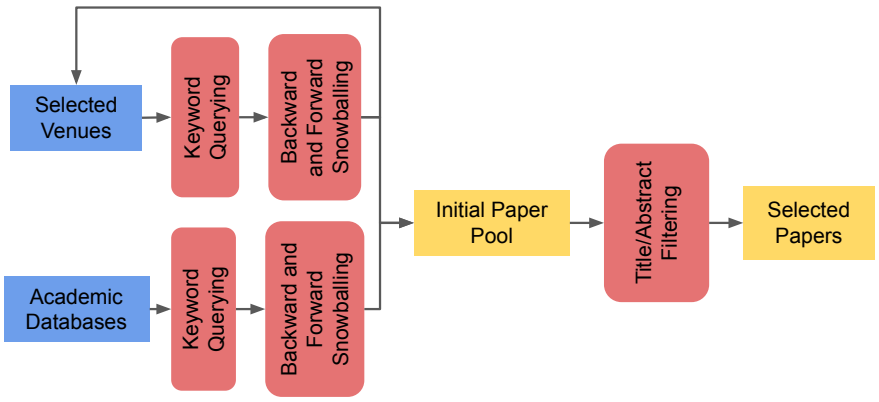

Figure 4.2.1: Methodology for identification of papers.

seizure detection research. In addition, it is a large-scale dataset comprising thousands of hours of multimodal recordings, including not only EEG but also EMG and ECG signals. Its large scale and multimodal structure also make it particularly suitable for building robust and generalizable models.

The next stage of the pipeline consists of preprocessing and feature extraction, which prepare the data for model training. This stage is crucial because raw signals are often noisy and, without adequate processing, they can obscure rather than reveal meaningful patterns. During preprocessing, the continuous recordings are segmented into fixed-length windows and filtered to remove noise and artifacts, ensuring that subsequent analysis focuses on physiologically meaningful signal components. Feature extraction is then applied across all three modalities (EEG, ECG, and EMG), resulting in a rich representation of the data. The extracted features encompass statistical measures, temporal characteristics, signal complexity measures, and spectral features derived from different frequency bands.

Once the data has been prepared, the extracted features are used to train and evaluate machine learning models. A variety of classifiers are explored to identify which approaches strike the best balance between predictive performance and interpretability. In particular, we focus on tree-based and ensemble-based methods such as Random Forest, XGBoost, LightGBM, CatBoost, and HistGradientBoosting. These models are well suited for tabular, multimodal data and have the advantage of handling nonlinear interactions and heterogeneous feature spaces. Beyond raw performance, they also lend themselves more naturally to interpretability techniques, making them an ideal choice for a clinical use case in which black-box predictions would not be acceptable. The emphasis here is not only on achieving competitive detection accuracy, but also on developing models that could be trusted and scrutinized by clinicians in real-world scenarios.

Finally, the workflow concludes with an explainability analysis, where we apply two complementary post-hoc techniques to interpret the trained models. First, we employ SHAP in all models, generating SHAP beeswarm plots that provide a global view of feature importance as well as the direction and magnitude of the contribution of each feature to the predictions. This allows us to identify not only which features are most influential, but also how their values affect the likelihood of detecting a seizure. In addition, we apply TE2Rules to the XGBoost and Random Forest models, extracting human-readable decision rules from these classifiers.

In summary, the methodology is designed not only to build accurate seizure detection models, but also to emphasize transparency and interpretability throughout the workflow. By combining benchmark data, multimodal feature representations, ensemble-based classifiers, and post-hoc interpretability tools, the approach balances predictive power with clinical relevance. A detailed description of each stage of the methodology, along with experimental configurations and results, is provided in Chapter 5.

Chapter 5

Experiments

Contents

5.1 Preliminaries

5.1.1 Dataset

SeizeIT2 Dataset

The models in this dissertation are trained and evaluated using the SeizeIT2 dataset [20]. The dataset includes recordings from 125 patients (51 female, 41%), encompassing approximately 11640 hours of wearable data, acquired across five distinct European Epilepsy Monitoring Units: University Hospital Leuven (Belgium), Freiburg University Medical Center (Germany), RWTH University of Aachen (Germany), Karolinska University Hospital (Sweden) and Coimbra University Hospital (Portugal). The University Hospital Leuven was the only center that enrolled pediatric patients. The dataset includes only data from patients with focal epilepsy who experienced one or more seizure episodes during the monitoring period. Four different modalities were recorded for most participants: bte-EEG, ECG, EMG and movement data. All participants' data within the dataset contain wearable bte-EEG. In 3% of the dataset, ECG, EMG and movement data were not included due to technical failures or errors in the setup. The dataset contains 886 recorded focal seizures with the wearable device. The mean duration of the recorded seizures was 58 seconds, ranging between 3 seconds and 16 minutes.

The SeizeIT2 project is an international multicenter dataset with more than 350 patients suffering from epilepsy and recorded both in home and hospital environments. The dataset is an open subset of the full project and was used for the Seizure Detection Challenge organized by KU Leuven in collaboration with Una Europa, aiming to develop innovative and robust machine learning (ML) frameworks for electroencephalography (EEG) data processing, in which the end use case is detection of epileptic seizures. According to the organizers' guidelines, the recordings from the first 96 subjects were designated for training, while the remaining subjects were used for evaluation. The final evaluation was carried out on a hidden test set.

For this application, it was not feasible to train on the entirety of the thousands of recorded hours, therefore, only a subset was used for training. In addition the dataset is very unbalanced containing few hours of seizure events compared to non seizure epochs. The constructed datasets included all periods identified as seizure events, along with randomly selected non-seizure periods. Three seizure-to-non-seizure ratios were considered: 1:2, 1:10, and 1:100. For the evaluation were used all the recordings that included every modality.

Data Augmentation

A frequent challenge in biomedical signal analysis, such as with EEG signals for seizure detection, is the limited quality and amount of available data [67]. In our dataset, the recordings originate from wearable devices, which provide lower-quality data, and the overall duration of seizure events is relatively limited. In this thesis, I experimented with datasets both with and without augmented seizure data.

The data augmentation technique I used was the same with the one used here [67].The mathematical technique Fourier Transform (FT) Surrogates was used, which transforms a function of time, such as an EEG signal, into a function of frequency. This transformed representation, or spectrum, offers an alternative view of the data and emphasizes different characteristics of the underlying signal. Fourier Transform (FT) surrogates are a particular form of data augmentation in which new surrogate signals are produced by randomizing the phases of the original EEG signal's Fourier Transform. Importantly, this process preserves the power spectrum, a measure of the energy distribution of the signal between frequencies, thus maintaining the overall structural properties of the original signal while modifying its temporal organization (such as the sequence and timing of events) [67].

The use of FT Surrogates served two primary purposes. Firstly, it increased the size of training datasets, creating more seizure samples. Secondly, this technique added diversity to the datasets, allowing the models to learn to recognize seizures in a broader range of circumstances.

Preprocessing and Feature Extraction

The preprocessing and feature extraction approaches used in this thesis are inspired by a study for Seizure detection from wearable data [67] and a study for prediction of post-traumatic stress disorder subtypes from Resting-state EEG [85].

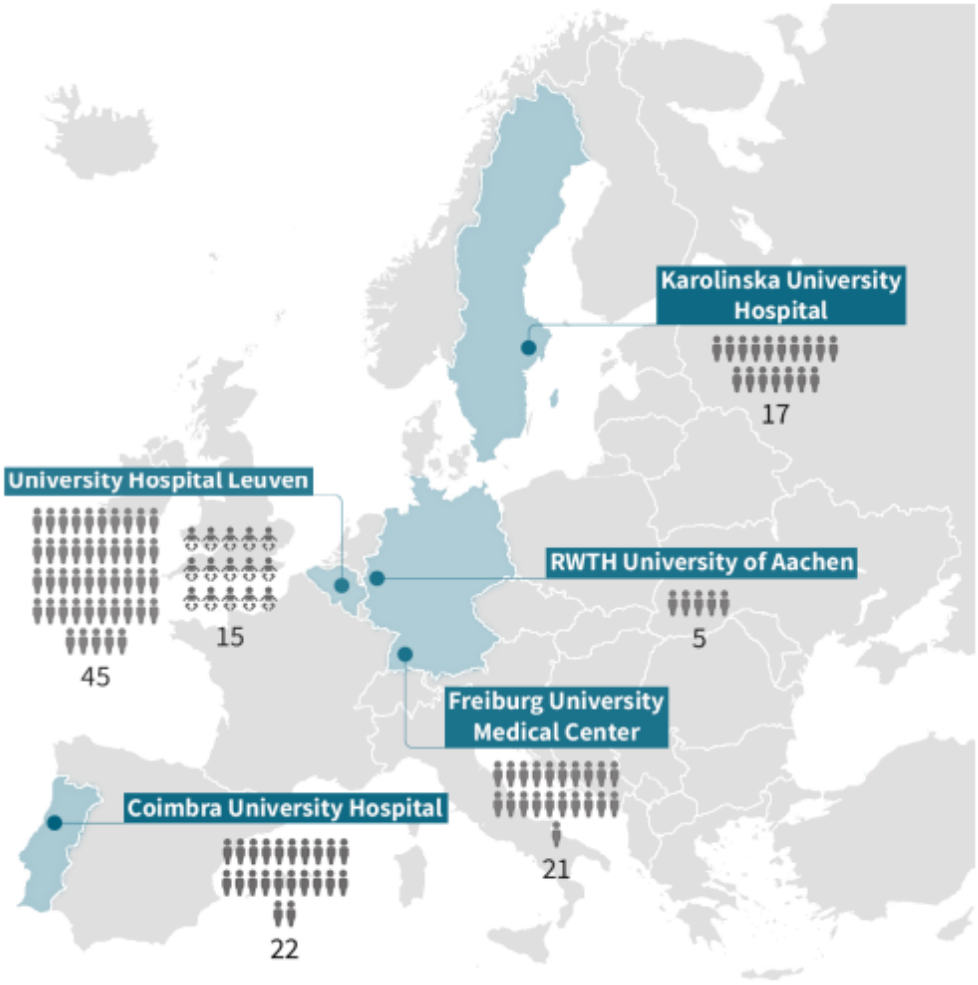

Figure 5.1.1: Number of participants per EMU included in the SeizeIT2

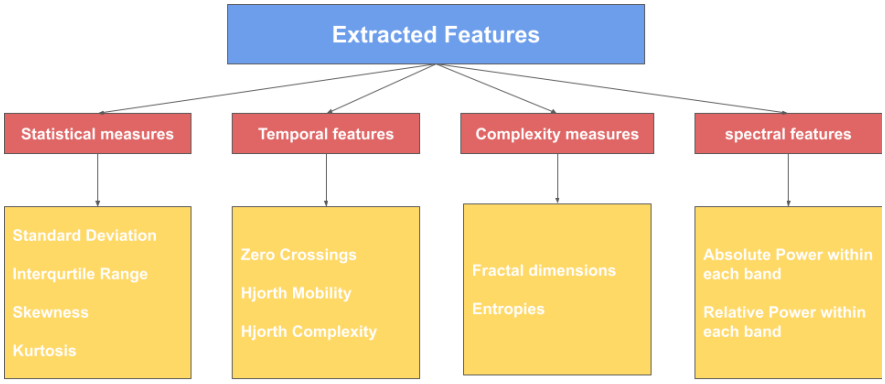

Figure 5.1.2: Extracted features

First, the EEG data were band-pass filtered between 1–100 Hz to retain the frequency components most relevant for seizure activity, while removing slow drifts and high-frequency noise. A 50 Hz notch filter was also applied to eliminate power-line interference commonly present in clinical recordings. After filtering,

the signals were segmented into non-overlapping 1-second epochs, providing consistent time windows for feature computation. Three seizure-to-non-seizure ratios (1:2, 1:10, and 1:100) were considered to address the inherent class imbalance, and for each ratio, experiments were conducted both with and without augmented seizure samples to assess the effect of data augmentation on model performance.

Following preprocessing, features were extracted from all modalities (EEG, ECG, and EMG) to create a rich representation of the data for machine learning models. The features were grouped into four main categories: statistical measures, temporal features, complexity measures, and spectral features. Together, these features capture complementary information on signal distribution, dynamics, nonlinearity, and frequency content. A summary of the extracted features is shown in Figure 5.2.1, with detailed descriptions provided in the following paragraphs.

Statistical measures used were Standard Deviation (STD), Interquartile Range (IQR), Skewness, and Kurtosis. STD and IQR are used to capture the variability within EEG signals. STD reflects the average deviation of the data from the mean, while IQR indicates the spread of the middle portion of the data. Skewness and kurtosis describe the characteristics of the signal's probability distribution: skewness measures its asymmetry, and kurtosis evaluates the weight of the distribution's tails compared to a normal distribution.

The temporal features analyzed include the number of zero crossings, Hjorth mobility, and Hjorth complexity. Zero crossings give an estimate of the signal's frequency content, while the Hjorth parameters serve as descriptive measures of signal characteristics. In particular, mobility reflects the average frequency or rate of change of the signal, whereas complexity indicates how closely the signal resembles a pure sine wave.

The complexity measures applied included fractal dimensions and entropies. Fractal dimensions capture how the level of detail in the data varies across different scales, while entropy quantifies the degree of randomness or unpredictability within the signal.

Spectral features, such as power, in different energy bands, such as Delta, Theta, Alpha, Beta, and Gamma were extracted. The EEG signal was divided into different frequency bands and absolute and relative power were measured within each band. The five canonical frequency bands: delta (1.25–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–49 Hz). Theta/beta ratio was also estimated.

The EMG signals were band-pass filtered between 20–450 Hz to remove motion artifacts and high-frequency noise, and a notch filter at 50 Hz was applied to eliminate power-line interference. The filtered signals were segmented into 1 s non-overlapping epochs for feature extraction.

Several widely used EMG descriptors were calculated. Root Mean Square (RMS) and Mean Absolute Value (MAV) summarize the signal amplitude and overall muscle activation level. Zero Crossing Rate (ZC) reflects the frequency content of the EMG activity, while Waveform Length (WL) quantifies the cumulative complexity of muscle activity over time. In addition, Standard Deviation (STD) and Variance capture the variability in the signal. Together, these features provide a compact characterization of EMG activity, capturing both intensity and dynamic properties of muscle contractions.

The ECG signals were preprocessed with a band-pass filter of 0.5–50 Hz to remove baseline drift and high-frequency noise, and a 50 Hz notch filter was applied to suppress power-line interference. The data were divided into 1 s epochs, from which a combination of statistical and physiological features were extracted.

Basic descriptors such as mean, Standard Deviation (STD), Variance, Root Mean Square (RMS), and Peak-to-Peak amplitude (PTP) capture overall signal morphology and variability. Higher-order statistics including skewness and kurtosis describe the distributional shape of the ECG waveform. Physiologically relevant features were derived through R-peak detection, enabling estimation of RR intervals and heart rate within the segment. From these, average RR interval, standard deviation of RR intervals, and instantaneous heart rate were computed, offering insight into short-term cardiac rhythm dynamics. These features together capture both morphological characteristics of the ECG waveform and temporal variability related to autonomic regulation.

5.1.2 Evaluation Metrics

The evaluation of our models was based on the scoring function that was also used in the Seizure Detection Challenge. The function is based on the metrics of sensitivity and False Alarm Ratio (FAR). These metrics

are particularly suitable for seizure detection tasks, which are highly imbalanced classification problems due to the rarity of seizure events compared to long periods of non-seizure activity.

Sensitivity

Sensitivity is measured on an event basis, meaning that the performance is assessed at the level of entire seizure events rather than individual prediction windows. This ensures that the clinical relevance of seizure detection is reflected in the evaluation, as missing an event is more critical than missing a single window. To compute sensitivity, the any-overlap method (OVLP) [179] is employed. According to OVLP, a True Positive (TP) is counted when the predicted hypothesis has any temporal overlap with a corresponding seizure event in the ground truth annotation. A False Negative (FN) occurs when no overlap exists. This method is permissive, as even partial overlaps count as successful detections, which typically results in higher sensitivities but underestimates the number of false detections.

False Alarm Ratio (FAR)

False alarms (FAs) correspond to spurious detections in which the model predicts a seizure event that does not overlap with any annotated seizure in the reference. Instead of reporting specificity, which is less informative in highly imbalanced datasets, the False Alarm Ratio (FAR) is used. FAR is computed as the number of false positives normalized by the recording duration and expressed as the number of false alarms per hour (FA/h). In our evaluation, FAs were computed using the epoch-based (EPOCH) scoring method [179]. In this method, both reference and hypothesis are discretized into non-overlapping epochs of fixed duration. Each epoch is assigned a seizure/non-seizure label, and errors are tabulated as insertions, deletions, or substitutions, with all errors weighted equally. This provides a more conservative estimate of false alarms compared to OVLP, reducing the risk of underreporting spurious detections.

Scoring Function

To balance the trade-off between high sensitivity and low false alarm rate, we adopted the combined scoring function defined in the challenge. Sensitivity is computed using the OVLP method, while FAR is estimated using the EPOCH-based approach. The final score is then calculated as a weighted combination of both metrics, with a weighting factor of 0.4 applied to the FAR to balance its influence relative to sensitivity. This results in a single performance score that rewards models capable of reliably detecting seizure events without producing excessive false alarms.

$$Score = \underbrace{Sensitivity(\%)}_{\text{OVLP}} - 0.4 * \underbrace{\frac{FA}{h}}_{\text{EPOCH}}$$

5.2 Model Experiments

All model training and evaluation were performed on an Amazon Web Services (AWS) EC2 instance equipped with a dedicated GPU to ensure sufficient computational resources for both machine learning and deep learning experiments. To maintain consistency across modalities, only recording runs containing the complete set of EEG, EMG, and ECG signals were used, thereby allowing the models to fully exploit multimodal information.

5.2.1 Machine Learning Models

The machine learning models were trained using the handcrafted feature sets described in Section 5.1. For each classifier, six variants were trained corresponding to different class ratios (seizure-to-non-seizure) and augmentation conditions. Specifically, for each ratio, one dataset contained only the original samples, while the other included synthetic data generated through augmentation. Multiple classification thresholds were tested for each configuration, and the threshold yielding the best overall performance was selected.

The models evaluated included XGBoost, LightGBM, CatBoost, Random Forest, and HistGradientBoosting, as well as two ensemble configurations. The first ensemble combined all five base models, whereas the second ensemble included only the top three performers: XGBoost, CatBoost, and LightGBM. Both ensembles

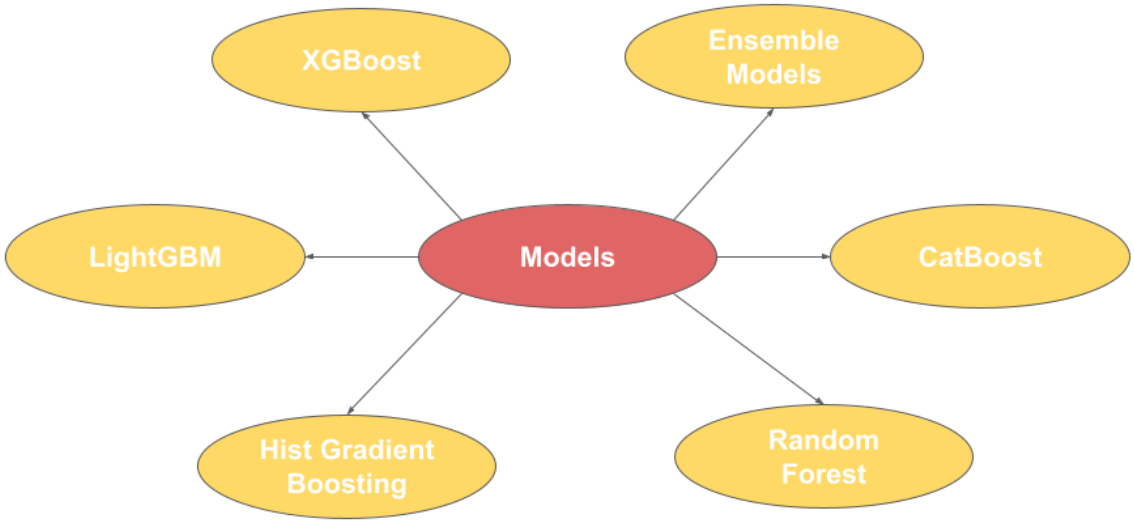

Figure 5.2.1: Overview of the machine learning models used in this study.

employed a **majority voting** mechanism: each model independently predicted a class label for a given input, and the label receiving the majority of votes was selected as the final decision. This strategy aimed to enhance generalization by leveraging model diversity, reducing variance, and mitigating the risk of overfitting to specific patterns or modalities.

In practice, the gradient boosting models (XGBoost, CatBoost, LightGBM) demonstrated strong adaptability to heterogeneous feature spaces, benefiting from their ability to model complex nonlinear dependencies while handling feature interactions automatically. Ensemble aggregation further improved robustness and interpretability, offering a principled way to consolidate multiple inductive biases inherent to different algorithms.

5.2.2 Deep Learning Models

In contrast to the machine learning approaches, the deep learning models were trained directly on the preprocessed multimodal time series, without reliance on handcrafted features. Although this approach was expected to capture temporal dependencies and nonlinear signal correlations more effectively, in practice the deep learning architectures underperformed compared to the feature-based machine learning pipelines. Their lower accuracy and generalization ability highlight the current challenges of applying generic deep models to limited or noisy biosignal datasets.

GDN

The Graph Deviation Network (GDN) proposed by [33] was initially designed for anomaly detection in multivariate industrial time series using a graph neural network (GNN) architecture. In our implementation, each channel from the EEG, ECG, EMG, and motion (MOV) signals was treated as a node in the GNN, allowing the model to learn inter-channel dependencies.

However, the GDN was not originally intended for biosignal analysis, and the low number of EEG channels (two per subject) limited the graph's representational richness. The resulting sparse connectivity hindered the model's ability to learn meaningful inter-node relationships. Consequently, despite its conceptual appeal for multimodal dependency modeling, the GDN proved unsuitable for our seizure detection task.

xLSTM

The xLSTM architecture, recently introduced by [18], extends the traditional Long Short-Term Memory (LSTM) design with improved gradient flow and modular flexibility. For this work, we employed the `xLSTMBlockStack` variant, optimized for non-language sequence modeling. The model was trained directly on the preprocessed EEG, ECG, and EMG sequences to learn temporal and cross-modal patterns.

Despite its promising architecture, the xLSTM exhibited significant overfitting, achieving high training accuracy but poor validation performance. This behavior suggests that, given the limited dataset size and signal variability, the model memorized training examples rather than learning generalizable temporal dynamics. Regularization and dropout adjustments offered marginal improvement, but overall, the model failed to generalize effectively to unseen samples, indicating that deep architectures may require substantially larger or more diverse datasets to outperform feature-based methods in this context.

5.2.3 Explainability

To interpret the internal decision processes of the trained models, two complementary explainability frameworks were employed: SHAP (SHapley Additive exPlanations) and TE2Rules [81]. These methods provided both feature-level and rule-level insights into model behavior, enhancing transparency and interpretability.

The SHAP library quantifies the contribution of each input feature to a model's prediction, based on Shapley values from cooperative game theory. This allows for a consistent, model-agnostic explanation of how individual features drive decisions across all samples. From SHAP, we used **beeswarm plots** (see Figure 5.3.7), which offer a dense yet interpretable visualization of feature importance and influence. Each dot represents one sample, with its horizontal position corresponding to the SHAP value (indicating impact magnitude and direction) and its color denoting the feature's actual value. Such plots reveal not only which features are most influential but also whether high or low feature values push predictions toward the seizure or non-seizure class.

In addition, TE2Rules was applied to the Random Forest and XGBoost models. Unlike SHAP, which provides additive feature attributions, TE2Rules extracts **explicit human-readable rules** that describe the sufficient and necessary conditions for classification. The algorithm uses an Apriori-based rule mining process to identify recurring patterns in the learned decision trees, enabling a direct logical interpretation of model behavior. Together, SHAP and TE2Rules form a comprehensive interpretability framework, combining quantitative feature relevance with qualitative, rule-based reasoning. This integration offers not only a better understanding of model decisions but also a means to validate their clinical plausibility in the context of multimodal seizure detection.

5.3 Results

5.3.1 Performance

This subsection presents the best performance scores achieved for each training ratio, comparing models trained with augmented data against those trained solely on the original samples. This comparison enables us to quantify the impact of class imbalance and to assess how effectively data augmentation improves generalization. Furthermore, the analysis provides insight into the relative robustness, adaptability, and discriminative capacity of the different classifiers.

As shown in Figure 5.3.1, the ensemble of XGBoost, CatBoost, and LightGBM achieved the highest overall performance, confirming that model ensembling enhances stability and predictive accuracy through the aggregation of diverse decision boundaries. Among the individual classifiers, XGBoost and CatBoost consistently outperformed LightGBM and HistGradientBoosting, which recorded the lowest scores across most configurations.

A key observation is that models trained with a 1:2 ratio (seizure to non-seizure) yielded the best performance on average. This ratio appears to strike an effective balance between the overrepresentation of non-seizure samples and the limited availability of seizure instances, thus reducing bias while preserving representativeness. The effect of augmentation, however, was model-dependent: some classifiers benefited substantially,

while others—particularly those with built-in robustness mechanisms against class imbalance—showed only marginal improvement or even slight degradation, likely due to noise introduced by synthetic samples.

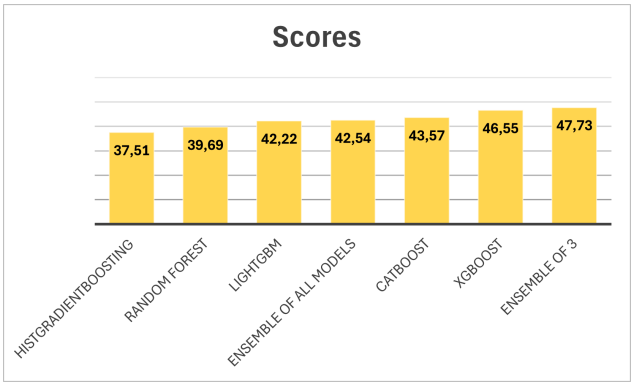

Figure 5.3.1: Best score achieved by each model across different training ratios and augmentation conditions.

XGBoost

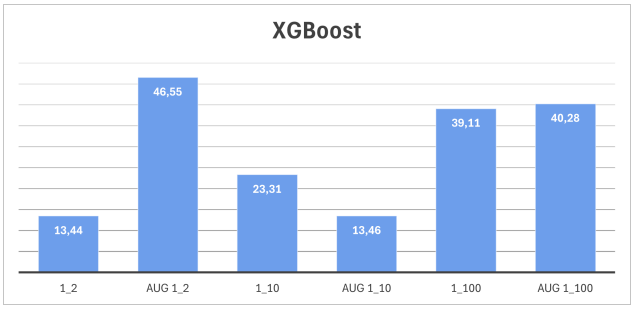

Figure 5.3.2: Scores of the XGBoost models across different training ratios and augmentation settings.

Figure 5.3.2 presents the best scores obtained by the XGBoost models under varying training ratios and augmentation conditions. Among all configurations, the model trained with a **1:2 ratio** and **augmented samples** achieved the highest performance. This optimal configuration reached a score of **46.55**, derived from a sensitivity of **0.61** and a False Alarm Ratio (FAR) of **35.10**, with a decision threshold set at **0.85**. In this context, the threshold value indicates that a probability of at least 85

XGBoost demonstrated mixed behavior across different ratios and data augmentation strategies. For datasets with **1:2** and **1:100** ratios, the inclusion of augmented data provided a clear performance benefit, suggesting that the model effectively utilized the added data diversity to generalize better. However, at the **1:10** ratio, augmentation slightly degraded performance, likely due to an increased imbalance in the learned decision boundaries. Overall, these results confirm the adaptability of XGBoost but also reveal its sensitivity to data composition, underscoring the importance of tuning augmentation strategies to the training ratio. The model's consistently high scores across multiple configurations reinforce its robustness and efficiency in handling heterogeneous multimodal data.

LightGBM

Figure 5.3.3 illustrates the performance of the LightGBM models under different training configurations. The best-performing model was trained using a **1:2 ratio** without augmented samples, achieving a score of **42.22**, with a sensitivity of **0.58** and a False Alarm Ratio of **43.64**, using a decision threshold of **0.85**. Similar to XGBoost, increasing the training ratio yielded better results overall, while extremely imbalanced configurations (1:100 ratio) failed to produce meaningful performance.

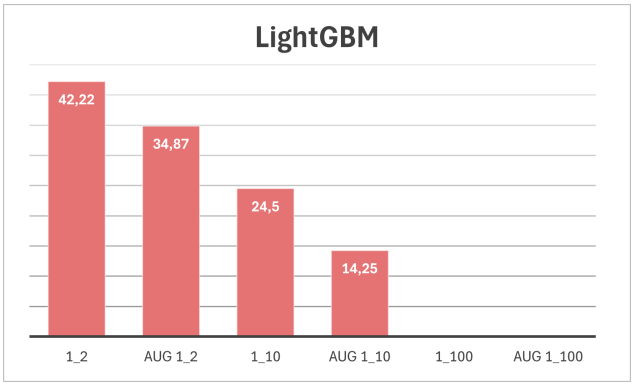

Figure 5.3.3: Scores of the LightGBM models across different training ratios and augmentation settings.

A key observation is that, in contrast to XGBoost, **data augmentation consistently reduced performance** in LightGBM across all tested ratios. This may be attributed to LightGBM's leaf-wise growth strategy and its sensitivity to data noise: augmented samples—especially if they introduce subtle inconsistencies—can distort the learned gradient distributions, reducing the model's discriminative precision. Thus, while LightGBM proved stable and competitive, its optimal performance was achieved under carefully curated, non-augmented datasets, emphasizing its reliance on data purity rather than synthetic diversity.

CatBoost

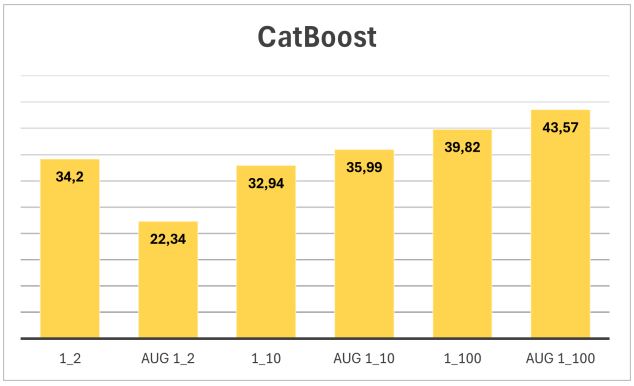

Figure 5.3.4: Scores of the CatBoost models across different training ratios and augmentation settings.

As shown in Figure 5.3.4, CatBoost exhibits a performance pattern opposite to that of LightGBM. In this case, the model's performance **improves as the training ratio decreases**, and the use of augmented data generally enhances its outcomes. The best-performing CatBoost model was trained with a **1:100 ratio** that included augmented samples, achieving a score of **43.57**, derived from a sensitivity of **0.57** and a False Alarm Ratio of **32.69**, with a decision threshold of **0.10**. This low threshold indicates a conservative decision policy, favoring sensitivity over specificity—an approach particularly advantageous when minimizing missed detections is more critical than avoiding false alarms.

Notably, CatBoost was the **only model** among those evaluated to achieve its best score under the most imbalanced (1:100) configuration. This distinctive behavior may stem from CatBoost's gradient-based handling of categorical and numerical feature interactions, which enables it to exploit even limited true samples effectively when supported by diverse augmented data. The model's robustness under imbalance highlights its strength in dealing with heterogeneous multimodal distributions and its superior ability to capture complex nonlinear dependencies across EEG and physiological modalities.

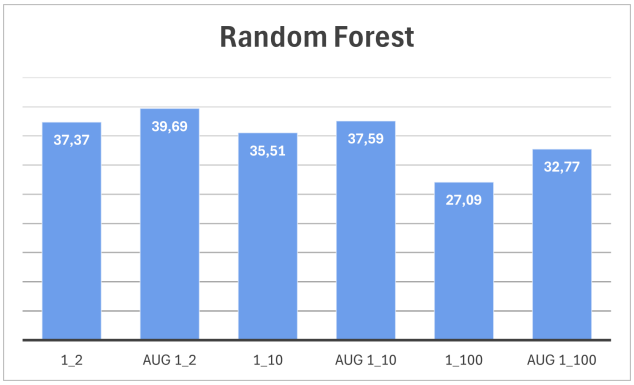

Figure 5.3.5: Scores of the Random Forest models across different training ratios and augmentation settings.

Random Forest

Figure 5.3.5 summarizes the performance of the Random Forest models under varying training ratios and data augmentation conditions. In this case, both increasing the training ratio and incorporating augmented data led to noticeable performance improvements. The best results were obtained with the model trained using a 1:2 ratio that included augmented samples, achieving an optimal balance between sensitivity and false alarm control. This configuration yielded the highest score of **39.69**, corresponding to a sensitivity of **0.58** and a False Alarm Ratio (FAR) of **38.5**, with a decision threshold set at **0.55**.

These findings suggest that Random Forest benefits from additional training data and synthetic variability, likely due to its ensemble averaging mechanism, which helps mitigate overfitting and enhances generalization. The performance trend also aligns with the model's relatively strong sensitivity observed in Section 5.3, reinforcing the value of balanced and augmented training sets in stabilizing decision boundaries across heterogeneous input modalities.

HistGradientBoosting

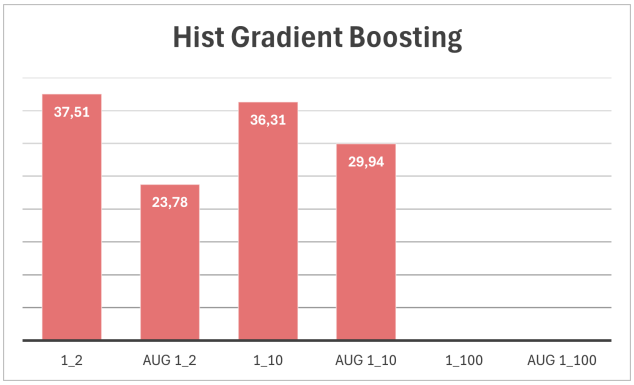

Figure 5.3.6: Scores of the HistGradientBoosting models across different training ratios and augmentation settings.

HistGradientBoosting produced the weakest overall performance among the evaluated models, exhibiting behavior similar to that of LightGBM but with greater instability across training configurations. As illustrated in Figure 5.3.6, both decreasing the training ratio and introducing augmented data led to performance degradation, suggesting that this model was less capable of leveraging additional synthetic samples. Models trained with the most imbalanced configuration (1:100 ratio) failed to learn effectively, producing near-random outcomes.

The best-performing HistGradientBoosting model was trained with a 1:2 ratio without augmentation, achieving a score of **37.51**, derived from a sensitivity of **0.54** and a False Alarm Ratio of **44.18**, with a decision

threshold of **0.90**. The consistently weaker performance of HistGradientBoosting compared to its boosting counterparts (XGBoost and LightGBM) likely reflects its limited flexibility in handling complex, high-dimensional relationships among EEG and physiological features. Moreover, its binning-based optimization, while computationally efficient, may have restricted its ability to capture fine-grained signal dynamics that are critical for this classification task.

Ensemble Models

Two ensemble configurations were evaluated to assess the combined predictive strength of the individual classifiers. The first ensemble integrated all five models—HistGradientBoosting, Random Forest, LightGBM, CatBoost, and XGBoost—through a majority voting scheme. This configuration achieved a score of **42.54**, with a sensitivity of **0.54** and a False Alarm Ratio of **39.59**.

The second ensemble, composed exclusively of the three best-performing models (XGBoost, CatBoost, and LightGBM), delivered the highest overall performance. It achieved a score of **47.73**, with a sensitivity of **0.65** and a False Alarm Ratio of **47.00**. For both ensembles, the decision thresholds of the constituent models were fine-tuned individually to maximize overall predictive efficiency.

The superior performance of the smaller, targeted ensemble underscores the advantage of selectively integrating high-performing classifiers while excluding weaker learners that may introduce decision noise. This result aligns with the SHAP-based feature analyses (Section 5.3), which demonstrated that these three models capture complementary yet robust signal representations. Collectively, the ensemble findings highlight the potential of hybrid model integration strategies for improving robustness and sensitivity in multimodal EEG–physiological classification tasks.

Overall, these results highlight two key insights: (1) gradient boosting methods consistently outperform traditional ensemble techniques for this task, and (2) controlled rebalancing of the data distribution (via ratio adjustment or augmentation) substantially enhances seizure detection performance by mitigating the class imbalance inherent in EEG-ECG datasets.

5.3.2 Explainability – SHAP

This subsection presents the results of the SHAP (SHapley Additive exPlanations) analysis, which was used to interpret the contribution of each feature to the model's predictions. The SHAP beeswarm plots revealed that the most influential features for seizure detection were the interquartile range (IQR), the peak-to-peak amplitude of the ECG signal, and the absolute power of the theta band. Notably, higher values of these features tend to increase the model's likelihood of predicting a seizure event.

The prominence of the IQR aligns with prior studies highlighting its role as a discriminative feature for EEG-based seizure classification. For example, [19] demonstrated that the IQR effectively separates normal, interictal, and ictal EEG segments, achieving nearly 100% classification accuracy. Similarly, changes in ECG morphology—particularly variations in QRS peak-to-peak amplitude—have been associated with ictal episodes, as shown in [165]. Furthermore, the observed importance of theta-band absolute power resonates with findings from [56], which identified theta-band increases as a consistent marker of epileptic activity across multiple studies.

Other relevant features included the RR interval (derived from ECG), ECG skewness, and heart rate. Lower RR intervals (corresponding to elevated heart rates) were strongly associated with seizure events, confirming the physiological link between epileptic activity and autonomic arousal. Consistent with clinical findings, studies such as [70] and [114] have documented shortened RR intervals and increased heart rates during ictal states, while [40] and [17] reported ictal tachycardia as one of the most frequent autonomic manifestations during seizures.

Overall, the SHAP analysis provides physiologically interpretable evidence supporting the multimodal nature of seizure biomarkers, combining EEG spectral power alterations with ECG-based indicators of autonomic dysregulation.

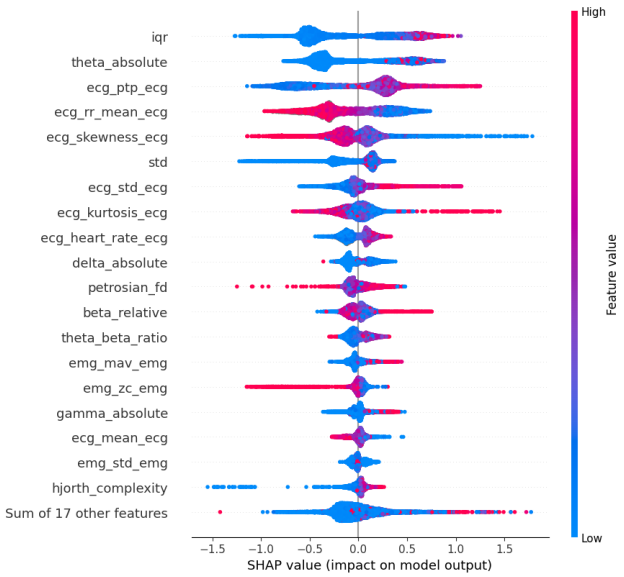

Figure 5.3.7: SHAP Beeswarm plot for the XGBoost model

XGBoost

As demonstrated in Section 5.3.1, the XGBoost model achieved the highest overall predictive performance among the evaluated classifiers. The SHAP analysis shown in Figure 5.3.7 provides deeper insight into the underlying feature contributions that drove this superior performance. The interquartile range (IQR) emerges as the most influential feature, aligning with the feature importance trends observed across most models. The IQR's dominance suggests that the variability and dispersion of the EEG signal are highly predictive of the target outcomes, possibly capturing the degree of neural signal irregularity associated with the studied mental states.

Interestingly, XGBoost diverges from other models in its prioritization of subsequent features. The absolute power of the theta band is identified as the second most important predictor, followed by the peak-to-peak amplitude of the ECG signal. This ranking highlights XGBoost's capacity to capture both neural and physiological dynamics, with theta activity often linked to cognitive control, drowsiness, or emotional engagement, and ECG amplitude reflecting autonomic reactivity. Notably, heart rate—while prominent in other models—appears less influential in XGBoost's internal feature hierarchy, indicating that XGBoost may rely more on fine-grained temporal variability (e.g., IQR, amplitude) rather than coarse global metrics (e.g., mean heart rate).

LightGBM

LightGBM exhibits a feature importance profile largely consistent with that of the other tree-based models, reinforcing the robustness of certain physiological and EEG-derived features across different boosting frameworks. Similar to XGBoost, the IQR and frequency-domain measures (such as theta power) remain among the top-ranked predictors, underscoring their stable contribution to model discrimination. However, a noteworthy distinction is the comparatively lower importance assigned to the standard deviation (STD) of the EEG signal. This suggests that LightGBM's partitioning strategy may rely more on distributional shape and specific frequency-related patterns rather than on general amplitude variability.

Overall, the LightGBM feature interpretation indicates that while its predictive structure mirrors that of XGBoost, its feature weighting slightly de-emphasizes statistical dispersion metrics. This may partially explain the subtle differences in performance and feature interpretability observed in Section 5.3.1.

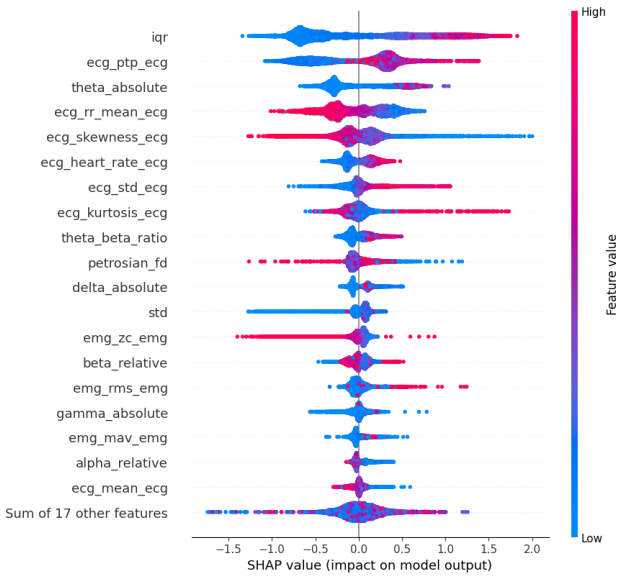

Figure 5.3.8: SHAP Beeswarm plot for the LightGBM model

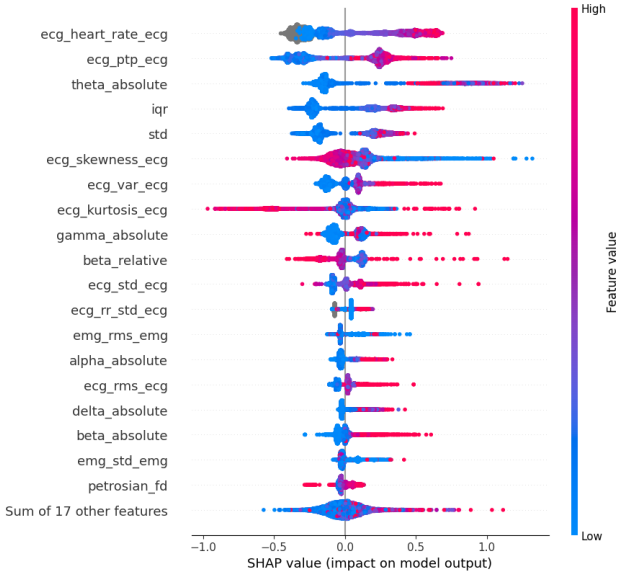

Figure 5.3.9: SHAP Beeswarm plot for the CatBoost model

CatBoost

In contrast to both XGBoost and LightGBM, CatBoost displays a notably distinct feature importance hierarchy, suggesting that it leverages different input dimensions to achieve its predictions. The SHAP beeswarm plot in Figure 5.3.9 reveals that heart rate is the most influential feature for CatBoost, indicating a stronger reliance on global physiological indicators rather than on EEG-derived measures. Meanwhile, the interquartile range (IQR) occupies the fourth position, and the standard deviation (STD) ranks fifth, both contributing less prominently to the model's decision-making process.

This shift in emphasis may be attributed to CatBoost's native handling of feature interactions and categorical encodings, which could amplify the relevance of smoother, low-dimensional physiological features (like heart rate) over high-dimensional EEG variability indices. Despite these differences, as reported in Section 5.3, CatBoost still achieved the second-best performance overall. This finding implies that multiple complementary feature hierarchies—emphasizing either physiological or neural signals—can effectively support robust

classification within this multimodal dataset.

Random Forest

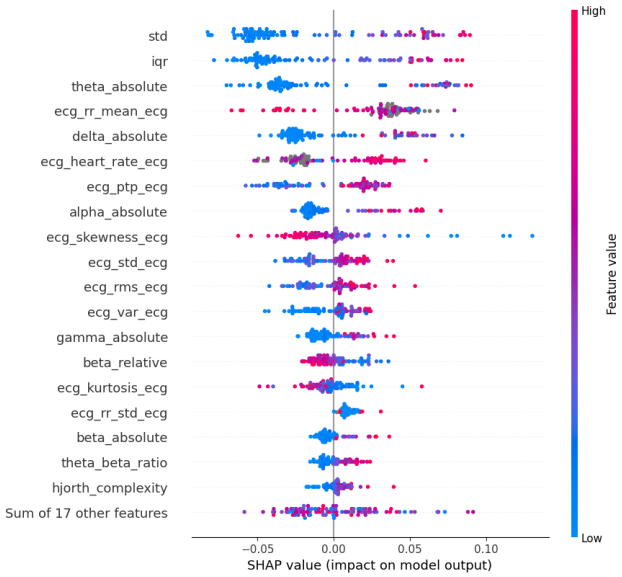

Figure 5.3.10: SHAP Beeswarm plot for the Random Forest model

The Random Forest model exhibits distinct differences in its feature importance distribution compared to the boosting-based classifiers. As shown in Figure 5.3.10, the most influential feature for Random Forest is the standard deviation (STD) of the EEG signal, followed by the interquartile range (IQR). This prioritization suggests that Random Forest relies more heavily on statistical dispersion metrics, emphasizing variability in EEG amplitude as a key discriminative factor. The prominence of these measures indicates that the ensemble of decision trees captures meaningful fluctuations and irregularities in neural activity, potentially linked to cognitive or affective state transitions.

Interestingly, the peak-to-peak amplitude of the ECG signal, which ranked highly in XGBoost and CatBoost, appears only in seventh position here, highlighting a reduced dependency on cardiovascular amplitude features. Conversely, Random Forest assigns greater importance to the absolute power of the delta and alpha frequency bands. This shift implies that the model captures more pronounced spectral information, possibly due to its ability to partition feature space based on multiple frequency components without relying on strong regularization, as boosting models do. Overall, Random Forest's interpretability profile reflects a broader sensitivity to both low-frequency EEG dynamics and amplitude variability, which may contribute to its comparatively balanced, though not top-ranked, predictive performance.

HistGradientBoosting

The feature importance profile derived from the HistGradientBoosting model remains largely aligned with the general trends observed across classifiers, reaffirming the robustness of key multimodal predictors. The interquartile range (IQR) and frequency-domain EEG features continue to play a significant role, though the standard deviation (STD) appears notably less influential in this model. This suggests that HistGradientBoosting relies less on simple measures of dispersion and instead emphasizes frequency-specific or distributional attributes.

Despite this general alignment with the overall feature patterns, the HistGradientBoosting classifier demonstrated the weakest predictive performance (see Section 5.3). This outcome likely reflects the model's lower representational flexibility compared to gradient-boosting variants such as XGBoost and LightGBM. While its internal feature weighting structure mirrors the others, its optimization and binning mechanisms may have limited its capacity to capture fine-grained nonlinearities across EEG and physiological inputs.

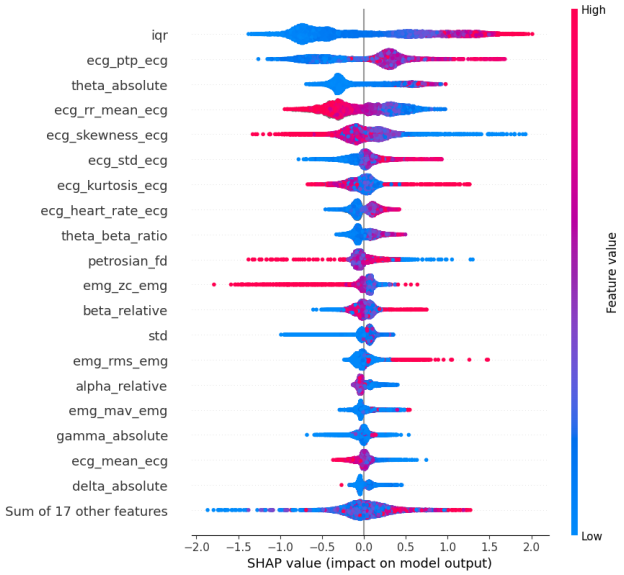

Figure 5.3.11: SHAP Beeswarm plot for the HistGradientBoosting model

Ensemble Models

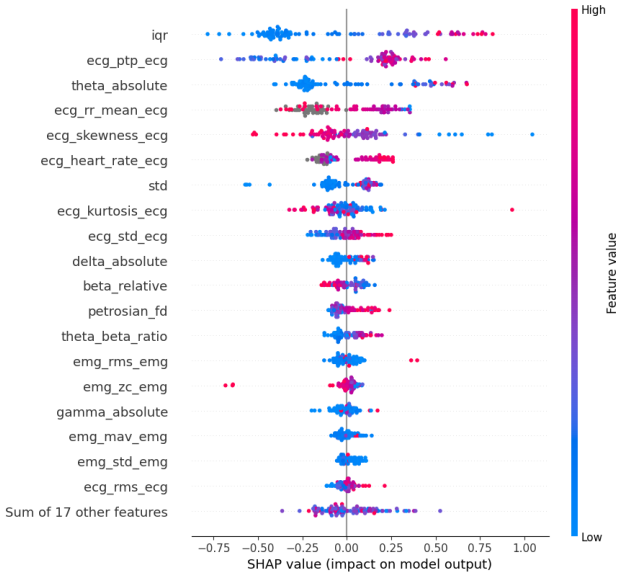

Figure 5.3.12: SHAP Beeswarm plot for the ensemble of all classifiers

To further explore model complementarity, SHAP analyses were performed on two ensemble configurations: one combining all classifiers, and another restricted to the three top-performing gradient-boosting models (XGBoost, LightGBM, and CatBoost). The beeswarm plots in Figures 5.3.12 and 5.3.13 illustrate that both ensembles preserve the dominant feature hierarchy observed in individual models, with IQR, theta power, and heart rate consistently emerging as key contributors. Minor variations occur among the lower-ranked features, reflecting the averaging of SHAP values across diverse model architectures.

As reported in Section 5.3, the ensemble restricted to XGBoost, LightGBM, and CatBoost outperformed the broader ensemble. This suggests that the combination of high-performing boosting models yields synergistic effects by integrating complementary decision boundaries while minimizing noise from weaker learners. The resulting ensemble thus balances interpretability and generalization, confirming the robustness of the

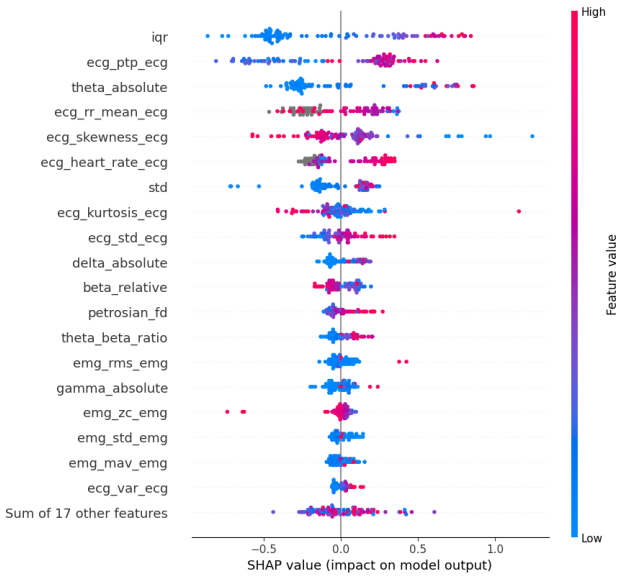

Figure 5.3.13: SHAP Beeswarm plot for the ensemble of XGBoost, LightGBM, and CatBoost

identified multimodal biomarkers across different learning paradigms.

5.3.3 Explainability – TE2Rules

In addition to feature-level explanations, this subsection reports interpretable rule sets extracted using the TE2Rules library [81]. This framework derives explicit decision rules that approximate the internal logic of tree ensemble models (such as Random Forest and XGBoost) using an Apriori-based rule mining approach. The resulting rule sets capture the necessary and sufficient feature combinations that lead to specific classifications.

Figures 5.3.14 and 5.3.15 illustrate representative rule sets derived from the Random Forest and XGBoost models. Examination of the Random Forest rules reveals that standard deviation (std) and interquartile range (IQR) appear most frequently, in line with SHAP's importance rankings. However, the rules also highlight additional features—such as skewness, the theta-to-beta ratio, and Hjorth complexity—that did not rank among the top SHAP contributors. These features may play context-dependent roles that emerge only through specific interactions captured by the rule mining process.

Similar findings are observed for XGBoost, where IQR and std again dominate, but rules involving theta-to-beta ratios and Hjorth complexity appear recurrently. This convergence between SHAP and TE2Rules underscores the consistency of certain biomarkers, while the divergences emphasize the complementary perspectives offered by additive feature attributions and rule-based reasoning. Together, these explainability methods not only enhance model transparency but also provide a richer, more nuanced understanding of how physiological and spectral features jointly contribute to seizure detection.

Collectively, the explainability analyses suggest that robust seizure detection relies on a combination of time-domain variability metrics (IQR, std), spectral features (theta-band power, theta-to-beta ratio), and ECG-derived measures of autonomic activity. These findings provide interpretable, physiologically grounded insights that not only validate model decisions but also strengthen confidence in their clinical relevance.

```
=== Extracted TE2Rules RF ===

std4 > 55.90328598022461 & iqr > 47.49892616271973 & iqr0 > 94.63816452026367 & iqr4 > 34.5535945892334 & iqr6 <= 0.10300000011920929 &
petrosian_fd <= 1.032990574836731
std2 > 72.9266586303711 & iqr > 22.76581859588623 & iqr1 <= 23.50987434387207 & iqr3 > 263.2537078857422 & iqr4 > 64.00311660766602 & iqr5 <=
0.7100000083446503 & iqr7 > 0.8960431516170502 & hjorth_mobility <= 1.0168043375015259
std > 18.579233169555664 & std2 > 97.21244812011719 & iqr1 <= 6.300330638885498 & iqr3 > 263.66168212890625 & iqr7 <= 0.6177054643630981
std <= 19.36530590057373 & std > 14.560649394989014 & std3 > 0.1083282046020031 & iqr7 <= -1.615802824497223 & iqr9 > 1820.10498046875
std0 > 138.94725799560547 & iqr > 24.698569297790527 & iqr4 <= 39.71879005432129 & iqr5 <= 0.6620000004768372 & skewness0 <= 3.560608744621277
std0 > 87.52128982543945 & std2 > 72.88996124267578 & std8 <= 83.60734558105469 & iqr3 > 288.22706604003906 & iqr5 <= 0.714000016450882 & iqr7 >
1.885946810245514 & iqr9 <= 12299.79931640625 & kurtosis <= 0.9613438844680786
theta_beta_ratio > 0.4980572313070297 & std > 19.742846488952637 & std4 > 53.52985763549805 & std5 <= 0.16413573920726776 & std6 <=
305.2804718017578 & iqr0 > 90.09090042114258 & iqr1 <= 14.919915676116943 & iqr4 > 40.67121696472168 & iqr9 <= 39973.880859375
iqr6 > 0.10300000011920929 & iqr8 > 104.7421646118164 & iqr9 <= 15782.47998046875 & skewness1 <= 4.796292543411255
std > 22.15066432952881 & std2 > 67.28566360473633 & iqr5 > 0.6860000193119049 & iqr8 <= 42.85501289367676 & hjorth_complexity >
2.9254664182662964
theta_beta_ratio <= 0.4912314713001251 & std <= 33.50930404663086 & std > 19.25168800354004 & std4 > 51.936065673828125 & iqr >
28.55723762512207 & iqr0 > 90.09090042114258 & iqr7 <= 4.885385990142822 & iqr8 > 39.48525428771973 & hjorth_mobility <= 0.7713758051395416
std2 > 38.34693908691406 & std7 <= 0.22397403419017792 & std8 > 111.73356246948242 & iqr1 > 9.588655471801758 & iqr3 > 294.79351806640625 & iqr7
<= 5.108145713806152 & iqr7 > -0.4570353031158447 & iqr8 > 44.13473129272461 & hjorth_complexity > 1.161778748035431 & approx_entropy >
1.0204318761825562
std > 19.36530590057373 & std2 <= 710.691650390625 & std4 > 79.35497283935547 & iqr > 40.088144302368164 & iqr1 <= 8.466097354888916 & iqr1 >
0.29392053186893463 & iqr3 <= 255.72965240478516 & iqr5 <= 0.6580000221729279 & iqr7 <= 1.4769538640975952 & skewness4 > 1031.1677856445312
std > 16.20181179046631 & std2 <= 69.74358367919922 & std6 <= 136.61009979248047 & iqr <= 30.59929084777832 & iqr4 > 65.09215545654297 & iqr5 <=
0.7180000245571136 & iqr8 > 40.33426094055176 & skewness1 > 5.961409568786621 & zero_crossings <= 57.25
std1 <= 0.45610281825065613 & std3 > 0.19292756170034409 & std4 <= 55.752336502075195 & std5 <= 0.11108389496803284 & std6 > 28.00527000427246 &
std7 <= 0.15089960396289825 & iqr > 21.032020568847656 & iqr5 > 0.7060000002384186 & hjorth_complexity > 1.4229558110237122
std0 > 53.89544868469238 & std2 <= 67.7069206237793 & std4 <= 67.90992736816406 & std4 > 6.302347898483276 & std7 > 0.09845180809497833 & iqr <=
21.421411514282227 & iqr1 <= 39.93844795227051 & iqr5 <= 0.49699999392032623 & iqr6 <= 0.0820000022649765 & iqr8 <= 107.63129425048828
std0 > 115.31373977661133 & iqr4 <= 107.24712753295898 & iqr6 > 0.10300000011920929 & iqr7 > 5.845348596572876
```

Figure 5.3.14: Example of extracted rules from the Random Forest model using TE2Rules.

```
=== Extracted TE2Rules (XGBoost) ===

std2 >= 67.7481918 & iqr1 < 10.5477562 & iqr5 < 0.712000012 & iqr8 >= 37.100956 & skewness1 < 583.087341 & petrosian_fd < 1.03210402
std1 < 0.654772043 & std2 >= 113.5644 & std4 >= 7.81451273 & std6 >= 4.45363188 & std7 >= 0.03924115 & iqr1 < 4.23302174 & iqr7 >= -3.35167527 &
skewness5 < 91.0
std0 >= 89.4768066 & std2 >= 40.2755356 & iqr3 >= 283.875854 & iqr5 < 0.600000024 & iqr7 >= 0.559303641
std >= 14.5811243 & std1 < 0.419865966 & std2 < 113.5644 & std5 >= 0.0315472297 & std7 >= 0.0518024713 & iqr3 < 738.832031 & iqr3 >= 339.627258
& iqr7 < -1.44384813
theta_beta_ratio < 0.611791015 & std2 >= 53.1654243 & iqr >= 42.1198654 & iqr1 < 6.92864084 & iqr5 < 0.783999979
std >= 14.5811243 & std2 >= 113.5644 & iqr < 41.5314636 & iqr >= 18.9989185 & hjorth_complexity >= 1.75268304 & petrosian_fd >= 1.02956975
iqr >= 18.9989185 & iqr3 < 903.465942 & iqr8 >= 111.605919 & skewness1 < 411.188599
std7 < 0.467306197 & iqr >= 24.0597782 & iqr2 < 4.11730576 & iqr3 < 348.000061 & iqr3 >= 264.837311 & iqr5 < 0.660000026 & skewness1
< 5.89941406
std3 >= 0.0475089252 & iqr1 >= 15.6237745 & iqr2 >= -4.42912054 & iqr3 < 799.746948 & iqr3 >= 345.13266 & iqr5 < 0.712000012 & iqr8 >= 23.132019
& skewness2 < 5.40126896 & petrosian_fd < 1.02944458
std0 >= 44.31036 & iqr >= 30.6013279 & iqr5 < 0.783999979 & iqr8 >= 66.9966049 & skewness1 < 411.188599 & skewness5 < 91.0 &
hjorth_complexity >= 1.15770948
std0 < 44.31036 & iqr3 < 930.040649 & iqr4 >= 105.815277 & iqr6 >= 0.0719999969 & skewness5 < 187.0
std1 < 0.581734478 & std2 < 102.329643 & std7 >= 0.0419003442 & iqr >= 24.0597782 & iqr3 < 264.837311 & iqr5 < 0.660000026 &
hjorth_complexity >= 2.14616656
std1 < 0.419865966 & std2 >= 81.307457 & std6 >= 2.21932149 & std9 >= 0.00354741537 & iqr >= 24.4599457 & iqr5 >= 0.691999972 & kurtosis
< -0.417636603
std >= 12.3525658 & iqr1 < 40.3446236 & iqr7 < 4.8598299 & iqr7 >= -1.49403417 & iqr8 >= 63.8544846 & skewness0 >= 54.9036064 & skewness1 >=
120.003464 & skewness5 >= 65.0
std4 < 4.23822069 & std8 >= 5.40889645 & iqr1 >= 13.782177 & iqr7 >= -3.35167527
std >= 12.9683733 & std8 < 452.743469 & iqr >= 41.5314636 & iqr1 >= -1.12345982 & iqr5 < 0.527999997
theta_beta_ratio < 0.963607311 & std3 >= 0.0475089252 & iqr3 >= 313.700653 & iqr7 < 3.90529156 & hjorth_complexity >= 1.20928526 & petrosian_fd
< 1.02456212
std1 < 0.456866741 & std1 >= 0.419865966 & std7 >= 0.0496475473 & iqr2 < 4.53908539 & iqr3 >= 264.837311 & iqr7 < 4.8598299
std4 >= 7.4728694 & std6 >= 4.45363188 & std7 < 0.467306197 & iqr2 < 4.11730576 & iqr3 >= 264.837311 & iqr7 < -3.16531134 & skewness1 >=
5.89941406
```

Figure 5.3.15: Example of extracted rules from the XGBoost model using TE2Rules.

Chapter 6

Conclusion

In this thesis, we proposed and evaluated a methodology based on handcrafted features extracted from EEG signals, demonstrating that such features can achieve strong performance even compared to modern deep learning models. The preprocessing and feature extraction steps were carefully designed to enhance the signal quality and extract physiologically meaningful information. The results revealed that, under specific conditions, traditional feature-based approaches outperform certain end-to-end neural architectures. This finding aligns with the outcomes of the Una Europa Seizure Detection Challenge (2023), where the winning approach followed a similar framework and reported superior performance relative to deep learning-based solutions [67].

A key reason for this observation lies in the characteristics of the competition dataset: EEG recordings were obtained from wearable devices with a limited number of channels, resulting in signals that are low in quality, contain high levels of noise, and are more susceptible to motion artifacts. These factors make feature engineering and preprocessing crucial for enhancing the signal-to-noise ratio and capturing the most discriminative aspects of brain activity. In contrast, deep learning models—although powerful in extracting representations from large and clean datasets—often require a substantial amount of data to generalize effectively. When data are scarce or noisy, handcrafted features rooted in physiological knowledge tend to provide better generalization and interpretability.

From a methodological perspective, this thesis reinforces the importance of domain knowledge and interpretability in EEG-based machine learning. Rather than prioritizing model complexity, effective solutions often depend on the quality of the preprocessing pipeline and the interpretability of the extracted features. This approach not only leads to robust performance but also facilitates transparency—an essential requirement in medical and neurophysiological applications.

Explainable AI for EEG Analysis. XAI in EEG analysis represents a rapidly evolving research frontier. The majority of studies to date employ *post-hoc* explanation methods, including visualization-based techniques such as saliency maps and Grad-CAM, as well as feature attribution tools like SHAP and LIME. Perturbation and occlusion-based methods also play an important role, providing fine-grained insights into model sensitivity. However, these approaches face several conceptual and practical limitations.

A central concern is faithfulness: whether the explanation genuinely reflects the model's internal reasoning or merely provides a plausible, human-readable narrative [37, 134]. Many post-hoc techniques, particularly those that rely on surrogate models (e.g., LIME), introduce approximations that can distort the model's true decision boundaries. Similarly, robustness remains an open issue. Explanations can vary dramatically with small perturbations to the input data, casting doubt on their reliability and reproducibility [59]. Moreover, recent analyses have revealed that the effectiveness of these methods depends strongly on the underlying model structure and can be biased by feature collinearity [136].

Beyond methodological weaknesses, there are also ethical and security concerns. Post-hoc explanations can be intentionally manipulated to present biased or unfair models as transparent—a phenomenon known as fairwashing [5, 148]. Such vulnerabilities highlight that adversarial manipulation can extend beyond models

themselves to their interpretability mechanisms, potentially misleading users and stakeholders. These issues have prompted growing support for inherently interpretable (ante-hoc) models, which provide transparency by design rather than through approximation [134].

Current Trends and Research Gaps. Our review indicates that XAI methods have been successfully applied to a wide range of EEG-based tasks, including *epilepsy and seizure detection*, *sleep monitoring*, *motor imagery and execution*, *emotion recognition*, *schizophrenia diagnosis*, *stroke prediction*, and *major depressive disorder*. This variety underscores the versatility of XAI approaches in neural signal analysis and their growing importance in clinical and cognitive neuroscience contexts.

However, the majority of these approaches remain task-specific, optimized for narrow problem domains without demonstrating cross-task generalization. As a result, it remains difficult to assess how well a given explainability method can transfer across applications. Another significant limitation is that most studies rely exclusively on EEG data, neglecting multimodal approaches that integrate EEG with other neuroimaging modalities such as MRI or fMRI. These multimodal perspectives could enable richer and more physiologically grounded explanations, but systematic research in this direction remains limited.

Furthermore, there is a notable absence of ground truth explanations and standardized benchmarks for evaluating XAI performance in EEG analysis. The lack of formal definitions for what constitutes a "correct" explanation makes it challenging to quantify or compare the fidelity of different methods [63]. Popular attribution methods can mistakenly ascribe importance to irrelevant features, leading to misleading interpretations that undermine clinical validity. This issue becomes particularly problematic in healthcare applications, where incorrect or unstable explanations could have ethical or diagnostic consequences.

Another persistent gap concerns human-centered validation. While many studies report quantitative measures of explainability, few incorporate qualitative evaluations from clinicians or domain experts. Such human-in-the-loop evaluation is vital for bridging the gap between algorithmic interpretability and real-world usability. Frameworks such as that proposed in [116] emphasize the need for co-design and iterative feedback between AI developers and domain experts, ensuring that explanations are meaningful and actionable in clinical practice.

Limitations. This thesis, while offering valuable insights, is subject to several limitations. First, the study focused primarily on a single dataset from the Una Europa Seizure Detection Challenge, which may limit the generalizability of the findings. Although the dataset represents realistic clinical conditions, additional datasets with different acquisition protocols and population demographics would be needed to fully validate the proposed framework.

Second, the feature-based approach, while interpretable, is constrained by the chosen feature set and may overlook subtle temporal or spatial dynamics that deep models could potentially capture. The results thus highlight a trade-off between interpretability and representational power. Furthermore, the evaluation of explainability in this work was primarily based on model-agnostic techniques (e.g., SHAP and te2rules), which, despite their utility, remain subject to the same limitations of faithfulness and robustness discussed earlier.

Finally, the human evaluation component was limited, as the explanations generated were not systematically validated by domain experts. Incorporating expert feedback would strengthen both the interpretive and clinical credibility of the framework. Future studies should therefore prioritize participatory evaluation and multi-criteria assessment of explainability, combining quantitative and qualitative perspectives.

Future Work Although deep learning models did not achieve the anticipated performance in our experiments, they remain powerful tools for pattern recognition and automatic feature extraction. Hybrid and semi-supervised architectures that combine handcrafted features with learned representations have shown promising results in related domains [180]. Future research could explore such hybrid approaches for seizure detection and other EEG-based tasks, leveraging both data-driven and knowledge-driven methodologies to achieve a balance between accuracy and interpretability.

Another promising direction involves expanding the framework toward multimodal explainability. Integrating EEG with complementary modalities such as fMRI, eye-tracking, or physiological signals (e.g., ECG,

EMG) could lead to more robust and neurophysiologically grounded interpretations of cognitive and clinical phenomena. Such multimodal XAI systems could reveal how brain dynamics correlate with behavior or emotional state in a more holistic manner.

In addition, future work should focus on developing standardized benchmarks and quantitative metrics for evaluating XAI in EEG analysis. Establishing ground truth annotations for feature importance or decision rationale—potentially through synthetic datasets or expert-labeled explanations—would allow more objective comparisons across models and studies.

Another key avenue is the incorporation of human-centered evaluation frameworks [116]. Collaborative studies involving neurologists, psychologists, and clinicians can provide essential feedback on the interpretability and usefulness of explanations. Such interdisciplinary validation would not only enhance trust but also guide the design of explanations that align with clinical reasoning processes.

Finally, future research should explore the ethical and regulatory implications of explainable AI in EEG analysis. As AI systems move closer to deployment in healthcare, ensuring transparency, fairness, and accountability will be critical. This includes designing systems that are resilient to manipulation (e.g., fairwashing) and sensitive to the social and ethical contexts in which they operate.

Summary. In conclusion, this thesis contributes to the understanding of how feature-based EEG analysis combined with explainable AI can offer both high performance and interpretability in seizure detection. The results emphasize that interpretability should not be treated as an afterthought but as a central design criterion in medical AI. By integrating robust signal processing, transparent modeling, and explainability, this work lays the foundation for future systems that are not only accurate but also trustworthy and clinically meaningful.

Chapter 7

Bibliography

[1] Abd-alrazaq, A. A. et al. "Detection of Sleep Apnea Using Wearable AI: Systematic Review and Meta-Analysis". In: *Journal of Medical Internet Research* 26 (2024). DOI: 10.2196/58187.

[2] Ahmad, I. "An efficient feature selection and explainable classification method for EEG-based epileptic seizure detection". In: *Journal of Information Security and Applications* (2024), p. 103654.

[3] Ahmad, I. et al. "A hybrid deep learning approach for epileptic seizure detection in EEG signals". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[4] Ahmad, I. et al. "A secure and interpretable AI for smart healthcare system: a case study on epilepsy diagnosis using EEG signals". In: *IEEE Journal of Biomedical and Health Informatics* (2024).

[5] Aı¨vodji, U. et al. "Fairwashing: the risk of rationalization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 161–170.

[6] Al Hammadi, A. Y. et al. "Explainable artificial intelligence to evaluate industrial internal security using EEG signals in IoT framework". In: *Ad Hoc Networks* 123 (2021), p. 102641.

[7] Ali, N. et al. "Optimizing Emotion Recognition in EEG Data: A Genetic Algorithm Approach with XAI Insights". In: *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE. 2024, pp. 1–6.

[8] Aminov, A. et al. "Acute single channel EEG predictors of cognitive function after stroke". In: *PloS one* 12.10 (2017), e0185841.

[9] Amrani, G. et al. "EEG signal analysis using deep learning: A systematic literature review". In: *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*. Fez, Morocco, 2021, pp. 1–8. DOI: 10.1109/ICDS53782.2021.9626707.

[10] Andrzejak, R. G. et al. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state". In: *Physical Review E* 64.6 (2001), p. 061907.

[11] Apicella, A. et al. "Toward the application of XAI methods in EEG-based systems". In: *arXiv preprint arXiv:2210.06554* (2022).

[12] Arias, J. T. and Astudillo, C. A. "Enhancing Schizophrenia Prediction Using Class Balancing and SHAP Explainability Techniques on EEG Data". In: *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*. IEEE. 2023, pp. 1–5.

[13] Arrieta, A. B. et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[14] Asif, M. et al. "Emotion recognition using temporally localized emotional events in EEG with naturalistic context: DENS# dataset". In: *IEEE Access* 11 (2023), pp. 39913–39925.

[15] Bach, S. et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[16] Banik, S. et al. "Assessment of Valance Emotional State Using EEG-EDA Coupling and Explainable Classifiers". In: *Digital Health and Informatics Innovations for Sustainable Health Care Systems*. IOS Press, 2024, pp. 953–957.

[17] Baumgartner, C., Lurger, S., and Leutmezer, F. "Autonomic symptoms during epileptic seizures". In: *Epileptic Disorders* 3.3 (2001), pp. 103–116.
[18] Beck, M. et al. "xlstm: Extended long short-term memory". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 107547–107603.
[19] Bedeeuzzaman, M., Farooq, O., and Khan, Y. U. "Automatic seizure detection using inter quartile range". In: *International Journal of Computer Applications* 44.11 (2012), pp. 1–5.
[20] Bhagubai, M. et al. *SeizeIT2: Wearable Dataset Of Patients With Focal Epilepsy*. 2025. DOI: https://doi.org/10.48550/arXiv.2502.01224. arXiv: 2502.01224 [eess.SP]. URL:
[21] Borra, D., Fantozzi, S., and Magosso, E. "Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination". In: *Neural Networks* 129 (2020), pp. 55–74.
[22] Bouazizi, S. and Ltifi, H. "Enhancing accuracy and interpretability in EEG-based medical decision making using an explainable ensemble learning framework application for stroke prediction". In: *Decision Support Systems* 178 (2024), p. 114126.
[23] Bouazizi, S. and Ltifi, H. "Novel diversified echo state network for improved accuracy and explainability of EEG-based stroke prediction". In: *Information Systems* 120 (2024), p. 102317.
[24] Breiman, L. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.
[25] Breiman, L. et al. *Classification and regression trees*. CRC press, 1984.
[26] Cai, H. et al. "A multi-modal open dataset for mental-disorder analysis". In: *Scientific Data* 9.1 (2022), p. 178.
[27] Chang, Y.-W. et al. "Development of an Al-Based Web Diagnostic System for Phenotyping Psychiatric Disorders". In: *Frontiers in Psychiatry* 11 (2020). DOI: 10.3389/fpsyt.2020.542394.
[28] Chen, T. and Guestrin, C. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
[29] Chien, J. and Danks, D. *Trustworthiness in Stochastic Systems: Towards Opening the Black Box*. 2025. arXiv: 2501.16461 [cs.CY]. URL:
[30] Chowdary, M. K., Nguyen, T. N., and Hemanth, D. "Deep learning-based facial emotion recognition for human–computer interaction applications". In: *Neural Computing and Applications* 35 (2021), pp. 23311–23328. DOI: 10.1007/s00521-021-06012-8.
[31] Cortes-Briones, J. et al. "Going deep into schizophrenia with artificial intelligence". In: *Schizophrenia Research* 245 (2021), pp. 122–140. DOI: 10.1016/j.schres.2021.05.018.
[32] Cui, W. et al. "Neuro-gpt: Developing a foundation model for eeg". In: *arXiv preprint arXiv:2311.03764* 107 (2023).
[33] Deng, A. and Hooi, B. "Graph neural network-based anomaly detection in multivariate time series". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 5. 2021, pp. 4027–4035.
[34] Detti, P., Vatti, G., and Zabalo Manrique de Lara, G. "EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings". In: *Processes* 8.7 (2020), p. 846.
[35] Dietterich, T. G. "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer. 2000, pp. 1–15.
[36] Dornhege, G. et al. "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms". In: *IEEE transactions on biomedical engineering* 51.6 (2004), pp. 993–1002.
[37] Doshi-Velez, F. and Kim, B. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).
[38] Došilović, F. K., Brčić, M., and Hlupić, N. "Explainable artificial intelligence: A survey". In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040.
[39] Dutt, M. et al. "SleepXAI: An explainable deep learning approach for multi-class sleep stage identification". In: *Applied Intelligence* 53.13 (2023), pp. 16830–16843.
[40] Eggleston, K. S., Olin, B. D., and Fisher, R. S. "Ictal tachycardia: the head–heart connection". In: *Seizure* 23.7 (2014), pp. 496–505.
[41] Ellis, C. A., Miller, R. L., and Calhoun, V. D. "A novel local explainability approach for spectral insight into raw eeg-based deep learning classifiers". In: *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2021, pp. 1–6.

[42] Ellis, C. A., Miller, R. L., and Calhoun, V. D. "A Model Visualization-based Approach for Insight into Waveforms and Spectra Learned by CNNs". In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2022, pp. 1643–1646.

[43] Ellis, C. A., Miller, R. L., and Calhoun, V. D. "A Systematic Approach for Explaining Time and Frequency Features Extracted by Convolutional Neural Networks From Raw Electroencephalography Data". In: *Frontiers in Neuroinformatics* 16 (2022), p. 872035.

[44] Ellis, C. A., Miller, R. L., and Calhoun, V. D. "A convolutional autoencoder-based explainable clustering approach for resting-state EEG analysis". In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2023, pp. 1–4.

[45] Ellis, C. A., Miller, R. L., and Calhoun, V. D. "Improving Explainability for Single-Channel EEG Deep Learning Classifiers via Interpretable Filters and Activation Analysis". In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2023, pp. 2474–2481.

[46] Ellis, C. A. et al. "A gradient-based approach for explaining multimodal deep learning classifiers". In: *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2021, pp. 1–6.

[47] Ellis, C. A. et al. "A novel activation maximization-based approach for insight into electrophysiology classifiers". In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2021, pp. 3358–3365.

[48] Ellis, C. A. et al. "Algorithm-agnostic explainability for unsupervised clustering". In: *arXiv preprint arXiv:2105.08053* (2021).

[49] Ellis, C. A. et al. "Explainable sleep stage classification with multimodal electrophysiology time-series". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 2363–2366.

[50] Ellis, C. A. et al. "Hierarchical neural network with layer-wise relevance propagation for interpretable multiclass neural state classification". In: *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2021, pp. 351–354.

[51] Ellis, C. A. et al. "Examining effects of schizophrenia on EEG with explainable deep learning models". In: *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2022, pp. 301–304.

[52] Ellis, C. A. et al. "Examining reproducibility of EEG schizophrenia biomarkers across explainable machine learning models". In: *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2022, pp. 305–308.

[53] Ellis, C. A. et al. "A Framework for Systematically Evaluating the Representations Learned by A Deep Learning Classifier from Raw Multi-Channel Electroencephalogram Data". In: *bioRxiv* (2023), pp. 2023–03.

[54] Ellis, C. A. et al. "Novel Approach Explains Spatio-Spectral Interactions in Raw Electroencephalogram Deep Learning Classifiers". In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE. 2023, pp. 1–5.

[55] Ellis, C. A. et al. "Identifying EEG Biomarkers of Depression with Novel Explainable Deep Learning Architectures". In: *bioRxiv* (2024).

[56] Faiman, I. et al. "Resting-state EEG for the diagnosis of idiopathic epilepsy and psychogenic nonepileptic seizures: A systematic review". In: *Epilepsy & Behavior* 121 (2021), p. 108047.

[57] Filandrianos, G. et al. "Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 9507–9525. DOI: 10.18653/v1/2023.findings-acl.606. URL:

[58] Friedman, J. H. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[59] Ghorbani, A., Abid, A., and Zou, J. "Interpretation of neural networks is fragile". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 3681–3688.

[60] Goldberger, A. L. et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *circulation* 101.23 (2000), e215–e220.

[61] Guidotti, R. et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5 (2019), pp. 1–42. DOI: 10.1145/3236009.

[62] Hashem, H. A. et al. "An integrated machine learning-based brain computer interface to classify diverse limb motor tasks: Explainable model". In: *Sensors* 23.6 (2023), p. 3171.
[63] Haufe, S. et al. "Explainable AI needs formal notions of explanation correctness". In: *arXiv preprint arXiv:2409.14590* (2024).
[64] Hochreiter, S. and Schmidhuber, J. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.
[65] Hoffman, R. R. et al. "Metrics for Explainable AI: Challenges and Prospects". In: *arXiv preprint arXiv:1812.04608* (2018).
[66] Hussain, I. et al. "An explainable EEG-based human activity recognition model using machine-learning approach and LIME". In: *Sensors* 23.17 (2023), p. 7452.
[67] Al-Hussaini, I. and Mitchell, C. S. "SeizFt: interpretable machine learning for seizure detection using wearables". In: *Bioengineering* 10.8 (2023), p. 918.
[68] Islam, M. et al. "EEG Channel Correlation Based Model for Emotion Recognition". In: *Computers in biology and medicine* 136 (2021), p. 104757. DOI: 10.1016/j.compbiomed.2021.104757.
[69] Islam, M. S. et al. "Explainable artificial intelligence model for stroke prediction using EEG signal". In: *Sensors* 22.24 (2022), p. 9859.
[70] Jeppesen, J. et al. "Detection of epileptic-seizures by means of power spectrum analysis of heart rate variability: a pilot study". In: *Technology and Health Care* 18.6 (2010), pp. 417–426.
[71] Jonas, S. et al. "EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features". In: *Human brain mapping* 40.16 (2019), pp. 4606–4617.
[72] Ke, G. et al. "LightGBM: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3146–3154.
[73] Kemp, B. et al. "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG". In: *IEEE Transactions on Biomedical Engineering* 47.9 (2000), pp. 1185–1194.
[74] Khan, F. A. et al. "Explainable fuzzy deep learning for prediction of epileptic seizures using EEG". In: *IEEE Transactions on Fuzzy Systems* (2024).
[75] Khessiba, S. et al. "Innovative deep learning models for EEG-based vigilance detection". In: *Neural Computing and Applications* 33.12 (2021), pp. 6921–6937.
[76] Kim, B., Khanna, R., and Koyejo, O. O. "Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016. URL:
[77] Koelstra, S. et al. "Deap: A database for emotion analysis; using physiological signals". In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.
[78] Kong, W. et al. "Weighted extreme learning machine for P300 detection with application to brain computer interface". In: *Journal of Ambient Intelligence and Humanized Computing* (2018), pp. 1–11. DOI: 10.1007/S12652-018-0840-1.
[79] Kostas, D., Aroca-Ouellette, S., and Rudzicz, F. "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data". In: *Frontiers in Human Neuroscience* 15 (2021), p. 653659.
[80] La Fiscal, L. et al. "Explainable AI for EEG Biomarkers Identification in Obstructive Sleep Apnea Severity Scoring Task". In: *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2023, pp. 1–6.
[81] Lal, G. R., Chen, X., and Mithal, V. "TE2Rules: Explaining Tree Ensembles using Rules". In: *arXiv preprint arXiv:2206.14359* (2022).
[82] Lawhern, V. J. et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of neural engineering* 15.5 (2018), p. 056013.
[83] Lee, C.-H. et al. "NeuroXAI: Adaptive, robust, explainable surrogate framework for determination of channel importance in EEG application". In: *Expert Systems with Applications* 261 (2025), p. 125364.
[84] Letham, B. et al. "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model". In: *The Annals of Applied Statistics* 9.3 (2015), pp. 1350–1371. DOI: 10.1214/15-AOAS848.
[85] Li, Q. et al. "Resting-state EEG functional connectivity predicts post-traumatic stress disorder subtypes in veterans". In: *Journal of neural engineering* 19.6 (2022), p. 066005.
[86] Liartis, J. et al. "Semantic Queries Explaining Opaque Machine Learning Classifiers." In: *DAO-XAI*. 2021.

[87] Liartis, J. et al. “Searching for explanations of black-box classifiers in the space of semantic queries”. In: *Semantic Web* 15.4 (2024), pp. 1085–1126.
[88] Loh, H. W. et al. “ADHD/CD-NET: automated EEG-based characterization of ADHD and CD using explainable deep neural network technique”. In: *Cognitive Neurodynamics* 18.4 (2024), pp. 1609–1625.
[89] Ludwig, S. A. “Explainability using SHAP for epileptic seizure recognition”. In: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. 2022, pp. 5305–5311.
[90] Lundberg, S. M. and Lee, S.-I. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. URL:
[91] Lyberatos, V. et al. “Perceptual musical features for interpretable audio tagging”. In: *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE. 2024, pp. 878–882.
[92] Lyberatos, V. et al. “Challenges and Perspectives in Interpretable Music Auto-Tagging Using Perceptual Features”. In: *IEEE Access* 13 (2025), pp. 60720–60732. DOI: 10.1109/ACCESS.2025.3555741.
[93] Lyberatos, V. et al. “Challenges and Perspectives in Interpretable Music Auto-Tagging Using Perceptual Features”. In: *IEEE Access* 13 (2025), pp. 60720–60732. DOI: 10.1109/ACCESS.2025.3555741.
[94] Lyberatos, V. et al. “Music interpretation and emotion perception: A computational and neurophysiological investigation”. In: *arXiv preprint arXiv:2506.01982* (2025).
[95] Lymperaiou, M. et al. “Towards Explainable Evaluation of Language Models on the Semantic Similarity of Visual Concepts”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by N. Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3639–3658. URL:
[96] Mansilla, D. et al. “Generalizability of electroencephalographic interpretation using artificial intelligence: An external validation study”. In: *Epilepsia* 65.10 (2024), pp. 3028–3037.
[97] Mansour, M., Khnaisser, F., and Partamian, H. “An explainable model for eeg seizure detection based on connectivity features”. In: *arXiv preprint arXiv:2009.12566* (2020).
[98] Marcus, C. L. et al. “A randomized trial of adenotonsillectomy for childhood sleep apnea”. In: *New England Journal of Medicine* 368.25 (2013), pp. 2366–2376.
[99] Mastromichalakis, O. M., Liartis, J., and Stamou, G. “Beyond One-Size-Fits-All: Adapting Counterfactual Explanations to User Objectives”. In: *arXiv preprint arXiv:2404.08721* (2024).
[100] Mastromichalakis, O. M., Liartis, J., and Stamou, G. “Beyond One-Size-Fits-All: How User Objectives Shape Counterfactual Explanations”. In: *XAI 2025: The 3rd World Conference on eXplainable Artificial Intelligence - Late-Breaking Work*. 2025.
[101] Mastromichalakis, O. M. et al. “Rule-based explanations of machine learning classifiers using knowledge graphs”. In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1. 2024, pp. 193–202.
[102] Mayor-Torres, J. M. et al. “Interpretable sincnet-based deep learning for emotion recognition from EEG brain activity”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2021.
[103] Mazurek, S. et al. “Explainable graph neural networks for EEG classification and seizure detection in epileptic patients”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.
[104] Menis Mastromichalakis, O. et al. “Semantic prototypes: Enhancing transparency without black boxes”. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024, pp. 1680–1688.
[105] Molnar, C. *Interpretable machine learning*. Lulu. com, 2020.
[106] Montavon, G., Samek, W., and Müller, K.-R. “Methods for Interpreting and Understanding Deep Neural Networks”. In: *Digital Signal Processing* 73 (2018), pp. 1–15. DOI: 10.1016/j.dsp.2017.10.011.
[107] Mumtaz, W. “MDD patients and healthy controls EEG data (new)”. In: *figshare, Dataset* (2016).
[108] Mumtaz, W. et al. “A wavelet-based technique to predict treatment outcome for major depressive disorder”. In: *PloS one* 12.2 (2017), e0171409.
[109] Murugan, T. K. and Kameswaran, A. “Employing convolutional neural networks and explainable artificial intelligence for the detection of seizures from electroencephalogram signal”. In: *Results in Engineering* 24 (2024), p. 103378.
[110] Nahmias, D. O. and Kontson, K. L. “Easy perturbation EEG algorithm for spectral importance (easyPEASI) a simple method to identify important spectral features of EEG in deep learning models”.

In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 2398–2406.

[111] Nauta, B., Willemsen, A., and Marsman, M. "Anecdotal Evidence and User Studies in Explainable AI: Evaluating Trust and Usability". In: *Proceedings of the 2022 Conference on Human Factors in Computing Systems (CHI)*. 2022.

[112] Obeid, I. and Picone, J. "The temple university hospital EEG data corpus". In: *Frontiers in neuroscience* 10 (2016), p. 196.

[113] Olejarczyk, E. and Jernajczyk, W. "Graph-based analysis of brain connectivity in schizophrenia". In: *PloS one* 12.11 (2017), e0188629.

[114] Opherk, C., Coromilas, J., and Hirsch, L. J. "Heart rate and EKG changes in 102 seizures: analysis of influencing factors". In: *Epilepsy research* 52.2 (2002), pp. 117–127.

[115] Pandey, P. and Miyapuram, K. P. "Nonlinear EEG analysis of mindfulness training using interpretable machine learning". In: *2021 IEEE International conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2021, pp. 3051–3057.

[116] Panigutti, C. et al. "Co-design of human-centered, explainable AI for clinical decision support". In: *ACM Transactions on Interactive Intelligent Systems* 13.4 (2023), pp. 1–35.

[117] Park, D. et al. "Spatio-temporal explanation of 3D-EEGNet for motor imagery EEG classification using permutation and saliency". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).

[118] Patakis, A. et al. *Semantic-Aware Interpretable Multimodal Music Auto-Tagging*. 2025. arXiv: 2505.17233 [cs.LG]. URL:

[119] Pathak, S. et al. "STQS: Interpretable multi-modal Spatial-Temporal-seQuential model for automatic Sleep scoring". In: *Artificial intelligence in medicine* 114 (2021), p. 102038.

[120] Pedregosa, F. et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12 (2011), pp. 2825–2830.

[121] Pérez-Velasco, S. et al. "Unraveling motor imagery brain patterns using explainable artificial intelligence based on Shapley values". In: *Computer Methods and Programs in Biomedicine* 246 (2024), p. 108048.

[122] Phadikar, S., Sinha, N., and Ghosh, R. "Unsupervised feature extraction with autoencoders for EEG based multiclass motor imagery BCI". In: *Expert Systems with Applications* 213 (Mar. 2023), p. 118901. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.118901. URL:

[123] Pilcevic, D. et al. "Performance evaluation of metaheuristics-tuned recurrent neural networks for electroencephalography anomaly detection". In: *Frontiers in Physiology* 14 (2023), p. 1267011.

[124] Poria, S. et al. "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances". In: *IEEE Access* 7 (2019), pp. 100943–100953. DOI: 10.1109/ACCESS.2019.2929050.

[125] Prokhorenkova, L. et al. "CatBoost: unbiased boosting with categorical features". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6638–6648.

[126] Quan, S. F. et al. "The sleep heart health study: design, rationale, and methods". In: *Sleep* 20.12 (1997), pp. 1077–1085.

[127] Raab, D., Theissler, A., and Spiliopoulou, M. "XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series". In: *Neural Computing and Applications* 35.14 (2023), pp. 10051–10068.

[128] Radwan, M., Lind, P. G., and Yazidi, A. "An Interpretable Graph Based Model for Classification Of EEG using Directional Functional Connectivity". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.

[129] Rajpura, P. and Meena, Y. K. "Towards Optimising EEG Decoding using Post-hoc Explanations and Domain Knowledge". In: *arXiv preprint arXiv:2405.01269* (2024).

[130] Ravindran, S., Akshay, and Contreras-Vidal, J. "An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth". In: *Scientific Reports* 13.1 (2023), p. 17709.

[131] Redline, S. et al. "The Childhood Adenotonsillectomy Trial (CHAT): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population". In: *Sleep* 34.11 (2011), pp. 1509–1517.

[132] Ribeiro, M. T., Singh, S., and Guestrin, C. ""Why Should I Trust You?" Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[133] Ronzhina, M. et al. “Sleep scoring using artificial neural networks”. In: *Sleep medicine reviews* 16.3 (2012), pp. 251–263.
[134] Rudin, C. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.
[135] Saeidi, M. et al. “Neural decoding of EEG signals with machine learning: a systematic review”. In: *Brain Sciences* 11.11 (2021), p. 1525.
[136] Salih, A. M. et al. “A perspective on explainable artificial intelligence methods: SHAP and LIME”. In: *Advanced Intelligent Systems* 7.1 (2025), p. 2400304.
[137] Samek, W. et al. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2021.
[138] Sánchez-Hernández, S. E. et al. “Evaluation of the Relation between Ictal EEG Features and XAI Explanations”. In: *Brain Sciences* 14.4 (2024), p. 306.
[139] Sancho, M. L. et al. “Identifying Reproducibly Important EEG Markers of Schizophrenia with an Explainable Multi-Model Deep Learning Approach”. In: *bioRxiv* (2024).
[140] Sattiraju, A. et al. “An Explainable and Robust Deep Learning Approach for Automated Electroencephalography-based Schizophrenia Diagnosis”. In: *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2023, pp. 255–259.
[141] Schirrmeister, R. T. et al. “Deep learning with convolutional neural networks for EEG decoding and visualization”. In: *Human brain mapping* 38.11 (2017), pp. 5391–5420.
[142] Selvaraju, R. R. et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.
[143] Senadheera, I. et al. “AI Applications in Adult Stroke Recovery and Rehabilitation: A Scoping Review Using AI”. In: *Sensors (Basel, Switzerland)* 24.20 (2024), p. 6585.
[144] Shen, J. et al. “Explainable depression recognition from EEG signals via graph convolutional network”. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2023.
[145] Shen, J. et al. “Exploring the intrinsic features of EEG signals via empirical mode decomposition for depression recognition”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2022), pp. 356–365.
[146] Shrikumar, A., Greenside, P., and Kundaje, A. “Learning important features through propagating activation differences”. In: *International conference on machine learning*. PMlR. 2017, pp. 3145–3153.
[147] Shrikumar, A. et al. “Not just a black box: Learning important features through propagating activation differences”. In: *arXiv preprint arXiv:1605.01713* (2016).
[148] Slack, D. et al. “Fooling lime and shap: Adversarial attacks on post hoc explanation methods”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 180–186.
[149] Sotirou, T. et al. “Musiclime: Explainable multimodal music understanding”. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.
[150] Stevenson, N. J. et al. “A dataset of neonatal EEG recordings with seizure annotations”. In: *Scientific data* 6.1 (2019), pp. 1–8.
[151] Stieger, J. R. et al. “Mindfulness improves brain–computer interface performance by increasing control over neural activity in the alpha band”. In: *Cerebral Cortex* 31.1 (2021), pp. 426–438.
[152] Sturm, I. et al. “Interpretable deep neural networks for single-trial EEG classification”. In: *Journal of neuroscience methods* 274 (2016), pp. 141–145.
[153] Sui, J. et al. “Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection”. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2014, pp. 3889–3892.
[154] Sujatha Ravindran, A. and Contreras-Vidal, J. “An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth”. In: *Scientific Reports* 13.1 (2023), p. 17709.
[155] Sundararajan, M., Taly, A., and Yan, Q. “Axiomatic Attribution for Deep Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 3319–3328. URL:
[156] Sylvester, S. et al. “SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI methods”. In: *Behavior Research Methods* 56.6 (2024), pp. 6067–6081.
[157] Tangermann, M. et al. “Review of the BCI competition IV”. In: *Frontiers in neuroscience* 6 (2012), p. 55.

[158] Tapia, C. G., Bozic, B., and Longo, L. "Investigating the Effect of Pre-processing Methods on Model Decision-Making in EEG-Based Person Identification". In: *World Conference on Explainable Artificial Intelligence*. Springer. 2023, pp. 131–152.

[159] Torres, J. M. M. et al. "Evaluation of interpretability for deep learning algorithms in EEG emotion recognition: A case study in autism". In: *Artificial intelligence in medicine* 143 (2023), p. 102545.

[160] Tran, L. V. et al. "Application of machine learning in epileptic seizure detection". In: *Diagnostics* 12.11 (2022), p. 2879.

[161] Tuncer, T. et al. "TATPat based explainable EEG model for neonatal seizure detection". In: *Scientific Reports* 14.1 (2024), p. 26688.

[162] Tveit, J. et al. "Automated interpretation of clinical electroencephalograms using artificial intelligence". In: *JAMA neurology* 80.8 (2023), pp. 805–812.

[163] Vaquerizo-Villar, F. et al. "An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea". In: *Computers in Biology and Medicine* 165 (2023), p. 107419.

[164] Vardhan, U. H., Femi, P. S., and Kala, A. "Human Stress Detection in and Through Sleep using Artificial Intelligence". In: *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)* (2023), pp. 1608–1612. DOI: 10.1109/ICESC57686.2023.10193562.

[165] Varon, C. et al. "Can ECG monitoring identify seizures?" In: *Journal of electrocardiology* 48.6 (2015), pp. 1069–1074.

[166] Vázquez, M. A., Maghsoudi, A., and Mariño, I. P. "An interpretable machine learning method for the detection of schizophrenia using EEG signals". In: *Frontiers in Systems Neuroscience* 15 (2021), p. 652662.

[167] Vieira, J. C. et al. "Using explainable artificial intelligence to obtain efficient seizure-detection models based on electroencephalography signals". In: *Sensors* 23.24 (2023), p. 9871.

[168] Wang, G., Deng, Z., and Choi, K.-S. "Detection of epilepsy with Electroencephalogram using rule-based classifiers". In: *Neurocomputing* 228 (2017), pp. 283–290.

[169] Wang, H. et al. *Rethinking Saliency Map: An Context-aware Perturbation Method to Explain EEG-based Deep Learning Model*. 2022. arXiv: 2205.14976 [cs.LG]. URL:

[170] Wang, P. et al. "A Comprehensive Survey on Emerging Techniques and Technologies in Spatio-Temporal EEG Data Analysis". In: *Chinese Journal of Information Fusion* 1.3 (Dec. 2024), pp. 183–211. ISSN: 2998-3363. DOI: 10.62762/cjif.2024.876830. URL:

[171] Wang, T. et al. "Bayesian Rule Sets for Interpretable Classification". In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016, pp. 1269–1274. DOI: 10.1109/ICDM.2016.0179.

[172] Wei, L. and Mooney, C. "An EEG-based Automatic Classification Model for Epilepsy with Explainable Artificial Intelligence". In: *Proceedings of the 2024 14th International Conference on Biomedical Engineering and Technology*. 2024, pp. 44–50.

[173] Yang, Y., Song, X., and He, H. "A Survey on Quantitative Metrics for Explainable AI". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.7 (2022), pp. 2905–2919.

[174] Zhang, J. et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59 (2020), pp. 103–126.

[175] Zhang, S. et al. "Visual explanations of deep convolutional neural network for EEG brain fingerprint". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.

[176] Zhang, X. et al. "Adversarial representation learning for robust patient-independent epileptic seizure detection". In: *IEEE journal of biomedical and health informatics* 24.10 (2020), pp. 2852–2859.

[177] Zheng, W.-L. and Lu, B.-L. "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks". In: *IEEE Transactions on autonomous mental development* 7.3 (2015), pp. 162–175.

[178] Zhou, X. et al. "An EEG channel selection framework for driver drowsiness detection via interpretability guidance". In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2023, pp. 1–5.

[179] Ziyabari, S. et al. "Objective evaluation metrics for automatic classification of EEG events". In: *arXiv preprint arXiv:1712.10107* (2017).

[180] Zong, J. et al. "FCAN–XGBoost: a novel hybrid model for EEG emotion recognition". In: *Sensors* 23.12 (2023), p. 5680.

[181] Zorzos, I. et al. “Applying neural networks with time-frequency features for the detection of mental fatigue”. In: *Applied Sciences* 13.3 (2023), p. 1512.