ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΔΙΚΤΥΩΣΗ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανίχνευση κρίσεων επιληψίας σε δεδομένα ηλεκτροεγκεφαλογράφου

Χρήστος Δ. Πατσούρας

***Επιβλέπουσα*: Αθανασία Αλωνιστιώτη,** Αναπληρώτρια Καθηγήτρια

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2020

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανίχνευση κρίσεων επιληψίας σε δεδομένα ηλεκτροεγκεφαλογράφου

Χρήστος Δ. Πατσούρας
Α.Μ.: M1612

ΕΠΙΒΛΕΠΟΥΣΑ: Αθανασία Αλωνιστιώτη, Αναπληρώτρια Καθηγήτρια ΕΚΠΑ

Δεκέμβριος 2020

ΠΕΡΙΛΗΨΗ

Βασικός τομέας ενασχόλησης της παρούσας διπλωματικής εργασίας είναι η ανίχνευση επιληπτικών κρίσεων με χρήση μεθόδων μηχανικής μάθησης. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από επιφανειακά ηλεκτροεγκεφαλογραφήματα (EEG). Πρόκειται για τη CHB-MIT βάση δεδομένων που διατίθεται δωρεάν στην πλατφόρμα PhysioNet. Στα πλαίσια της υλοποίησης εξετάστηκε όλη η διαδικασία διαχείρισης των δεδομένων από τη λήψη τους, την εξαγωγή χαρακτηριστικών (μέση τιμή, διακύμανση, ασυμμετρία, κύρτωση, τυπική απόκλιση, διάμεσος, διασχίσεις μηδενικού άξονα, ενεργός τιμή, εύρος μεταβολής, εντροπία δείγματος, ισχύς μέσω PSD στις συχνότητες δέλτα, θήτα (theta), άλφα (alpha), βήτα (beta), γάμμα (gamma), μέγιστη αλληλοσυσχέτιση) σε αυτά, την κανονικοποίησή τους (z-score), τη μείωση διαστάσεων (PCA) με διατήρηση τής εγγενούς πληροφορίας τους, την εξισορρόπηση των δειγμάτων (Cluster Centroids, ADASYN), επιληπτικών και μη, έως την εκπαίδευση, τη βελτιστοποίηση (αναζήτηση πλέγματος) και την εφαρμογή των ταξινομητών (SVM, kNN, Απλοϊκός Bayes, Δέντρα Απόφασης, Τυχαίο Δάσος, LDA, Λογιστική Παλινδρόμηση, Νευρωνικό Δίκτυο με LSTM), την αξιολόγηση τους (ακρίβεια, ευαισθησία, ειδικότητα, αξιοπιστία, βαθμολογία F1, συντελεστής συσχέτισης Matthews, συντελεστής κ του Cohen) και τη σύγκριση των αποτελεσμάτων. Διενεργούνται τρία διαφορετικά πειράματα είτε χρησιμοποιώντας τις μετρήσεις από όλα τα ηλεκτρόδια είτε τμήμα αυτών. Η βασική διαφορά της μεθόδου μας σε σχέση με τη βιβλιογραφία είναι ότι εξετάζονται τα αποτελέσματα της γενίκευσης σε αντίθεση με τις εστιασμένες στον κάθε ασθενή μεθόδους που συνήθως συναντάται. Η υλοποίηση όλων των παραπάνω γίνεται μέσω της γλώσσας Python, που είναι η δημοφιλέστερη για εφαρμογές μηχανικής μάθησης, και της πλατφόρμας Jupyter.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ανίχνευση Κρίσεων Επιληψίας, Ηλεκτροεγκεφαλογράφημα, Εγκεφαλική Δραστηριότητα, Εξαγωγή Χαρακτηριστικών, Αλγόριθμοι Ταξινόμησης

ABSTRACT

The main field of work of this thesis is the detection of seizures using machine learning methods. The data we used came from scalp electroencephalograms (EEGs). This is the CHB-MIT database, which is available for free, from the PhysioNet platform. In the context of the implementation, the whole process of data management was examined from their download, the extraction of characteristics (mean, variance, skewness, kurtosis, standard deviation, median, zero crossings, root mean square, peak to peak, sample entropy, power via PSD in the delta, theta, alpha, beta, gamma frequencies, maximum correlation) from them, their normalization (z-score), the reduction of dimensions (PCA) by preserving their inherent information, the balancing of epileptic and non-epileptic samples (Cluster Centroids, ADASYN) to training, optimization (grid search) and classifier implementation (SVM, kNN, Naive Bayes, Decision Trees, Random Forest, LDA, Logistic Regression, Neural Network with LSTM), their evaluation (accuracy, sensitivity/recall, specificity, precision, F1 score, Matthews correlation coefficient, Cohen's Kappa coefficient) and comparison of results. Three different experiments are performed either by using the measurements of all the electrodes or part of them. The main difference of our method in relation to the bibliography is that the results of the generalization of the methods are examined in contrast to the focused ones on each patient that is usually encountered. All of the above is done using Python, which is the most popular of machine learning applications, and the Jupyter platform.

SUBJECT AREA: Machine Learning

KEYWORDS: Seizure Detection, EEG, Brain Activity, Feature Extraction, Classification Algorithms

Στους φίλους και τους συνεργάτες στο όμορφο ταξίδι των φοιτητικών μου χρόνων

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας Διπλωματικής Εργασίας, την ανάθεση ενός θέματος στα μέτρα των ενδιαφερόντων μου, αλλά και για την αρωγή στα πρώτα μου επαγγελματικά βήματα θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια Νάνσυ Αλωνιστιώτη.

Θα ήταν παράλειψη μου να μην εξάρω την ιδιαίτερη συμβολή, τη διαρκή και εμπεριστατωμένη καθοδήγηση και την εν γένει υποστήριξη των φίλων μου, συμφοιτητών και συνεργατών της καθηγήτριας, Λίνας Μαγουλά και Νικόλα Κουρσιουμπά.

Τέλος, ιδιαίτερα σημαντική υπήρξε η ενθάρρυνση και η κατανόηση φίλων, συναδέλφων και οικογένειας στην περάτωση και των μεταπτυχιακών σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα Διπλωματική εργασία εκπονήθηκε στην Αθήνα κατά το ακαδημαϊκό έτος 2019-2020. Αποτελεί απαραίτητη προϋπόθεση για τη λήψη του μεταπτυχιακού διπλώματος στη Δικτύωση Υπολογιστών του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών.

Κατά τη διάρκεια των φοιτητικών μου σπουδών και παρακολουθώντας πληθώρα μαθημάτων διαφορετικών μεταξύ τους, τόσο σε προπτυχιακό όσο και σε μεταπτυχιακό επίπεδο, ένα από τα θέματα που μου κέντρισε σημαντικά το ενδιαφέρον ήταν ο τρόπος που λειτουργεί ο ανθρώπινος εγκέφαλος, ο τρόπος που αντιλαμβάνεται, επεξεργάζεται και αντιδρά στα ερεθίσματα. Παράλληλα, η πρόοδος, η ευρεία διάδοση και οι πολλαπλοί τομείς εφαρμογής της μηχανικής μάθησης μου κέντρισε την επιθυμία να ασχοληθώ περαιτέρω με το αντικείμενο. Ο συνδυασμός των παραπάνω δύο ενδιαφερόντων μου είχε ως αποτέλεσμα την επιλογή αυτής της εργασίας ως καταληκτική για την ολοκλήρωση των μεταπτυχιακών σπουδών μου.

1. ΕΙΣΑΓΩΓΗ

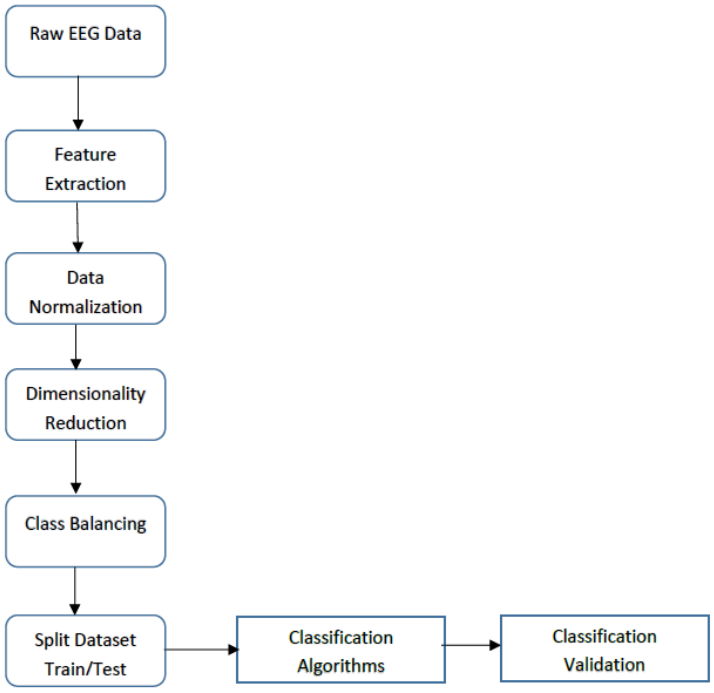
Ξεκινώντας τη διπλωματική εργασία, θεώρησα χρήσιμο στην εισαγωγική ενότητα να εξηγηθεί ποιο είναι το θέμα που εξετάζεται, η δομή της λύσης, η αρχιτεκτονική των πειραμάτων που διεξήχθησαν καθώς και να δοθεί μια πρώτη ιδέα στον αναγνώστη για το τι πρόκειται να διαβάσει στα επόμενα κεφάλαια.

1.1 Αντικείμενο ενασχόλησης

Αντικείμενο ενασχόλησης τής εργασίας είναι η ανίχνευση επιληπτικών κρίσεων σε δεδομένα προερχόμενα από ηλεκτροεγκεφαλογράφημα. Τα δεδομένα συλλέχθηκαν στο παιδικό νοσοκομείο Μασαχουσέτης (CHB) από το MIT και είναι ελεύθερα διαθέσιμα στην πλατφόρμα PhysioNet. Για τη διαδικασία ανίχνευσης χρησιμοποιήθηκαν κάποιοι δημοφιλείς αλγόριθμοι μηχανικής μάθησης που προτείνονται από τη βιβλιογραφία για παρεμφερή πειράματα και αξιολογήθηκαν με επτά διαφορετικές μετρικές. Για την υλοποίηση όλης αυτής της διαδικασίας χρησιμοποιήθηκε η γλώσσα Python 3.7 και η πλατφόρμα Jupyter. Σκοπός είναι η σύγκριση μερικών από τις μεθόδους που έχουν προταθεί στη βιβλιογραφία και η γενίκευσή τους από εστιασμένες στον κάθε ασθενή σε σύνολο δεδομένων που περιέχει περισσότερους ασθενείς. Ο κώδικας θα είναι διαθέσιμος στο GitHub μετά τη βαθμολόγηση της εργασίας.

1.2 Αρχιτεκτονική Πειράματος

Σε αυτό το σημείο, είναι σημαντικό να κατανοήσει ο αναγνώστης τον τρόπο που προσεγγίστηκε το πρόβλημα στην παρούσα διπλωματική εργασία. Στο Σχήμα 1 παρουσιάζονται τα διαδοχικά στάδια που ακολουθήθηκαν ώστε να παραχθεί το τελικό αποτέλεσμα, που είναι η σύγκριση των διάφορων αλγορίθμων στο πρόβλημα . Στις επόμενες υποενότητες της Ενότητας 1.2 θα μιλήσουμε ακροθιγώς για τα επιμέρους τμήματα του σχήματος 1, τα οποία θα αναλυθούν εκτενέστερα, με έμφαση στη μεθοδολογία και παρουσίαση αποτελεσμάτων στα κεφάλαια που ακολουθούν.



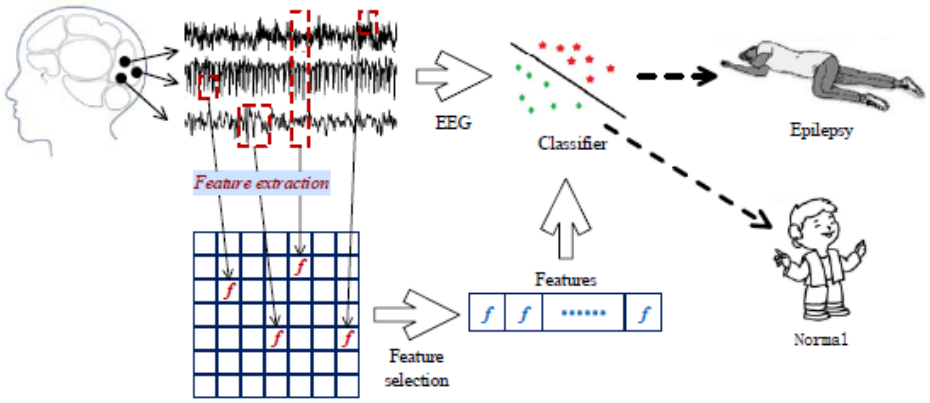
Σχήμα 1: Αρχιτεκτονική Πειράματος

1.2.1 Ανεπεξέργαστα EEG δεδομένα

Το πρώτο επίπεδο της αρχιτεκτονικής που παρουσιάζεται στο Σχήμα 1 αφορά τη συλλογή τού συνόλου ανεπεξέργαστων δεδομένων (raw data). Όπως έχει ήδη αναφερθεί από την Ενότητα 1.1, το σύνολο δεδομένων είναι το CHB-MIT Scalp EEG Database. Πρόκειται για δεδομένα που συλλέχτηκαν στο Παιδικό Νοσοκομείο της Βοστώνης από το MIT σε 23 νέους από 1,5 έως 22 ετών που αντιμετώπιζαν κρίσεις επιληψίας και είναι διαθέσιμα ελεύθερα στην πλατφόρμα PhysioNet. Τα δεδομένα είναι δειγματοληπτημένα στα 256 Hz σε ανάλυση 16 bit με τα περισσότερα δείγματα να περιέχουν δεδομένα από 23 κανάλια αισθητήρων (24 ή 26 σε μερικές περιπτώσεις). Σε αυτά έγινε μια προεπεξεργασία κρατώντας μόνο τα κανάλια που είναι κοινά σε όλους τους ασθενείς και επιλέγοντας μόνο τις καταγραφές που περιείχαν και επιληπτικές κρίσεις. Περισσότερες λεπτομέρειες για τα δεδομένα και τη διαχείρισή τους θα ακολουθήσουν στο Κεφάλαιο 3.

1.2.2 Μέθοδοι Εξαγωγής Χαρακτηριστικών

Επόμενο επίπεδο της αρχιτεκτονικής μας είναι εκείνο της εξαγωγής χαρακτηριστικών. Στόχος είναι ο μετασχηματισμός τού συνόλου δεδομένων σε ένα άλλο, τροποποιημένο κατά τέτοιο τρόπο, ώστε να βελτιστοποιηθεί η πληροφορία που παίρνουμε, να μειωθεί ο πλεονασμός δεδομένων και να αξιοποιηθούν καλύτερα οι δυνατότητες των αλγορίθμων ταξινόμησης μηχανικής μάθησης.



Εικόνα 1: Συνολική διαδικασία ταξινόμησης σήματος EEG

Για τη μέγιστη αξιοποίηση και την ακριβέστερη περιγραφή τού συνόλου δεδομένων παίρνουμε χαρακτηριστικά στο πεδίο του χρόνου και του φάσματος. Στα πειράματα που αυτό είναι εφικτό χρησιμοποιούνται και χαρακτηριστικά συσχέτισης των καναλιών. Για να εξαχθούν τα χαρακτηριστικά, τεμαχίζουμε το σήμα σε μη επικαλυπτόμενα παράθυρα των 2 δευτερολέπτων. Όπως προαναφέρθηκε στην υποενότητα 1.2.1, το σήμα είναι δειγματοληπτημένα στα 256 Hz. Άρα, για κάθε $256\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau\alpha/sec \cdot 2\,\mathrm{sec} = 512\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau\alpha$ παράγεται μια τιμή χαρακτηριστικού. Στις υποενότητες 1.2.2.1 – 1.2.2.3 θα καταγράψουμε τα χαρακτηριστικά που υπολογίσαμε. Αναλυτικότερα σε αυτά θα αναφερθούμε στο Κεφάλαιο 4.

1.2.2.1 Χαρακτηριστικά στο πεδίο του χρόνου

Τα χαρακτηριστικά που υπολογίστηκαν στο πεδίο του χρόνου είναι:

- Αριθμητική μέση τιμή (Mean)
- Διακύμανση (Variance)
- Ασυμμετρία (Skewness)

- Κύρτωση (Kurtosis)
- Τυπική Απόκλιση (Standard Deviation)
- Διάμεσος (Median)
- Διασχίσεις μηδενικού άξονα (Zero Crossings)
- Ενεργός τιμή σήματος (Root Mean Square)
- Εύρος μεταβολής (Peak to Peak)
- Εντροπία Δείγματος (Sample Entropy)

1.2.2.2 Φασματικά χαρακτηριστικά

Όσον αφορά τα φασματικά χαρακτηριστικά, υπολογίστηκε η ισχύς μέσω PSD στις παρακάτω συχνότητες:

- Δέλτα (0.5-4 Hz)
- Θήτα (4-8 Hz)
- Άλφα (8-12 Hz)
- Βήτα (12-30 Hz)
- Γάμμα (30-100 Hz)

Ο προσεγγιστικός τρόπος υπολογισμού τής φασματικής πυκνότητας ισχύος (power spectral density) γίνεται μέσω της μεθόδου Welch και το ολοκλήρωμα του PSD στις παραπάνω συχνότητες, που δίνει το εμβαδόν και κατ' επέκταση την ισχύ τους, υπολογίζεται μέσω του κανόνα του Simpson.

1.2.2.3 Διμερή χαρακτηριστικά

Το μοναδικό διμερές χαρακτηριστικό που υπολογίστηκε είναι η μέγιστη αλληλοσυσχέτιση (maximum cross-correlation). Η μέγιστη συσχέτιση είναι η εξάρτηση μεταξύ ενός ζεύγους καναλιών EEG, λαμβάνοντας υπόψη τις χρονικές καθυστερήσεις μετατοπίζοντας ένα από τα δύο σήματα. Αυτό το χαρακτηριστικό χρησιμοποιήθηκε σε ένα μόνο από τα τρία πειράματα για λόγους που θα εξηγηθούν στο Κεφάλαιο 4.

1.2.3 Κανονικοποίηση Χαρακτηριστικών

Η κανονικοποίηση (normalization) είναι μια διαδικασία μετασχηματισμού δεδομένων, κατά την οποία, αριθμητικές τιμές αντικαθίστανται με άλλες, πιο «κατάλληλες», έτσι ώστε όλες οι μεταβλητές να ανήκουν στο ίδιο εύρος τιμών. Η κανονικοποίηση εφαρμόζεται στα χαρακτηριστικά αφαιρώντας το μέσο όρο και κλιμακώνοντας στη μοναδιαία διακύμανση.

Η τυπική τιμή ενός δείγματος x υπολογίζεται ως $z = {(x - u)}/{s}$, όπου x η μέση τιμή του δείγματος, s η τυπική απόκλιση και z η νέα τιμή του δείγματος. Αυτή είναι η λεγόμενη z-score κανονικοποίηση.

1.2.4 Μείωση Διαστάσεων

Τα διανύσματα χαρακτηριστικών που παράγονται από το στάδιο εξαγωγής χαρακτηριστικών περιέχουν πολλές τυχαίες μεταβλητές και μεταφέρουν μεγάλη

πληροφορία. Ωστόσο, έχει μεγάλη σημασία να βρεθεί μια συσχέτιση μεταξύ αυτών των μεταβλητών και να μειωθεί ο αριθμός της τυχαιότητας υπό εξέταση. Αυτή η διαδικασία μπορεί να επιτευχθεί εφαρμόζοντας μια μέθοδο μείωσης διαστάσεων. Η μείωση διαστάσεων επιλέγει τα πιο σημαντικά στοιχεία τού χώρου των χαρακτηριστικών, ώστε να τα διατηρήσουμε και πετώντας τα υπόλοιπα στοιχεία. Έτσι, κρατάμε τη σημαντικότερη πληροφορία, μειώνοντας παράλληλα και το υπολογιστικό κόστος και το ενδεχόμενο υπερπροσαρμογής (overfitting). Στην περίπτωσή μας, αποφασίζουμε να εκτελέσουμε μια μη εποπτευόμενη τεχνική γραμμικού μετασχηματισμού που ονομάζεται Principal Component Analysis (PCA).

1.2.4.1 PCA

Το PCA είναι μια στατιστική μέθοδος που χρησιμοποιεί τη διαδικασία γραμμικού, ορθογώνιου μετασχηματισμού για να μετασχηματίσει ένα σύνολο χαρακτηριστικών υψηλότερης διάστασης, που θα μπορούσε ενδεχομένως να συσχετιστεί σε ένα σύνολο κατώτερων διαστάσεων γραμμικών μη συσχετισμένων χαρακτηριστικών. Είναι ουσιαστικά μια γρήγορη και ευέλικτη μη εποπτευόμενη μέθοδος για μείωση των διαστάσεων στα δεδομένα, αλλά μπορεί επίσης να είναι χρήσιμη ως εργαλείο οπτικοποίησης, φιλτραρίσματος θορύβου, εξαγωγής χαρακτηριστικών και πολλά άλλα.

Ένα ουσιαστικό μέρος της χρήσης του PCA στην πράξη είναι η ικανότητα εκτίμησης του πλήθους συνιστωσών απαιτούνται για την περιγραφή των δεδομένων. Αυτό μπορεί να προσδιοριστεί εξετάζοντας την αθροιστική αναλυτική αναλογία διακύμανσης ως συνάρτηση τού αριθμού των στοιχείων. Η καμπύλη αυτή ποσοτικοποιεί το μέρος τής συνολικής διακύμανσης που περιέχεται στα πρώτα N στοιχεία. Με βάση αυτήν την καμπύλη, μπορούμε να βρούμε τον ελάχιστο αριθμό συνιστωσών που απαιτούνται ώστε να διατηρήσουμε την επιθυμητή ποσότητα πληροφορίας.

1.2.5 Εξισορρόπηση Κλάσεων

Τα μη ισορροπημένα σύνολα δεδομένων είναι μια συνηθισμένη δυσκολία στα προβλήματα ταξινόμησης μηχανικής μάθησης, όπου υπάρχει μια δυσαναλογία δειγμάτων σε κάθε κλάση. Σε αυτήν την περίπτωση, πολλοί αλγόριθμοι τείνουν μόνο να προβλέπουν την επικρατούσα κλάση, επομένως, έχουν μεγάλη εσφαλμένη ταξινόμηση τής μειονεκτούσας τάξης σε σύγκριση με την πλειοψηφική τάξη.

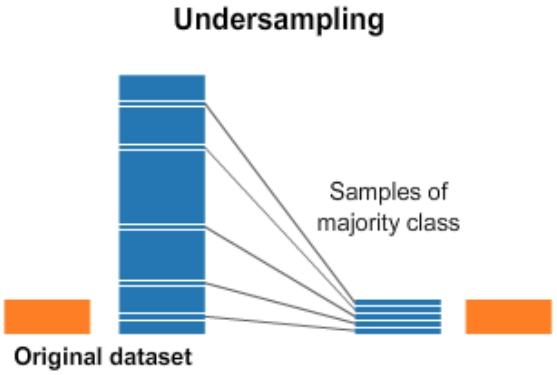
Τα μη επιληπτικά δείγματα στο CHB-MIT σύνολο δεδομένων ξεπερνούν το 98% των συνολικών δειγμάτων, εξετάζοντας μόνο τις καταγραφές που έχουν ένα τμήμα με επιληπτική κρίση. Για να αποφύγουμε την προκατάληψη των ML αλγορίθμων πρέπει να έχουμε ακριβώς τον ίδιο αριθμό δειγμάτων και στις 2 κλάσεις. Αυτό θα επιτευχθεί μέσω της υποδειγματοληψίας της επικρατούσας κλάσης και της υπερδειγματοληψίας της μειονεκτούσας κλάσης. Οι λόγοι που επιλέχθηκε ο συνδυασμός των δύο τεχνικών και όχι η μία από αυτές για να εξισορροπηθεί το σύνολο δεδομένων θα παρουσιαστούν στο Κεφάλαιο 4, όπου θα αναλυθεί εκτενέστερα η μεθοδολογία.

1.2.5.1 Υποδειγματοληψία επικρατούσας κλάσης

Για την υποδειγματοληψία της επικρατούσας κλάσης δοκιμάστηκαν τρεις μεθοδολογίες:

- Cluster Centroids
- Near Miss
- Random Undersampler

Ως προκαθορισμένη μέθοδος υποδειγματοληψίας επελέγη η Cluster Centroids, αλλά διατίθεται ο κώδικας και για τις άλλες δύο. Σκοπός της υποδειγματοληψίας είναι να δώσουμε μια συγκεκριμένη τιμή, που ορίζεται στο config αρχείο ως undersampling_rate, στο λόγο τού αριθμού των στοιχείων της μειονεκτούσας κλάσης προς τον αριθμό των στοιχείων της επικρατούσας κλάσης μετά την υποδειγματοληψία. Ουσιαστικά και οι τρεις μεθοδολογίες στοχεύουν στη συγκεκριμένη αναλογία, αλλά την επιτυγχάνουν με διαφορετικό τρόπο, επιλέγοντας διαφορετικά δείγματα από την επικρατούσα τάξη.



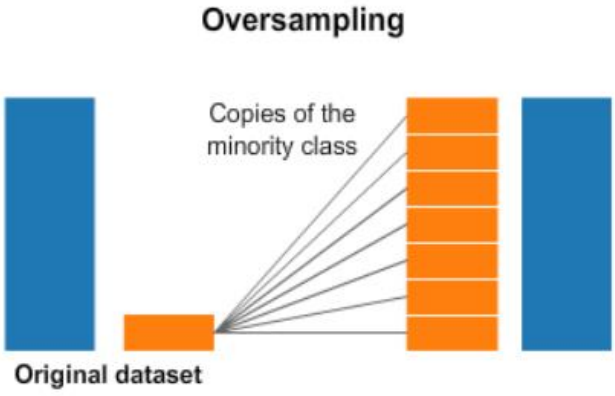
Σχήμα 2: Υποδειγματοληψία επικρατούσας κλάσης

1.2.5.2 Υπερδειγματοληψία μειονεκτούσας κλάσης

Για την υπερδειγματοληψία της μειονεκτούσας κλάσης δοκιμάστηκαν δυο μεθοδολογίες:

- ADASYN
- SMOTE

Ως προκαθορισμένη μέθοδος υπερδειγματοληψίας επελέγη η ADASYN, αλλά διατίθεται ο κώδικας και για τη μέθοδο SMOTE. Αφού υποδειγματοληπτήσουμε την επικρατούσα κλάση, στόχος είναι να εξισώσουμε τα στοιχεία των δύο κλάσεων, δημιουργώντας συνθετικά δεδομένα με βάση τα υπάρχοντα της μειονεκτούσας κλάσης.



Σχήμα 3: Υπερδειγματοληψία μειονεκτούσας κλάσης

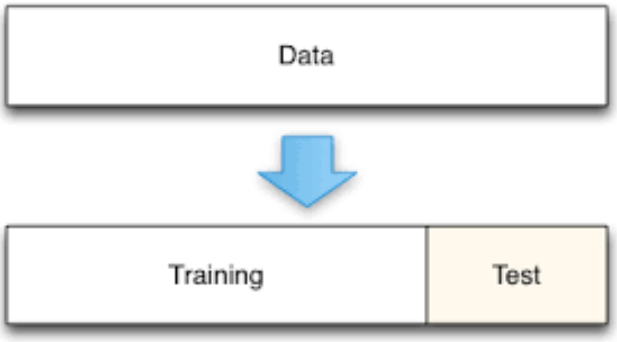
1.2.6 Διαχωρισμός δεδομένων εκπαίδευσης και δεδομένων αξιολόγησης

Αφού παράχθηκε ένα ισορροπημένο σύνολο δεδομένων, αυτό πρέπει να χωριστεί σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης ώστε να μπορέσουμε να εφαρμόσουμε τους αλγορίθμους εποπτευόμενης μηχανικής μάθησης για την ταξινόμηση των δειγμάτων

σε επιληπτικά ή όχι. Θα δοκιμάσουμε δύο διαφορετικές μεθοδολογίες και θα συγκρίνουμε τα αποτελέσματα των μετρικών. Η πρώτη μεθοδολογία είναι ο απλός διαχωρισμός των δεδομένων σε δύο τμήματα με το μεγαλύτερο ποσοστό να πηγαίνει στα δεδομένα εκπαίδευσης. Η δεύτερη μεθοδολογία είναι η διασταυρούμενη επικύρωση, στην οποία χωρίζουμε το σύνολο δεδομένων σε κ ίσα κομμάτια, τρέχουμε κ επαναλήψεις με ένα διαφορετικό τμήμα αξιολόγησης κάθε φορά και τα υπόλοιπα κ-1 χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Θα επανέλθουμε στο συγκεκριμένο θέμα στο Κεφάλαιο 5.

1.2.6.1 Διαχωρισμός συνόλου δεδομένων

Ο διαχωρισμός τού συνόλου δεδομένων χωρίζει το σύνολο δεδομένων σε δύο άλλα σύνολα, το σύνολο εκπαίδευσης και το σύνολο αξιολόγησης. Συνήθως τα δεδομένα που χρησιμοποιούνται για εκπαίδευση των ταξινομητών είναι σημαντικά περισσότερα από εκείνα που χρησιμοποιούνται για την αξιολόγησή τους.



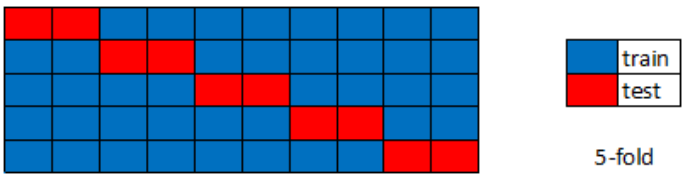
Σχήμα 4: Διαχωρισμός συνόλου δεδομένων σε δεδομένα εκπαίδευσης και αξιολόγησης

1.2.6.2 Διασταυρούμενη επικύρωση

Το σύνολο δεδομένων χωρίζεται σε κ μικρότερα σύνολα. Ακολουθείται η ακόλουθη διαδικασία για κάθε ένα από τα κ τμήματα:

- Ένα μοντέλο εκπαιδεύεται χρησιμοποιώντας κ-1 τμήματα ως δεδομένα εκπαίδευσης.
- Το μοντέλο που προκύπτει επικυρώνεται στο υπόλοιπο μέρος των δεδομένων (δηλαδή, χρησιμοποιείται ως σύνολο αξιολόγησης για τον υπολογισμό ενός μέτρου απόδοσης, όπως η ακρίβεια).

Μετά το τέλος της παραπάνω διαδικασίας, οι μετρικές αξιολόγησης που θέλουμε να υπολογίσουμε προκύπτουν ως ο μέσος όρος των μετρικών που υπολογίσαμε σε κάθε ένα από τα κ βήματα. Αυτή η προσέγγιση μπορεί να είναι υπολογιστικά ακριβή, αλλά παράγει πιο αξιόπιστα αποτελέσματα, αφού όλα τα δεδομένα χρησιμοποιούνται και στα δύο υποσύνολα σε κάποιο από τα βήματα.



Σχήμα 5: Διασταυρούμενη επικύρωση

1.2.7 Αλγόριθμοι Ταξινόμησης

Το πιο σημαντικό σημείο όλης της αρχιτεκτονικής είναι η ταξινόμηση ενός δείγματος ως επιληπτικό ή όχι. Αφού υπολογίσαμε τα χαρακτηριστικά και διαμορφώσαμε το σύνολο δεδομένων, εδώ θα χρησιμοποιήσουμε κάποιους αλγορίθμους μηχανικής μάθησης για να προβλέψουμε που ανήκει το κάθε δείγμα αξιολόγησης, βασιζόμενοι και στα δεδομένα εκπαίδευσης. Μελετώντας τη βιβλιογραφία, οδηγηθήκαμε στο συμπέρασμα ότι οι πιο συνηθισμένοι και με τα καλύτερα αποτελέσματα αλγόριθμοι σε EEG, οι οποίοι και εφαρμόστηκαν, είναι οι ακόλουθοι:

- Μηχανές Διανυσμάτων Υποστήριξης (SVM)
- Κ Εγγύτεροι Γείτονες (kNN)
- Απλοϊκός Bayes (Naïve Bayes)
- Δέντρα Απόφασης (Decision Trees)
- Τυχαίο Δάσος (Random Forest)
- Ανάλυση Γραμμικής Διάκρισης (LDA)
- Λογιστική Παλινδρόμηση (Logistic Regression)
- Νευρωνικά Δίκτυα με LSTM (Long Short Term Memory)

Για το θεωρητικό υπόβαθρο των ως άνω αλγορίθμων αλλά και την υλοποίηση τους θα μάθουμε περισσότερα στο Κεφάλαιο 5.

1.2.8 Μέθοδοι Αξιολόγησης

Αφού δοκιμαστούν οι αλγόριθμοι ταξινόμησης, το τελευταίο βήμα είναι να αξιολογηθούν τα αποτελέσματα που παράγουν, δηλαδή το πόσο καλά διαχωρίζουν τα επιληπτικά δείγματα. Η επιλογή μας ήταν πάλι να υλοποιήσουμε τις πιο συχνά εμφανιζόμενες στη βιβλιογραφία μετρικές:

- Ακρίβεια (Accuracy)
- Ευαισθησία/Ανάκληση (Sensitivity/Recall)
- Ειδικότητα (Specificity)
- Αξιοπιστία (Precision)
- Βαθμολογία F1 (F1 score)
- Συντελεστής Συσχέτισης Matthews (Matthews Correlation Coefficient)
- Συντελεστής κ του Cohen (Cohen’s Kappa Coefficient)

Για το πως υπολογίζονται αυτές οι τιμές και το τι συμπεράσματα βγάζουμε από αυτές θα μιλήσουμε στο Κεφάλαιο 6.

1.3 Διάρθρωση κειμένου

Αφού μιλήσαμε περιληπτικά σε αυτό το κεφάλαιο για το αντικείμενο ενασχόλησης της εργασίας και τη βασική αρχιτεκτονική, για να έχουμε ένα πλήρες εισαγωγικό κεφάλαιο, θα πρέπει να αναφέρουμε εν τάχει τι ακολουθεί στα επόμενα κεφάλαια. Στο Κεφάλαιο 2 γίνεται αναφορά σε κάποιες θεωρητικές έννοιες-κλειδιά της εργασίας και συγκεκριμένα στις επιληπτικές κρίσεις, στο EEG και στις έννοιες της τεχνητής νοημοσύνης, της

μηχανικής μάθησης, της βαθιάς μάθησης και τη σύνδεσή τους με το EEG. Στο Κεφάλαιο 3 θα μιλήσουμε για το σύνολο δεδομένων, πως καταλήξαμε σε αυτό, πως το προσπελάσαμε και πως το διαχειριστήκαμε. Το Κεφάλαιο 4 περιγράφει όλα τα βήματα τής μετατροπής του συνόλου δεδομένων σε ένα κανονικοποιημένο και ισορροπημένο σύνολο χαρακτηριστικών στο οποίο μπορούμε να εφαρμόσουμε τα πειράματά μας. Στο Κεφάλαιο 5 θα αναλύσουμε το θεωρητικό υπόβαθρο και την υλοποίηση των αλγορίθμων ταξινόμησης των δειγμάτων σε επιληπτικά ή μη. Στο Κεφάλαιο 6, ο αναγνώστης μπορεί να βρει τις μετρικές αξιολόγησης της ταξινόμησης, πως υπολογίζονται και τι συμπέρασμα βγάζουμε από κάθε μία από αυτές. Στο Κεφάλαιο 7 θα αναφερθούμε στα τρία διαφορετικά πειράματα που υλοποιήσαμε και θα παραθέσουμε τα αποτελέσματά τους. Τέλος, στο Κεφάλαιο 8 θα βγάλουμε τα τελικά συμπεράσματα, θα αναδείξουμε τα προβλήματα που προέκυψαν και θα δώσουμε μια βάση για μελλοντική έρευνα.

2. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Σκοπός αυτής της ενότητας είναι η παρουσίαση των βασικών εννοιών της πτυχιακής εργασίας και της μεταξύ τους συσχέτισης. Μια ματιά στο τι συμβαίνει στο ανθρώπινο σώμα και εγκέφαλο κατά την επιληπτική κρίση, στο πως καταγράφουμε την εγκεφαλική δραστηριότητα και πως αυτή η πληροφορία μπορεί να συνδυαστεί με τη μηχανική μάθηση, παρέχει μια βάση για την κατανόηση του θέματος που πραγματεύεται η εργασία.

2.1 Κρίση Επιληψίας

Η ασθένεια στην οποία οι ασθενείς υποφέρουν από επιληπτικές κρίσεις που προκαλούνται από διαταραχή τής λειτουργίας τού εγκεφάλου ονομάζεται επιληψία. Η επιληψία είναι η τρίτη πιο κοινή διαταραχή τού εγκεφάλου μετά το Αλτσχάιμερ και το εγκεφαλικό που επηρεάζει περίπου το 1% του παγκόσμιου πληθυσμού (πάνω από 50 εκατομμύρια ανθρώπους) και από αυτούς περίπου το 0,2% χάνουν τη ζωή τους. Υπάρχουν πολλές πιθανές αιτίες επιληψίας, μία εκ των οποίων είναι μια μοριακή μετάλλαξη, η οποία οδηγεί σε ακανόνιστη νευρωνική συμπεριφορά ή μετανάστευση νευρώνων. Εναλλακτικά, μπορεί να αναπτυχθεί ως αποτέλεσμα εγκεφαλικού τραύματος όπως σοβαρό χτύπημα στο κεφάλι, εγκεφαλικό επεισόδιο, εγκεφαλική λοίμωξη ή εγκεφαλική κακοήθεια. Αν και η κύρια αιτία της επιληψίας παραμένει άγνωστη, η έγκαιρη διάγνωση μπορεί να είναι χρήσιμη για τη θεραπεία της. Οι ασθενείς με επιληψία μπορούν να την αντιμετωπίσουν με φάρμακα ή χειρουργικές επεμβάσεις. Ωστόσο, αυτές οι μέθοδοι δεν είναι πλήρως αποτελεσματικές. Δυστυχώς, οι επιληπτικές κρίσεις που δεν μπορούν να θεραπευτούν πλήρως περιορίζουν την ενεργό ζωή του ασθενούς. Σε αυτές τις περιπτώσεις, οι ασθενείς δεν μπορούν ανεξάρτητα να εργαστούν και να κάνουν κάποια δραστηριότητα. Αυτό οδηγεί σε κοινωνική απομόνωση ατόμων και οικονομικές δυσκολίες.

Οι επιληπτικές κρίσεις εμφανίζονται λόγω διαταραχής στη λειτουργικότητα τού εγκεφάλου που μπορεί να επηρεάσει την υγεία του ασθενούς. Ενδεχόμενες αιτίες είναι η διακοπή της παροχής του αίματος, ο υψηλός πυρετός, η στέρηση ύπνου, η έλλειψη οξυγόνου, κάποιος τραυματισμός, λοιμώξεις του εγκεφάλου και δηλητηρίαση. Η πρόβλεψη των επιληπτικών κρίσεων πριν από την έναρξη τους είναι αρκετά χρήσιμη για την αποτροπή τους με φαρμακευτική αγωγή πριν αυτή συμβεί. Οι επιληπτικές κρίσεις έχουν τέσσερις διαφορετικές καταστάσεις:

1. Κατάσταση πριν από την κρίση (preictal state): Εμφανίζεται πριν από την έναρξη της κρίσης
2. Κατάσταση κρίσης (ictal state): Ξεκινά με την εμφάνιση και τελειώνει με την επίθεση της επιληπτικής κρίσης
3. Κατάσταση μετά την κρίση (postictal state): Ξεκινά μετά την κατάσταση κρίσης
4. Μεσοκρισική κατάσταση (interictal state): Ξεκινά μετά την postictal κατάσταση τής πρώτης κρίσης και τελειώνει πριν από την έναρξη της κατάστασης διαδοχικών κρίσεων

Υψηλό ποσοστό επιληπτικών ασθενών στις αναπτυσσόμενες χώρες και φτωχοί ασθενείς στις ανεπτυγμένες χώρες παραμένουν χωρίς θεραπεία. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, περίπου τα τρία τέταρτα των ατόμων με επιληψία σε χώρες χαμηλού εισοδήματος ενδέχεται να μην λάβουν κατάλληλη θεραπεία. Κατά συνέπεια είναι σημαντική η πρόληψη τους.

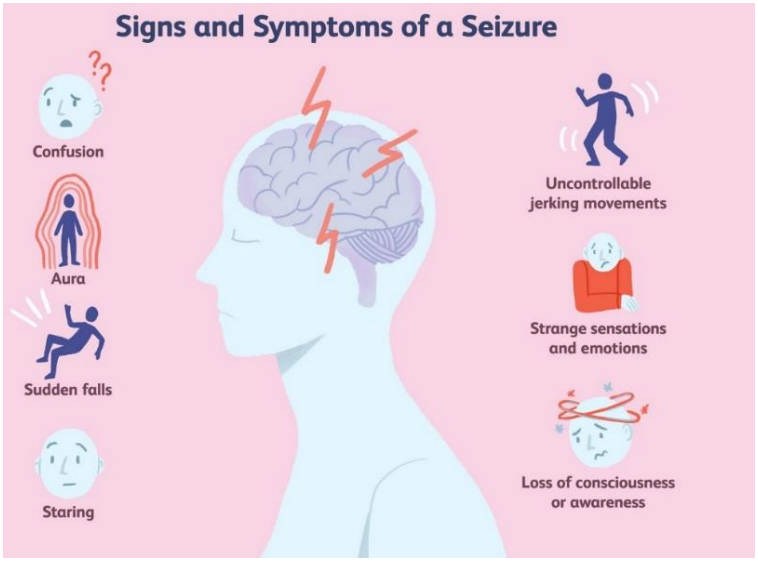
Η πρόβλεψη των επιληπτικών κρίσεων έχει προσελκύσει αυξανόμενη προσοχή ως μία από τις πιο απαιτητικές προσπάθειες ανάλυσης δεδομένων για τη βελτίωση της ζωής των

ασθενών με ανθεκτική στα φάρμακα επιληψία. Παρά την εισαγωγή νέων φαρμάκων τις τελευταίες δεκαετίες, το ένα τρίτο των ατόμων με επιληψία εξακολουθούν να έχουν επιληπτικές κρίσεις παρά τη θεραπεία. Ωστόσο, ακόμη και όταν οι επιληπτικές κρίσεις ελέγχονται καλά, η ποιότητα ζωής των ασθενών μειώνεται σημαντικά από το άγχος που σχετίζεται με την απρόβλεπτη φύση των επιληπτικών κρίσεων και τις συνέπειές τους.

Οι επιληπτικές κρίσεις μπορούν να λάβουν πολλές διαφορετικές μορφές και επηρεάζουν διαφορετικούς ανθρώπους με διαφορετικούς τρόπους. Η σοβαρότητα των επιληπτικών κρίσεων μπορεί να ποικίλει από το γέλιο έως τον ξαφνικό απροσδόκητο θάνατο στην επιληψία (SUDEP). Οι επιθέσεις μπορούν να διαρκέσουν από δευτερόλεπτα έως λεπτά. Οι επιληπτικές κρίσεις μπορεί να οδηγήσουν σε απώλεια προσοχής ή σε σπασμούς ολόκληρου του σώματος. Τα συμπτώματα κατά τη διάρκεια μιας επιληπτικής κρίσης είναι συνήθως στερεοτυπικά (εμφανίζονται με τον ίδιο ή παρόμοιο τρόπο κάθε φορά), επεισοδιακά (έρχονται και φεύγουν) και μπορεί να είναι απρόβλεπτα. Όπως προείπαμε, τα συμπτώματα δεν είναι κοινά σε όλους τους ασθενείς, αλλά μπορούμε να διακρίνουμε τρεις γενικές κατηγορίες συμπτωμάτων, οι οποίες, βέβαια, δεν είναι απαραίτητο να εμφανίζονται σε όλους τους ασθενείς:

1. <u>Πριν την έναρξη της κρίσης</u>: Μερικοί άνθρωποι μπορεί να βιώσουν συναισθήματα, αισθήσεις ή αλλαγές στη συμπεριφορά ώρες ή μέρες πριν από μια κρίση. Αυτά τα συναισθήματα γενικά δεν αποτελούν μέρος τής κρίσης, αλλά μπορεί να προειδοποιήσουν ένα άτομο ότι μπορεί να προκληθεί κρίση. Δεν έχουν όλοι αυτά τα σημάδια, αλλά, αν υπάρχουν, μπορούν να βοηθήσουν ένα άτομο να αλλάξει τη δραστηριότητά του, να φροντίσει να πάρει τα φάρμακά του, να χρησιμοποιήσει μια θεραπεία διάσωσης και να λάβει μέτρα για την πρόληψη τραυματισμού. Από την άλλη πλευρά, ορισμένοι άνθρωποι μπορεί να μην γνωρίζουν την αρχή και επομένως δεν έχουν καμία προειδοποίηση. Η αύρα είναι το πρώτο σύμπτωμα μιας κρίσης και θεωρείται μέρος της κρίσης. Συχνά η αύρα είναι ένα απερίγραπτο συναίσθημα. Άλλες φορές είναι εύκολο να αναγνωριστεί και μπορεί να είναι μια αλλαγή στο συναίσθημα, την αίσθηση, τη σκέψη ή τη συμπεριφορά που είναι παρόμοια κάθε φορά που εμφανίζεται μια κρίση. Μια αύρα μπορεί να συμβεί πριν από μια αλλαγή στη συνείδηση, ωστόσο, πολλοί άνθρωποι δεν έχουν αύρα ή προειδοποίηση. Στους τελευταίους η κρίση ξεκινά με απώλεια συνείδησης. Μερικά από τα πιο συνηθισμένα συμπτώματα πριν από μια κρίση είναι το Déjà vu (συναίσθημα ότι ένα άτομο, μέρος ή πράγμα είναι οικείο, αλλά δεν το έχετε ξαναδεί ποτέ), το Jamais vu (αίσθημα ότι ένα άτομο, μέρος ή πράγμα είναι νέο ή άγνωστο, αλλά δεν είναι), κάποιες περίεργες μυρωδιές, ήχοι, γεύσεις και συναισθήματα, απώλεια ή θόλωση όρασης, ζάλη, ναυτία, πονοκέφαλος και μούδιασμα σε κάποια μέρη του σώματος.
2. <u>Κατά τη διάρκεια της κρίσης</u>: Είναι η περίοδος από τα πρώτα συμπτώματα (συμπεριλαμβανομένης της αύρας) έως το τέλος της κρίσης. Αυτό σχετίζεται με την ηλεκτρική δραστηριότητα της κρίσης στον εγκέφαλο. Μερικές φορές τα ορατά συμπτώματα διαρκούν περισσότερο από μια κρίση σε ένα EEG. Αυτό οφείλεται στο γεγονός ότι ορισμένα από τα ορατά συμπτώματα μπορεί να είναι συνέπειες μιας κρίσης ή να μην σχετίζονται καθόλου με την επιληπτική δραστηριότητα. Ορισμένα συνηθισμένα συμπτώματα αυτής της κατηγορίας είναι απώλεια συνείδησης (συχνά αναφέρεται και ως “black out”), η σύγχυση, οι περίοδοι αμνησίας, η απόσπαση προσοχής ή το απλανές βλέμμα, η λιποθυμία, τα προβλήματα σε ακοή, όσφρηση και γεύση, οι λάμψεις, οι παραισθήσεις, το μούδιασμα, η δυσκολία στην ομιλία, τα προβλήματα στην κατάποση, οι σπασμοί, οι αυξημένοι καρδιακοί παλμοί σε συνδυασμό με δυσκολίες στην αναπνοή, η ακράτεια σάλιων, ούρων ή κοπράνων, κάποιες επαναλαμβανόμενες μη σκόπιμες κινήσεις (αυτοματισμοί), ενδεχόμενες ξαφνικές πτώσεις και πολλά άλλα.

3. Μετά το τέλος της κρίσης: Καθώς η κρίση τελειώνει, εμφανίζεται η φάση μετά την κρίση (postictal, όπως την αναφέραμε παραπάνω). Αυτή είναι η περίοδος αποκατάστασης μετά την κρίση. Μερικοί άνθρωποι αναρρώνουν αμέσως, ενώ άλλοι μπορεί να χρειαστούν λεπτά έως ώρες για να αισθανθούν όπως συνήθως. Ο τύπος της επιληπτικής κρίσης, καθώς και το ποιο μέρος του εγκεφάλου επηρεάζει η κρίση, καθορίζει την περίοδο ανάρρωσης - πόσο καιρό μπορεί να διαρκέσει και τι μπορεί να συμβεί κατά τη διάρκεια αυτής. Κάποια συνηθισμένα συμπτώματα αυτής της κατηγορίας είναι οι αργές αντιδράσεις (είτε σε κινήσεις είτε σε ομιλία) ή η αδυναμία απάντησης, το αίσθημα νύστας ή ταραχής του ασθενούς, η απώλεια μνήμης, το αίσθημα ζάλης, ναυτίας ή πονοκεφάλου σε συνδυασμό με αίσθημα ανησυχίας ή σύγχυσης, οι μικροτραυματισμοί, όπως μώλωπες, εκδορές, σπασίματα ή τραυματισμοί στο κεφάλι, αν υπήρξε πτώση κατά τη διάρκεια της κρίσης, η αδυναμία, η εξάντληση ή η δίψα κ.α.



Εικόνα 2: Σημάδια και συμπτώματα κρίσης επιληψίας

Οι επιληπτικές κρίσεις μπορούν να προβλεφθούν εντοπίζοντας την αρχή της preictal κατάστασης, ενώ η ανίχνευση τους γίνεται προσδιορίζοντας την ictal κατάσταση. Αν και στη βιβλιογραφία υπάρχει σημαντική ενασχόληση με την πρόβλεψη των κρίσεων επιληψίας, στόχος της παρούσας εργασίας είναι ο εντοπισμός τους.

2.2 Ηλεκτροεγκεφαλογράφημα

Η έναρξη της επιληπτικής κρίσης και η ανίχνευση συμβάντων επιτυγχάνονται συνήθως μέσω ανάλυσης τού ηλεκτροεγκεφαλογραφήματος. Το ηλεκτροεγκεφαλογράφημα (EEG) είναι ένα εργαλείο μέτρησης για την καταγραφή τής ηλεκτρικής δραστηριότητας τού εγκεφάλου που παρατηρείται λόγω της χημικής διακύμανσης. Ουσιαστικά είναι μια πολυκαναλική καταγραφή τής ηλεκτρικής δραστηριότητας που δημιουργείται από συλλογές νευρώνων στον εγκέφαλο. Αυτό σημαίνει ότι διαφορετικά κανάλια αντικατοπτρίζουν τη δραστηριότητα σε διαφορετικές περιοχές του εγκεφάλου.

Από τη σκοπιά των μαθηματικών, το EEG είναι μια χρονική ακολουθία διανυσμάτων $X(t)$ διάστασης M, όπου M το πλήθος των καναλιών ηλεκτροδίων. Κάθε συνιστώσα $x_i(t)$ είναι ένα μόνο κανάλι ηλεκτροδίου με σταθερό ρυθμό δειγματοληψίας. Μια κοινή υπόθεση είναι ότι τα μοτίβα συγχρονισμού των εγκεφαλικών κυμάτων είναι διαφορετικά στις φάσεις πριν

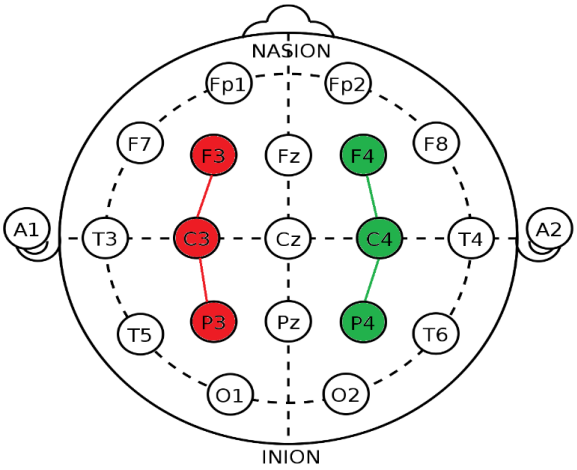
την επιληπτική κρίση και στη μεσοκρισική φάση, γεγονός που βοηθά στο να είμαστε σε θέση να τις διαχωρίσουμε.

Το EEG έχει πολλά πλεονεκτήματα έναντι άλλων μεθόδων καταγραφής της εγκεφαλικής δραστηριότητας, η χρονική ανάλυσή του είναι υψηλότερη και μετρά άμεσα τη δραστηριότητα. Αν και είναι κοινώς αποδεκτό ότι συχνά τα EEG παρουσιάζουν θόρυβο, σφάλματα και ατέλειες στο παραγόμενο σήμα, τα προαναφερθέντα πλεονεκτήματα το καθιστούν ένα εξαιρετικά χρήσιμο και διαδεδομένο κλινικό εργαλείο.

Όταν το EEG μετριέται χρησιμοποιώντας μη επεμβατικά ηλεκτρόδια που είναι τοποθετημένα στο τριχωτό της κεφαλής ενός ατόμου, αναφέρεται ως επιφανειακό EEG. Όταν μετριέται χρησιμοποιώντας ηλεκτρόδια τοποθετημένα στην επιφάνεια του εγκεφάλου ή στα βάθη του αναφέρεται ως ενδοκρανιακό EEG. Το επιφανειακό EEG είναι ένα μη επεμβατικό εργαλείο χαμηλού κόστους που μπορεί να χρησιμοποιηθεί για μακροπρόθεσμη αξιολόγηση της πορείας ενός ασθενούς. Επειδή, λοιπόν, τα επιφανειακά EEG είναι πιο συνηθισμένα και τα περισσότερα ελεύθερα σύνολα δεδομένων περιέχουν τέτοιου είδους καταγραφές, θα ασχοληθούμε μόνο με αυτά.

2.2.1 Επιφανειακό ηλεκτροεγκεφαλογράφημα

Το επιφανειακό ηλεκτροεγκεφαλογράφημα (scalp EEG) είναι μια μη επεμβατική μέθοδος μέτρησης των ηλεκτρικών κυμάτων που δημιουργούνται από τη δραστηριότητα των δεκάδων εκατομμυρίων νευρώνων του εγκεφάλου. Η ως άνω μέθοδος καταγράφεται συνήθως μέσω ηλεκτροδίων που είναι συμμετρικά τοποθετημένα στο τριχωτό της κεφαλής, όπως φαίνεται στην Εικόνα 3. Ένα σήμα EEG (εναλλακτικά γνωστό ως κανάλι) σχηματίζεται λαμβάνοντας τη διαφορά δυναμικών ανάμεσα σε δύο ηλεκτρόδια. Για παράδειγμα, το κανάλι FP1 - F7 σχηματίζεται λαμβάνοντας τη διαφορά δυναμικού μεταξύ των ηλεκτροδίων FP1 και F7. Κάθε κανάλι EEG συνοψίζει τη δραστηριότητα που εντοπίζεται σε μια περιοχή του εγκεφάλου. Παραδείγματος χάρη, το κανάλι FP1 - F7 αντανακλά τη νευρική δραστηριότητα που προέρχεται από τον μετωπιαίο λοβό του αριστερού ημισφαιρίου. Η έναρξη μιας εστιακής κρίσης συνεπάγεται μια αλλαγή δραστηριότητας στα λίγα κανάλια EEG του τριχωτού της κεφαλής που βρίσκονται πάνω ή κοντά στην περιοχή του εγκεφάλου που προκαλεί επιληπτική κρίση. Από την άλλη πλευρά, η έναρξη μιας γενικευμένης κρίσης συνεπάγεται δραστηριότητα σε όλα τα κανάλια EEG του τριχωτού της κεφαλής.

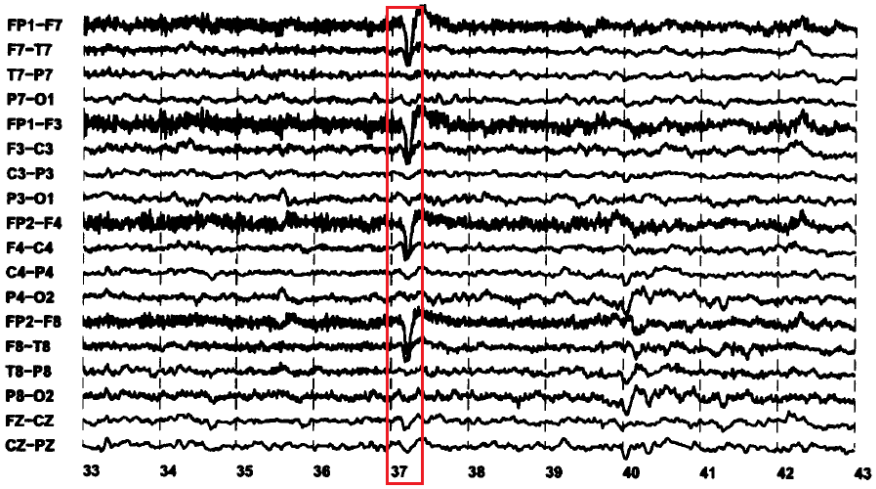


Εικόνα 3: Σύστημα τοποθέτησης ηλεκτροδίων 10-20

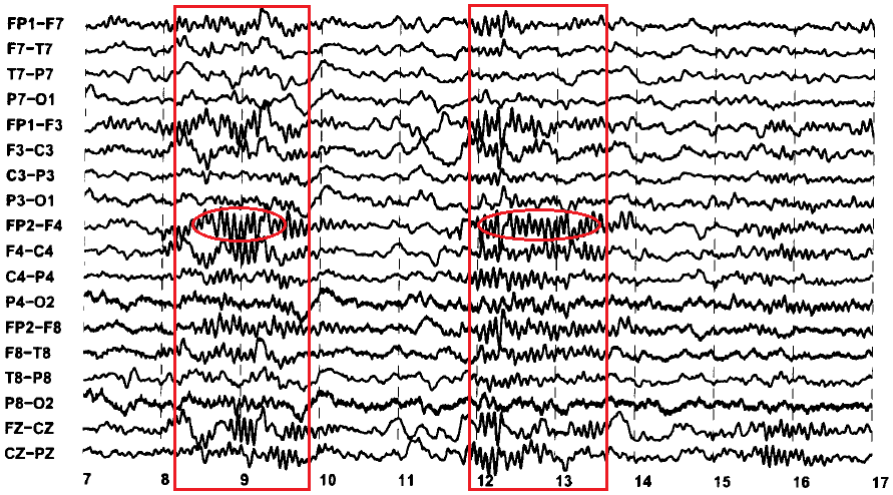
Η φυσική τής παραγωγής EEG περιορίζει τόσο την προέλευση όσο και τα χαρακτηριστικά της νευρικής δραστηριότητας που είναι ορατά μέσα στο EEG του τριχωτού της κεφαλής. Συγκεκριμένα, οι νευρώνες που συμβάλλουν περισσότερο στο επιφανειακό EEG είναι εκείνοι που βρίσκονται πλησιέστερα στην επιφάνεια του τριχωτού της κεφαλής. Αντίθετα, η δραστηριότητα των νευρώνων που έχουν θαφτεί μέσα σε βαθιές εγκεφαλικές δομές δεν είναι παρατηρήσιμη. Επιπλέον, το εγκεφαλονωτιαίο υγρό και το κρανίο που περιβάλλουν τον εγκέφαλο δρουν ως εξασθενητές που μειώνουν σε μεγάλο βαθμό το πλάτος των νευρικών ταλαντώσεων υψηλότερης συχνότητας. Μια σημαντική συνέπεια αυτών των φυσικών περιορισμών είναι ότι ορισμένοι τύποι επιληπτικών κρίσεων, συγκεκριμένα εκείνοι που περιλαμβάνουν μια μικρή, βαθιά περιοχή εντός του εγκεφάλου, δεν μπορούν να παρατηρηθούν μέσω του επιφανειακού EEG.

Οι ηλεκτροεγκεφαλογράφοι περιγράφουν την εγκεφαλική δραστηριότητα όσον αφορά τη χωρική κατανομή της στο τριχωτό της κεφαλής (μετωπική, οπίσθια, πλευρική και διμερή) καθώς και την κυρίαρχη συνιστώσα της συχνότητας. Μια EEG κυματομορφή ταξινομείται ότι έχει μια δέλτα συνιστώσα εάν το κυρίαρχο συστατικό της συχνότητας f είναι $\leq 4Hz$, μια θήτα συνιστώσα εάν $4 < f < 8\,Hz$, μια άλφα συνιστώσα όταν $8 \leq f < 12Hz$, μια βήτα συνιστώσα όταν $12 \leq f < 30Hz$ ή μια γάμμα συνιστώσα όταν $\mathrm{f} \geq 30\mathrm{Hz}$. Για παράδειγμα, η άλφα κυματομορφή στην ομιλία αναφέρεται σε ρυθμό 10Hz που εμφανίζεται πιο έντονα στα οπίσθια κανάλια όταν ένα άτομο κλείνει τα μάτια του και χαλαρώνει.

Η δραστηριότητα τού επιφανειακού EEG διαμορφώνεται από την κατάσταση εγρήγορσης ενός ατόμου. Συγκεκριμένα, η κυρίαρχη συχνότητα και η χωρική κατανομή τής ηλεκτρομαγνητικής δραστηριότητας κατά τη διάρκεια που κάποιος είναι ξύπνιος είναι διαφορετική από αυτήν κατά τον ύπνο. Για παράδειγμα, το Σχήμα 6 απεικονίζει την ηλεκτρομαγνητική δραστηριότητα ενός ξύπνιου ατόμου που διεκόπη από ένα ανοιγόκλειμα ματιών στα 37 δευτερόλεπτα. Το EEG αυτό αποτελείται κυρίως από δραστηριότητα χαμηλής συχνότητας, αλλά το ανοιγόκλειμα των ματιών οδηγεί σε μια κάμψη προς τα κάτω του σήματος κυρίως στα κανάλια FP1-F7, FP1-F3, FP2-F4 και FP2-F8. Το σχήμα 7 απεικονίζει τη δραστηριότητα EEG που καταγράφηκε κατά τη διάρκεια του ύπνου. Η ταλάντωση 11Hz, που παρατηρείται πιο έντονα στο κανάλι FP2-F4 μεταξύ 8-10 και 12-14 δευτερολέπτων, είναι γνωστή ως άτρακτοι ύπνου.



Σχήμα 6: Δραστηριότητα καναλιών EEG σε ανοιγόκλειμα ματιών

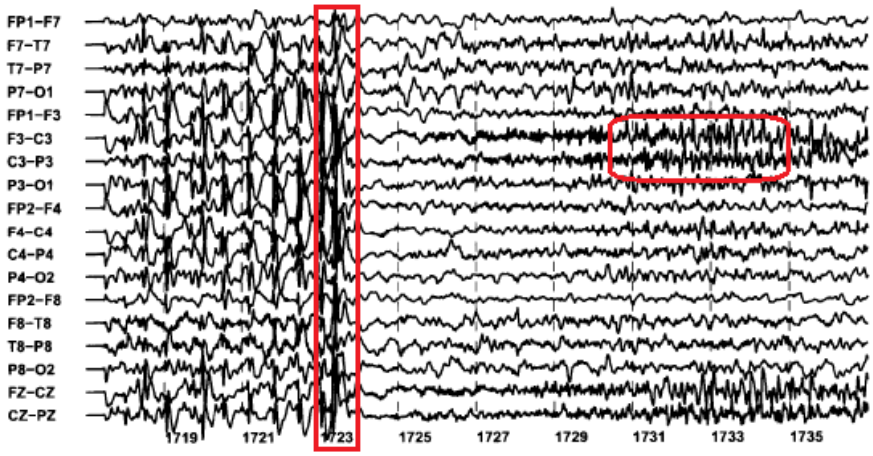


Σχήμα 7: Δραστηριότητα καναλιών EEG κατά τη διάρκεια του ύπνου

2.2.2 Επιληπτικές κρίσεις στο επιφανειακό ηλεκτροεγκεφαλογράφημα

Μέσα στο επιφανειακό EEG, οι επιληπτικές κρίσεις εκδηλώνονται ως ξαφνική ανακατανομή τής φασματικής ενέργειας σε ένα σύνολο καναλιών EEG. Η ανακατανομή της φασματικής ενέργειας προκαλείται από τον υπερσυγχρονισμό των νευρώνων εντός ενός επιληπτικού νευρικού δικτύου και συνίσταται σε μια εμφάνιση ή εξαφάνιση συνιστωσών συχνότητας εντός της ζώνης 0-25Hz. Ωστόσο, το ποιες φασματικές συνιστώσες εξαφανίζονται ή ξεχωρίζουν ποικίλλει ανάλογα με τους ασθενείς. Επιπλέον, τα EEG κανάλια που καταδεικνύουν τη φασματική αλλαγή ενέργειας ποικίλλουν επίσης μεταξύ των ασθενών καθώς είναι συνάρτηση τής εγκεφαλικής θέσης προέλευσης μιας επιληπτικής κρίσης.

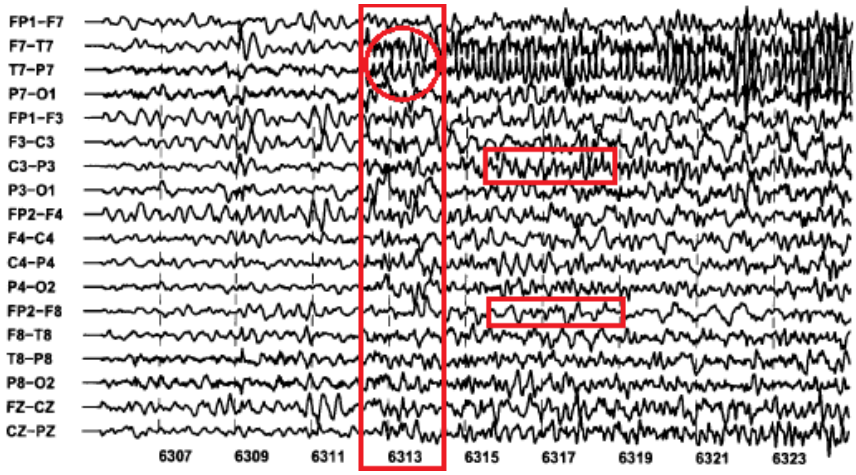
Για παράδειγμα, τα Σχήματα 8 και 9 απεικονίζουν κρίσεις από τους ασθενείς Α και Β αντίστοιχα. Η κρίση στο Σχήμα 8 ξεκινά στα 1723 δευτερόλεπτα και αποτελείται από ισοπέδωση του σήματος EEG σε όλα τα κανάλια ακολουθούμενη από την εμφάνιση ενός ρυθμού στη ζώνη βήτα στα κανάλια F3-C3, C3-P3. Στη συνέχεια, κατά τη διάρκεια μερικών δευτερολέπτων, το πλάτος αυτού του ρυθμού αυξάνεται καθώς η συχνότητά του μειώνεται και εγκαθίσταται μέσα στη θήτα.



Σχήμα 8: Επιφανειακό EEG με κρίση επιληψίας ασθενούς Α

Η επιληπτική κρίση στο Σχήμα 9 ξεκινά στα 6313 δευτερόλεπτα με την έναρξη ενός ρυθμού θήτα που είναι πιο εμφανής στα κανάλια F7-T7, T7-P7. Άλλα κανάλια EEG

εμφανίζουν επίσης μια αλλαγή μετά την έναρξη της επιληπτικής κρίσης. Το κανάλι C3-P3 αναπτύσσει έναν ρυθμό θήτα ενώ το κανάλι FP2-F8 αναπτύσσει έναν ρυθμό δέλτα ζώνης.



Σχήμα 9: Επιφανειακό EEG με κρίση επιληψίας ασθενούς B

Τα προηγούμενα παραδείγματα απεικονίζουν τη μεταβλητότητα στη φασματική και χωρική υπογραφή της επιληπτικής και της μη επιληπτικής δραστηριότητας στους ασθενείς. Αυτή η μεταβλητότητα είναι ο πρωταρχικός λόγος για τον οποίο οι ανιχνευτές που δεν είναι εξειδικευμένοι στον ασθενή παρουσιάζουν κακά αποτελέσματα για επεισόδια κρίσεων. Αντιθέτως, για ένα συγκεκριμένο άτομο, οι επιληπτικές κρίσεις που προκύπτουν από την ίδια εγκεφαλική θέση παρουσιάζουν παρόμοια κλινική συμπτωματολογία και χωρικά και φασματικά χαρακτηριστικά.

Το ηλεκτροεγκεφαλογράφημα φέρει μεγάλο αριθμό πολύπλοκων πληροφοριών που είναι πολύτιμες για την ανίχνευση των επιληπτικών κρίσεων. Ενώ υπάρχουν πολλά που πρέπει να γίνουν κατανοητά για τις επιληπτικές κρίσεις, γνωρίζουμε ότι υπάρχουν ορισμένα χαρακτηριστικά κοινά στις περισσότερες από αυτές. Κατά τη διάρκεια μιας επιληπτικής κρίσης, υπάρχει αυξημένη και συγχρονισμένη δραστηριότητα στον εγκέφαλο που είναι γνωστή ως υπερσυγχρονισμός. Η ανίχνευση τόσο της αύξησης της δραστηριότητας όσο και της παρουσίας υπερσυγχρονισμού μπορεί να παρέχει έναν αποτελεσματικό τρόπο ανίχνευσης τής παρουσίας επιληπτικής κρίσης. Δεδομένου ότι η ισχύς ενός σήματος EEG αυξάνεται - μερικές φορές δραματικά - κατά τη διάρκεια των περισσότερων δραστηριοτήτων μιας επιληπτικής κρίσης, αυτή η συμπεριφορά, που παρατηρείται και μετριέται τόσο στο τομέα του χρόνου όσο και σε αυτόν της συχνότητας, μπορεί να παράξει τα πιο δημοφιλή χαρακτηριστικά που χρησιμοποιούνται για την ανίχνευση μη φυσιολογικής εγκεφαλικής δραστηριότητας στη βιβλιογραφία.

Στα πλαίσια της παρούσας εργασίας, παρά τις δυσκολίες που αναλύθηκαν στην τελευταία παράγραφο, θα επιχειρήσουμε να εφαρμόσουμε τις τεχνικές και τους τρόπους εντοπισμού μιας κρίσης επιληψίας, μερικοί εκ των οποίων αναφέρθηκαν στην ενότητα 1.2, και θα αξιολογήσουμε τα αποτελέσματα.

2.3 EEG και Μηχανική Μάθηση

Σε αυτήν την ενότητα θα ασχοληθούμε με κάποιες έννοιες όπως η τεχνητή νοημοσύνη, η μηχανική μάθηση και η βαθιά μάθηση. Η κατανόηση των βασικών αρχών τους θα μας

βοηθήσει στην κατανόηση όλων των αλγορίθμων και των τεχνικών που χρησιμοποιήθηκαν στα πλαίσια αυτής της διπλωματικής εργασίας.

2.3.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη (Artificial Intelligence) είναι ο τομέας της επιστήμης των υπολογιστών που ασχολείται με την ανάπτυξη συστημάτων τα οποία παρουσιάζουν ευφυή χαρακτηριστικά παρόμοια με εκείνα της ανθρώπινης σκέψης. Μέσα από την κατανόηση των δυνατοτήτων της νοημοσύνης, ο άνθρωπος προσπάθησε να προσδώσει το χαρακτηριστικό του αυτό σε μηχανές, ώστε να του προσφέρουν επιπρόσθετες υπηρεσίες, πέρα από τη χρήση τους ως εργαλεία αποθήκευσης τεράστιων ποσοτήτων πληροφορίας. Η νοημοσύνη μπορεί να χαρακτηριστεί ως μια σύνθετη πνευματική λειτουργία με την οποία το άτομο μαθαίνει, κατανοεί και αντιμετωπίζει νέες καταστάσεις. Έτσι, και τα συστήματα AI μπορούν να αναλύουν σύνθετα ή και μεγάλα δεδομένα, να ανακαλύπτουν τυχόν σχέσεις που υπάρχουν μεταξύ τους, αλλά και πρότυπα και μοτίβα σε αυτά, συμβάλλοντας με τον τρόπο αυτό στη λήψη καλύτερων αποφάσεων.

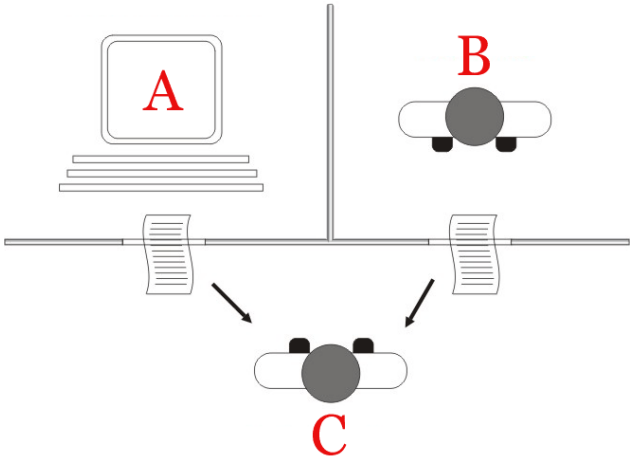
Η τεχνητή νοημοσύνη εμφανίστηκε στα μέσα του 20ου αιώνα και αποτελεί ένα διεπιστημονικό κράμα. Στη θεμελίωση της συνέδραμαν επιστήμες όπως η φιλοσοφία, τα μαθηματικά (λογική, θεωρία υπολογισμού, πιθανότητες), η οικονομία, η ψυχολογία, οι νευροεπιστήμες και φυσικά η πληροφορική με το σχεδιασμό των κατάλληλων αλγορίθμων και την αξιοποίηση των δυνατοτήτων των υπολογιστών.

Οι δυο βασικότερες προσεγγίσεις της Τεχνητής Νοημοσύνης είναι οι εξής:

- Συμβολική: Στηρίζεται στο γεγονός ότι κατανοώντας τον τρόπο λειτουργία τής ανθρώπινης νοημοσύνης μπορούμε να την προσεγγίσουμε με αλγορίθμους και συστήματα χρησιμοποιώντας σύμβολα για την αναπαράσταση των εννοιών καθώς και των σχέσεων μεταξύ τους. Ο μαθηματικός κλάδος της λογικής εξυπηρετεί στην αναπαράσταση της γνώσης και αποτελεί κλασικό παράδειγμα εφαρμογής αυτής της κατηγορίας.
- Μη συμβολική ή Υπολογιστική: Στηρίζεται στην παραγωγή τής ευφυούς συμπεριφοράς μέσω της μίμησης βιολογικών διεργασιών, όπως είναι η εξέλιξη των ειδών και η λειτουργία του εγκεφάλου. Τα τεχνητά νευρωνικά δίκτυα και οι γενετικοί αλγόριθμοι είναι κάποιες από τις τεχνικές που χρησιμοποιούνται στη συγκεκριμένη κατηγορία.

Ένα από τα πρώτα πειράματα που έγιναν και καταδεικνύουν τον τρόπο λειτουργίας της τεχνητής νοημοσύνης προτάθηκε από τον σπουδαίο μαθηματικό και εκ των θεμελιωτών τής επιστήμης των υπολογιστών, Alan Turing, στη δημοσίευση του Computing Machinery and Intelligence το 1950. Η δοκιμασία Turing, γνωστή και ως παιχνίδι μίμησης, είναι μια δοκιμασία τής ικανότητας ενός μηχανήματος να επιδεικνύει ευφυή συμπεριφορά, ισοδύναμη με την ανθρώπινη και δυσδιάκριτη από αυτή. Ο Turing πρότεινε ότι ένας ανθρώπινος αξιολογητής θα κρίνει τις φυσικές γλωσσικές συνομιλίες μεταξύ ενός ανθρώπου και ενός μηχανήματος που έχει σχεδιαστεί για να δημιουργεί ανθρώπινες αντιδράσεις. Ο αξιολογητής θα γνώριζε ότι ένας από τους δύο συμμετέχοντες στη συνομιλία είναι μια μηχανή και όλοι οι συμμετέχοντες θα διαχωρίζονταν ο ένας από τον άλλο. Η συνομιλία θα περιοριζόταν σε ένα μόνο κανάλι που θα υποστηρίζει μόνο κείμενο, όπως πληκτρολόγιο υπολογιστή και οθόνη, οπότε το αποτέλεσμα δεν θα εξαρτιόταν από την ικανότητα του μηχανήματος να αποδώσει λέξεις ως ομιλία. Εάν ο αξιολογητής δεν μπορεί αξιόπιστα να ξεχωρίσει τη μηχανή από τον άνθρωπο, η μηχανή λέγεται ότι έχει περάσει το τεστ, αλλιώς αποτυγχάνει και επιδέχεται περιθώρια βελτίωσης. Τα αποτελέσματα των δοκιμών δεν εξαρτώνται από

την ικανότητα του μηχανήματος να δίνει σωστές απαντήσεις σε ερωτήσεις, μόνο πόσο κοντά οι απαντήσεις του μοιάζουν με αυτές που θα έδινε ένας άνθρωπος.



Εικόνα 4: Δοκιμασία Τεχνητής Νοημοσύνης Turing (Παιχνίδι Μίμησης)

Η τεχνητή νοημοσύνη χρησιμοποιείται σε πολλούς τομείς, Ενδεικτικά παραδείγματα είναι η μηχανική μάθηση, η μηχανική όραση, η ρομποτική, η επεξεργασία φυσικής γλώσσας, η αναπαράσταση γνώσης, η ανάλυση δεδομένων, η εξαγωγή συμπερασμάτων, η ασφάλεια και η πρόληψη απάτης, η γεωργία, το εμπόριο και οι αγορές, η κατασκευή και παραγωγή, η υγεία και πάρα πολλές ακόμα πτυχές της καθημερινής ζωής. Η ένταξη του AI επέφερε σημαντικά επιτεύγματα στην ανθρωπότητα, αυτοματοποιώντας και διευκολύνοντας την καθημερινότητά πολλών ανθρώπων, φυσικά με επιπτώσεις που ακόμη ερευνώνται.

2.3.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning) είναι ένας υποτομέας τής Τεχνητής Νοημοσύνης - και δη από τους παλαιότερους – που επιτρέπει στους υπολογιστές να αντιμετωπίζουν εργασίες που έχουν πραγματοποιηθεί, μέχρι τώρα, μόνο από ανθρώπους. Η έρευνα σε αυτόν τον τομέα στοχεύει στη δημιουργία μηχανών ικανών να μαθαίνουν και να βελτιώνονται από την εμπειρία χωρίς να είναι ρητά προγραμματισμένες. Ειδικότερα, αξιοποιώντας ένα σύνολο δεδομένων, συνήθως γνώση που έχει αποκτηθεί από προηγούμενη εκτέλεση μιας διαδικασίας, κατασκευάζονται μοντέλα ή πρότυπα τα οποία, αν ακολουθηθούν, βελτιώνουν την απόδοσή των συστημάτων. Για τη δημιουργία ενός προτύπου αναπτύσσονται αλγόριθμοι, βασισμένοι σε ένα σύνολο εκπαίδευσης, προκειμένου τα δεδομένα να ταξινομηθούν σε κατηγορίες ή να αναγνωρισθούν τυχόν τυποποιημένες μορφές σε αυτά ή να γίνει πρόβλεψη για κάποια τάση ή συμπεριφορά αυτών. Ο πρωταρχικός στόχος είναι να επιτρέπεται στους υπολογιστές να μαθαίνουν αυτόματα χωρίς ανθρώπινη παρέμβαση ή βοήθεια και να προσαρμόζουν ανάλογα τις ενέργειες.

Ο πιο διαδεδομένος τρόπος διαχωρισμού σε είδη μάθησης περιλαμβάνει τις ακόλουθες τρεις κατηγορίες:

- Εποπτευόμενη Μάθηση (Supervised Learning): Περιλαμβάνει τη μάθηση από παραδείγματα που ανήκουν σε καθορισμένες κατηγορίες. Ο αλγόριθμος δημιουργεί μια συνάρτηση λαμβάνοντας ως εισόδους ένα σύνολο στιγμιότυπών εκπαίδευσης με γνωστές εξόδους μαθαίνοντας μέσω της σύγκρισης της εξόδου του με τις σωστές εξόδους και βρίσκοντας τα λάθη. Το σύστημα καλείται να

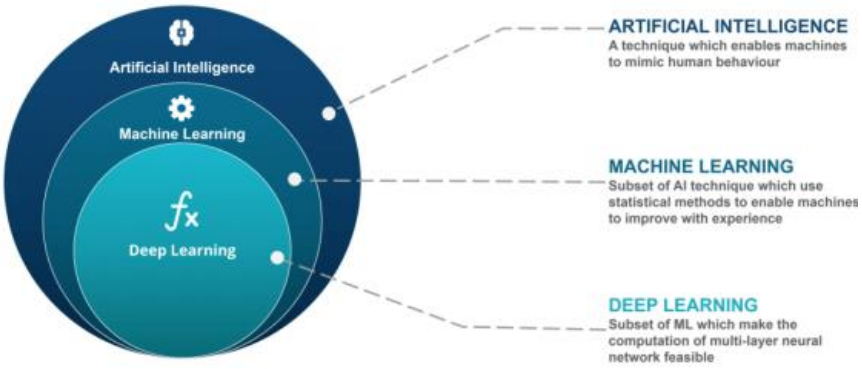
ανακαλύψει τις κοινές ιδιότητες των αντικειμένων κάθε κατηγορίας. Οι εποπτευόμενοι αλγόριθμοι μηχανικής μάθησης μπορούν να εφαρμόσουν ό,τι έχουν μάθει στο παρελθόν σε νέα δεδομένα χρησιμοποιώντας επισημασμένα παραδείγματα για την πρόβλεψη μελλοντικών γεγονότων. Ξεκινώντας από την ανάλυση ενός γνωστού συνόλου δεδομένων κατάρτισης, ο αλγόριθμος εκμάθησης παράγει μια συναγόμενη συνάρτηση για να κάνει προβλέψεις σχετικά με τις τιμές εξόδου. Στόχος της είναι η γενίκευση της συνάρτησης ώστε να απεικονίζονται και δεδομένα εισόδου με άγνωστη έξοδο. Η συγκεκριμένη μέθοδος βρίσκει εφαρμογή σε προβλήματα ταξινόμησης, πρόγνωσης και διερμηνείας.

- Μη Εποπτευόμενη Μάθηση (Unsupervised Learning): Χρησιμοποιείται σε δεδομένα που δεν είναι γνωστές οι κατηγορίες τους, δηλαδή δεν υπάρχει «σωστή απάντηση» στο πρόβλημα του διαχωρισμού σε κατηγορίες. Στη συγκεκριμένη περίπτωση το σύστημα καλείται να ανακαλύψει και να δημιουργήσει μόνο του τις κατηγορίες βασιζόμενο στις κοινές ιδιότητες των αντικειμένων. Ο αλγόριθμος προσπαθεί να ανακαλύψει τυχόν συσχετίσεις μεταξύ των στιγμιότυπών εισόδου με άγνωστη έξοδο προκειμένου να βρεθούν οι δομικοί σχηματισμοί τους. Η μη εποπτευόμενη μάθηση μελετά πώς τα συστήματα μπορούν να συμπεράνουν μια συνάρτηση για να περιγράψουν μια κρυφή δομή από δεδομένα χωρίς κατηγορία. Η συγκεκριμένη κατηγορία βρίσκει εφαρμογή σε προβλήματα ανάλυσης συσχετισμών και ομαδοποίησης.
- Ημιεποπτευόμενη Μάθηση (Semi-Supervised Learning): Βρίσκεται κάπου μεταξύ της εποπτευόμενης και της μη εποπτευόμενης μάθησης, δεδομένου ότι χρησιμοποιεί τόσο δεδομένα με καθορισμένες κατηγορίες όσο και χωρίς για εκπαίδευση - συνήθως μια μικρή ποσότητα δεδομένων με κατηγορίες και μια μεγάλη ποσότητα δεδομένων χωρίς, διότι τα δεδομένα χωρίς κατηγορία είναι λιγότερο ακριβά και απαιτούν λιγότερη προσπάθεια απόκτησης. Τα συστήματα που χρησιμοποιούν αυτήν τη μέθοδο είναι σε θέση να βελτιώσουν σημαντικά την ακρίβεια τής μάθησης. Συνήθως, η ημιεποπτευόμενη μάθηση επιλέγεται όταν τα ληφθέντα δεδομένα με κατηγορία απαιτούν εξειδικευμένες μεθόδους και πόρους για να εκπαιδευτούν, οπότε το κόστος που σχετίζεται με την κατηγοριοποίηση είναι πολύ υψηλό για να επιτρέψει πλήρη εκπαίδευση. Διαφορετικά, η απόκτηση δεδομένων χωρίς κατηγορία γενικά δεν απαιτεί πρόσθετους πόρους.

Για τις απαιτήσεις της παρούσας διπλωματικής εργασίας αξιοποιήθηκε η κατηγορία αλγορίθμων της επιβλεπόμενης μάθησης. Αυτό θα φανεί ακόμα πιο ξεκάθαρο αργότερα μιλώντας για το σύνολο δεδομένων, τη διαχείρισή του και τα προσδοκόμενα αποτελέσματα.

2.3.3 Βαθιά Μάθηση

Η βαθιά μάθηση είναι μια τεχνική μηχανικής μάθησης που διδάσκει στους υπολογιστές να κάνουν ό,τι έρχεται φυσικά στον άνθρωπο, να μαθαίνει δηλαδή με το παράδειγμα. Επιτυγχάνει αποτελέσματα που δεν ήταν δυνατό πριν. Στη βαθιά μάθηση, ένα μοντέλο υπολογιστή μαθαίνει να εκτελεί εργασίες ταξινόμησης απευθείας από τα δεδομένα, όπως εικόνες, κείμενο ή ήχο. Τα μοντέλα βαθιάς μάθησης μπορούν να επιτύχουν ακρίβεια τελευταίας τεχνολογίας, μερικές φορές υπερβαίνοντας την ανθρώπινη απόδοση. Τα μοντέλα εκπαιδεύονται χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων με ετικέτες και αρχιτεκτονικές νευρωνικών δικτύων που περιέχουν πολλά επίπεδα.



Εικόνα 5: Σχέση Τεχνητής Νοημοσύνης με Μηχανική και Βαθιά Μάθηση

Ενώ η θεωρία της βαθιάς μάθησης σχηματίστηκε για πρώτη φορά στη δεκαετία του 1980, υπάρχουν δύο κύριοι λόγοι που μόλις πρόσφατα έγινε χρήσιμη:

1. Η βαθιά μάθηση απαιτεί μεγάλες ποσότητες δεδομένων με κατηγορία. Για παράδειγμα, η ανάπτυξη αυτοκινήτου χωρίς οδηγό απαιτεί εκατομμύρια εικόνες και χιλιάδες ώρες βίντεο.
2. Η βαθιά μάθηση απαιτεί σημαντική υπολογιστική ισχύ. Οι υψηλής απόδοσης κάρτες γραφικών έχουν παράλληλη αρχιτεκτονική που είναι αποτελεσματική για βαθιά μάθηση. Όταν συνδυάζεται με clusters ή cloud computing, αυτό οδηγεί στο να μειωθεί ο χρόνος εκπαίδευσης για ένα δίκτυο βαθιάς μάθησης από εβδομάδες σε ώρες ή λιγότερο.

Η βαθιά μάθηση έχει αλλάξει ριζικά τη μηχανική μάθηση σε πολλούς τομείς (π.χ. όραση υπολογιστή, ομιλία κ.λπ.) παρέχοντας γενικού σκοπού ευέλικτα μοντέλα που μπορούν να λειτουργήσουν με ανεπεξέργαστα δεδομένα και να μάθουν τους κατάλληλους μετασχηματισμούς για ένα πρόβλημα που αντιμετωπίζεται. Αυτά τα μοντέλα μπορούν να χρησιμοποιούν μεγάλες ποσότητες δεδομένων για να μάθουν άμεσα χαρακτηριστικά και να συλλάβουν τη δομή των δεδομένων με αποτελεσματικό τρόπο ώστε να μπορούν στη συνέχεια να μεταφερθούν ή/και να προσαρμοστούν σε διαφορετικές εργασίες. Αυτή η ικανότητα μάθησης από άκρο σε άκρο ταιριάζει απόλυτα με τις απαιτήσεις της ανάλυσης EEG.

Μερικοί τομείς στους οποίους χρησιμοποιείται η βαθιά μάθηση είναι:

- η αυτοματοποιημένη οδήγηση: ανίχνευση σημάτων στοπ, φαναριών, πεζών
- αεροδιαστημική και άμυνα: εντοπισμός από δορυφόρους, προσδιορισμός ασφαλών ζωνών
- Ιατρική έρευνα: έρευνες καρκίνου
- Βιομηχανικός αυτοματισμός: βελτίωση ασφάλειας εργαζομένων, βελτίωση ταχύτητας παραγωγής
- Ηλεκτρονικά: αυτόματη μετάφραση ακοής και ομιλίας

2.3.4 Συσχέτιση EEG με Μηχανική Μάθηση

Η παγκόσμια αγορά EEG αναπτύσσεται σημαντικά, καθώς τα EEG δεδομένα χρησιμοποιούνται όλο και περισσότερο σε προληπτικές διαγνωστικές διαδικασίες. Δεδομένου ότι η χειροκίνητη ανίχνευση επιληπτικών κρίσεων σε συνεχώς παρακολουθούμενα ηλεκτροεγκεφαλογραφήματα είναι μια πολύ χρονοβόρα διαδικασία

και απαιτεί εκπαιδευμένο εξειδικευμένο προσωπικό, οι προσπάθειες ανάπτυξης αυτόματης ανίχνευσης επιληπτικών κρίσεων είναι ποικίλες και συνεχείς. Τα εργαλεία τεχνητής νοημοσύνης και μηχανικής μάθησης είναι ο ιδανικός σύντροφος για την αυτοματοποίηση, την επέκταση και τη βελτίωση της ανάλυσης δεδομένων EEG.

Οι προσεγγίσεις μηχανικής μάθησης εφαρμόζονται έντονα στο πρόβλημα της πρόβλεψης επιληπτικών κρίσεων λόγω της ικανότητάς τους να ταξινομούν καταστάσεις επιληπτικών κρίσεων από μεγάλο αριθμό δεδομένων και να παρέχουν χρήσιμα αποτελέσματα για τους νευρολόγους. Αυτά τα μοντέλα μηχανικής εκμάθησης περιλαμβάνουν την απόκτηση σήματος EEG, την προεπεξεργασία του σήματος, την εξαγωγή χαρακτηριστικών από τα σήματα και τέλος την ταξινόμηση μεταξύ των διαφορετικών καταστάσεων τής επιληπτικής κρίσης.

Η μηχανική μάθηση έχει φέρει επανάσταση στο πεδίο τής πρόβλεψης επιληπτικών κρίσεων προσφέροντας εργαλεία για την αντιμετώπιση τής υψηλής πολυπλοκότητας των σημάτων EEG και επιτρέποντας την εύκολη αξιολόγηση πολλαπλών παραλλαγών και χώρων χαρακτηριστικών υψηλότερης τάξης, ώστε να διακρίνονται τα κρυμμένα χαρακτηριστικά. Μέχρι τα τελευταία χρόνια, οι παραδοσιακές τεχνικές μηχανικής μάθησης (δηλαδή αλγόριθμοι μη βαθιάς μάθησης) ήταν η μόνη βιώσιμη επιλογή στην ανάλυση EEG και στην πραγματικότητα συνεχίζουν να χρησιμοποιούνται εκτενώς σε συνδυασμό με διάφορους αλγόριθμους εξαγωγής και επιλογής χαρακτηριστικών.

Σε μια σχετικά νεότερη τάση, οι αλγόριθμοι βαθιάς μάθησης (deep learning) έχουν βρει εφαρμογές στην ιατρική επεξεργασία εικόνας και σήματος, λόγω των εξελίξεων και της διαθεσιμότητας υπολογιστικής ισχύος και μεγάλων δεδομένων, δείχνοντας υψηλές δυνατότητες και σημαντικό αντίκτυπο καθώς, στις περισσότερες περιπτώσεις, η απόδοσή τους υπερβαίνει αυτές που είχαν προηγουμένως επιτευχθεί με παραδοσιακές τεχνικές μηχανικής μάθησης. Καθώς τα δεδομένα EEG αυξάνουν τον όγκο και την πολυπλοκότητα τους, οι αλγόριθμοι βαθιάς μάθησης αρχίζουν να αποδεικνύουν τις ικανότητές τους στο χειρισμό τής χαοτικής φύσης των σημάτων EEG και ανοίγουν νέες ευκαιρίες σε απαιτητικές βιοϊατρικές εφαρμογές, όπως είναι και η πρόληψη επιληπτικών κρίσεων.

3. ΔΕΔΟΜΕΝΑ

Στην ενότητα αυτή θα παρουσιάσουμε την πορεία αναζήτησης, επιλογής και λήψης τού συνόλου δεδομένων που χρησιμοποιήσαμε στη διπλωματική εργασία. Όπως προαναφέρθηκε, πρόκειται για το CHB-MIT Scalp EEG Database. Θα εξετάσουμε τους τρόπους λήψης τού συνόλου και θα δώσουμε αρκετές εναλλακτικές λύσεις. Τέλος, θα αιτιολογήσουμε πόσα από τα δεδομένα και τα κανάλια χρησιμοποιήσαμε και το λόγο για τον οποίο απορρίψαμε τα υπόλοιπα.

3.1 Αναζήτηση Συνόλου Δεδομένων

Από την εκτεταμένη βιβλιογραφία που χρησιμοποιήθηκε στα πλαίσια της εργασίας προέκυψε πληθώρα διαθέσιμων EEG συνόλων δεδομένων. Για διάφορους λόγους, τα περισσότερα εξ αυτών απορρίφθηκαν και έμεινε ουσιαστικά ως μόνο κατάλληλο το CHB-MIT. Τα κριτήρια που τέθηκαν για το σύνολο δεδομένων που αναζητούσαμε ήταν να περιέχει δεδομένα EEG από ανθρώπινο εγκέφαλο, μετρήσεις από πολλαπλά κανάλια ώστε να μπορούμε να εξάγουμε χαρακτηριστικά, αρκετά δεδομένα ώστε να είναι αξιόπιστες οι δοκιμές και να διατίθεται ελεύθερα. Παρακάτω ακολουθεί συνοπτική αναφορά στα σύνολα δεδομένων που εξετάστηκαν και στους λόγους που απορρίφθηκαν:

- *European Epilepsy Database*: Η Ευρωπαϊκή βάση επιληπτικών δεδομένων αναπτύχθηκε στα πλαίσια χρηματοδότησης της ΕΕ για το έργο EPILEPSIAE από τρία κέντρα επιληψίας στην Κοΐμπρα, το Παρίσι και το Φράιμπουργκ. Είναι η μεγαλύτερη σε τάξη μεγέθους διαθέσιμη συλλογή μακροχρόνιων μετρήσεων EEG και περιέχει τόσο ενδοκρανιακά όσο και επιφανειακά δεδομένα. Επί του παρόντος, η βάση δεδομένων της ΕΕ περιέχει σύνολα δεδομένων EEG από περισσότερους από 250 ασθενείς με επιληψία, 50 από αυτούς με ενδοκρανιακές εγγραφές με έως και 122 κανάλια. Κάθε σύνολο δεδομένων παρέχει δεδομένα EEG για συνεχή χρόνο εγγραφής περίπου 150 ωρών (>5 ημέρες) κατά μέσο όρο με ρυθμό δείγματος από 250Hz έως 2500Hz. Προφανώς απορρίφθηκε επειδή δεν είναι δωρεάν, αλλά και γιατί τα δεδομένα είναι εξαιρετικά ογκώδη για να τα διαχειριστούμε με τους περιορισμένους πόρους της παρούσας διπλωματικής εργασίας.
- *Freiburg EEG Database*: Μαζί με το CHB MIT, το σύνολο δεδομένων τού Freiburg είναι το δημοφιλέστερο στη βιβλιογραφία. Η βάση δεδομένων περιέχει ενδοκρανιακά EEG δεδομένα 21 ασθενών που πάσχουν από επιθετικές εστιακές επιληψίες. Τα EEG δεδομένα αποκτήθηκαν χρησιμοποιώντας το Neurofile NT σύστημα ψηφιακού βίντεο με 128 κανάλια, ρυθμό δειγματοληψίας 256Hz και μετατροπέα αναλογικού προς ψηφιακό 16bit. Πλέον, όμως, δεν είναι διαθέσιμο διότι έχει ενσωματωθεί στην Ευρωπαϊκή βάση επιληπτικών δεδομένων.
- *Bonn University*: Το συγκεκριμένο σύνολο δεδομένων δεν είναι πια διαθέσιμο από τη σελίδα του Πανεπιστημίου της Βόννης, καθώς και αυτό έχει ενταχθεί στην Ευρωπαϊκή βάση επιληπτικών δεδομένων. Παρόλα αυτά υπάρχει ακόμα δειγματοληπτημένο μέρος του στο UC Irvine Machine Learning Repository. Ο λόγος που δεν επελέγη το συγκεκριμένο είναι ότι έχει 5 διαφορετικές κλάσεις δεδομένων – αν και μόνο μία από αυτές περιέχει επιληπτικά δεδομένα, οπότε οι άλλες 4 μπορούν να συνενωθούν σε μία νέα κλάση με μη επιληπτικά δεδομένα – και κυρίως γιατί το τμήμα τού συνόλου δεδομένων που βρήκαμε είχε λίγα δεδομένα και για μικρό χρονικό διάστημα.
- *TUH (*Temple University Hospital*) EEG dataset*: Η συλλογή του TUH είναι η μεγαλύτερη διαθέσιμη βάση δεδομένων κλινικών δεδομένων EEG παγκοσμίως.

Περιλαμβάνει περισσότερα από 25.000 αρχεία EEG και περισσότερους από 14.000 ασθενείς. Τα πρωτογενή σήματα αποτελούνται από καταγραφές που κυμαίνονται μεταξύ 20 και 128 καναλιών δειγματοληπτημένα στα 250Hz τουλάχιστον χρησιμοποιώντας έναν μετατροπέα αναλογικού προς ψηφιακό 16-bit. Οι ηλικίες των ασθενών κυμαίνονται από 18 έως 90 ετών και το σύνολο των δεδομένων αγγίζει τα 1.8 TB. Αν και ελεύθερα διαθέσιμο, γίνεται εύκολα αντιληπτό ότι είναι ανέφικτο να το διαχειριστούμε με τους διαθέσιμους πόρους. Είναι δύσκολη ακόμα και η λήψη του και οι ίδιοι προτείνουν να τους αποσταλεί από τον ενδιαφερόμενο αποθηκευτικό μέσο ώστε να γίνει εκεί η εγγραφή.

- *American Epilepsy Society Seizure Prediction Challenge dataset*: Πρόκειται για ένα σύνολο δεδομένων από έναν διαγωνισμό του Kaggle το 2014. Αυτό το σύνολο δεδομένων έχει ενδοκρανιακά δεδομένα (iEEG) από πέντε σκύλους και δύο ανθρώπους με 48 επιληπτικές κρίσεις και πάνω από 627 ώρες μεσοκρισικής εγγραφής. Καταγράφηκαν δεδομένα iEEG από 16 εμφυτευμένα ηλεκτρόδια με ρυθμό δειγματοληψίας 400Hz. Ο βασικός ανασταλτικός παράγοντας είναι ότι υπάρχουν δεδομένα τόσο από ανθρώπους όσο και από σκύλους που δε βολεύει για την παραγωγή γενικευμένων ταξινομητών.

Από τα παραπάνω αντιλαμβανόμαστε ότι κανένα άλλο σύνολο δεδομένων από αυτά της βιβλιογραφίας πλην του CHB MIT δεν ικανοποιεί τις προϋποθέσεις που είχαν τεθεί εξαρχής. Συνεπώς, η επιλογή αποτέλεσε ουσιαστικά μονόδρομο.

3.2 Σύνολο Δεδομένων CHB MIT

Αυτή η βάση δεδομένων, που συλλέχθηκε στο Παιδικό Νοσοκομείο της Βοστώνης, αποτελείται από EEG εγγραφές από νεαρά άτομα με επιθετικές κρίσεις επιληψίας. Οι ασθενείς παρακολουθήθηκαν για αρκετές ημέρες μετά την απόσυρση τού φαρμάκου προκειμένου να χαρακτηριστούν οι κρίσεις τους και να αξιολογηθεί η υποψηφιότητά τους για χειρουργική επέμβαση.

Οι καταγραφές, που ομαδοποιήθηκαν σε 23 περιπτώσεις, συλλέχθηκαν από 22 άτομα (5 άνδρες ηλικίας 3–22 και 17 γυναίκες ηλικίας 1,5-19). Το δείγμα chb21 λήφθηκε 1,5 χρόνια μετά το δείγμα chb01 από την ίδια ασθενή. Ο Πίνακας 1 περιέχει το φύλο, την ηλικία και μερικά στοιχεία της EEG καταγραφής κάθε ασθενούς. Το δείγμα chb24 προστέθηκε σε αυτήν τη συλλογή τον Δεκέμβριο του 2010 και δε δίνονται για αυτό ηλικία και φύλο.

Κάθε περίπτωση περιέχει μεταξύ 9 και 42 συνεχόμενα αρχεία edf (βλ. Παράρτημα Ι) από ένα μόνο ασθενή. Οι περιορισμοί υλικού οδήγησαν σε κενά μεταξύ διαδοχικά αριθμημένων αρχείων edf, κατά τη διάρκεια των οποίων τα σήματα δεν καταγράφηκαν. Στις περισσότερες περιπτώσεις, τα κενά είναι της τάξης των 10 δευτερόλεπτων ή λιγότερο, αλλά περιστασιακά υπάρχουν πολύ μεγαλύτερα κενά. Προκειμένου να προστατευθεί το απόρρητο των ασθενών, όλες οι προστατευμένες πληροφορίες υγείας στα πρωτότυπα αρχεία edf έχουν αντικατασταθεί από υποκατάστατες πληροφορίες στα αρχεία που παρέχονται εδώ. Οι ημερομηνίες στα αρχικά αρχεία edf έχουν αντικατασταθεί από υποκατάστατες ημερομηνίες, αλλά οι χρονικές σχέσεις μεταξύ των μεμονωμένων αρχείων που ανήκουν σε κάθε περίπτωση έχουν διατηρηθεί. Στις περισσότερες περιπτώσεις, τα αρχεία edf περιέχουν ακριβώς μία ώρα ψηφιοποιημένων σημάτων EEG, αν και αυτά που ανήκουν στην περίπτωση chb10 έχουν διάρκεια δύο ωρών και αυτά που ανήκουν στις περιπτώσεις chb04, chb06, chb07, chb09 και chb23 έχουν διάρκεια τεσσάρων ωρών.

Πίνακας 1: Περιγραφή συνόλου δεδομένων CHB-MIT

Case	Gender	Age (years)	# of EEG channels	Avg Length of Seizures (s)	# seizures	Duration of Recordings (hh:mm:ss)
1	F	11	18	63.14	7	40:33:08
2	M	11	18	57.33	3	35:15:59
3	F	14	18	57.43	7	38:00:06
4	M	22	18	94.5	4	156:03:54
5	F	7	18	111	5	39:00:10
6	F	1.5	18	15.3	10	66:44:06
7	F	14.5	18	108.33	3	67:03:08
8	M	3.5	18	183.8	5	20:00:23
9	F	10	18	69	4	67:52:18
10	M	3	18	63.86	7	50:01:24
11	F	12	18	268.67	3	34:47:37
12	F	2	18	37	27	20:41:40
13	F	3	18	44.58	12	33:00:00
14	F	9	18	21	8	26:00:00
15	M	16	18	104.94	20	40:00:36
16	F	7	18	8.44	10	19:00:00
17	F	12	18	97.67	3	21:00:24
18	F	18	18	52.83	6	35:38:05
19	F	19	18	78.67	3	29:55:46
20	F	6	18	36.75	8	27:36:06
21	F	13	18	49.75	4	32:49:49
22	F	9	18	68	3	31:00:11
23	F	6	18	56.67	7	26:33:30
24	–	–	18	31.94	16	21:17:47
Total:					**185**	**979:56:07**

Όλα τα σήματα δειγματοληπτήθηκαν στα 256 δείγματα ανά δευτερόλεπτο με ανάλυση 16bit. Τα περισσότερα αρχεία περιέχουν 23 σήματα/κανάλια EEG (24 ή 26 σε μερικές περιπτώσεις), από αυτά, όμως, μόνο τα 18 υπάρχουν σε όλες τις περιπτώσεις. Το διεθνές σύστημα τοποθέτησης ηλεκτροδίων 10-20 EEG (Εικόνα 3) χρησιμοποιήθηκε για αυτές τις καταγραφές. Σε μερικές εγγραφές καταγράφονται και άλλα σήματα, όπως ένα σήμα ECG στα τελευταία 36 αρχεία που ανήκουν στην περίπτωση chb04 και ένα σήμα VNS στα 18 τελευταία αρχεία που ανήκουν στην περίπτωση chb09. Σε ορισμένες περιπτώσεις, έως και 5 "εικονικά" σήματα (που ονομάστηκαν "-") ήταν διάσπαρτα μεταξύ των σημάτων EEG για να αποκτήσουν μια ευανάγνωστη μορφή εμφάνισης. Αυτά τα εικονικά σήματα μπορούν να αγνοηθούν.

Το αρχείο RECORDS περιέχει μια λίστα με όλα τα 664 αρχεία edf που περιλαμβάνονται σε αυτήν τη συλλογή και το αρχείο RECORDS-WITH-SEIZURES παραθέτει τα 129 από αυτά τα αρχεία που περιέχουν μία ή περισσότερες επιληπτικές κρίσεις. Συνολικά, αυτά

τα αρχεία περιλαμβάνουν 198 κρίσεις (182 στο αρχικό σύνολο 23 ασθενών). Τα αρχεία seizures περιέχουν το δευτερόλεπτο της αρχής της κρίσης και το μήκος της σε δευτερόλεπτα. Επιπλέον, τα αρχεία με το όνομα chbnn-Summary.txt περιέχουν πληροφορίες σχετικά με το μοντάζ που χρησιμοποιείται για κάθε εγγραφή και τον χρόνο που έχει παρέλθει σε δευτερόλεπτα από την αρχή κάθε αρχείου edf έως την αρχή και το τέλος κάθε επιληπτικής κρίσης που περιέχεται σε αυτό.

3.3 Λήψη Συνόλου Δεδομένων

Το συνολικό μη συμπιεσμένο μέγεθος τού διαθέσιμου συνόλου δεδομένων είναι 42.6 GB. Η πλατφόρμα PhysioNet προτείνει την πρόσβαση στα αρχεία με τους ακόλουθους 4 δωρεάν τρόπους:

- Κατέβασμα συμπιεσμένου αρχείου από το διαθέσιμο σύνδεσμο πατώντας εδώ.
- Μέσω του Google Cloud Storage Browser πατώντας εδώ.
- Πάλι από το Google Cloud αλλά μέσω του εργαλείου gsutil

  gsutil -m cp -r gs://chbmit-1.0.0.physionet.org DESTINATION
- Μέσω terminal και της εντολής wget

  wget -r -N -c -np https://physionet.org/files/chbmit/1.0.0/

Προσωπική επιλογή για τη λήψη τού συνόλου δεδομένων ήταν η εντολή wget. Όπως έχουμε προαναφέρει στην ενότητα 1.2.5, λαμβάνοντας υπόψη μόνο τα αρχεία που περιέχουν επιληπτικές κρίσεις, το ποσοστό των μη επιληπτικών δειγμάτων ξεπερνά το 98%. Αντιλαμβάνεται κανείς ότι αν χρησιμοποιήσουμε και τα υπόλοιπα αρχεία, το μόνο που θα καταφέρουμε είναι να μεγαλώσουμε αυτό το ποσοστό και να χρειαστεί να παράξουμε ακόμα περισσότερα δεδομένα της επιληπτικής κλάσης έχοντας ως βάση ακριβώς τα ίδια δείγματα για την εφαρμογή των αλγορίθμων υπερδειγματοληψίας. Ουσιαστικά θα έχουμε πολλά παραπλήσια δείγματα που δε θα μας δίνουν περισσότερη πληροφορία.

Για τους παραπάνω λόγους θα χρησιμοποιήσουμε μόνο τα αρχεία που περιέχουν επιληπτικά δεδομένα. Για να κατεβάσουμε μόνο αυτά, αναπτύξαμε το αρχείο DownloadDataset.ipynb. Ο κώδικας παίρνει το url τού συνόλου δεδομένων και μέσω της εντολής wget κατεβάζει για τον πρώτο ασθενή το αρχείο MD5SUMS που περιέχει όλα τα ονόματα των αρχείων ανά ασθενή. Από αυτά κατεβάζει μόνο τα αρχεία edf που έχουν και το αντίστοιχο seizures αρχείο. Επαναλαμβάνει την ίδια διαδικασία για όλους τους ασθενείς. Έτσι, στο τέλος θα έχουμε ένα φάκελο για κάθε ασθενή που περιέχει μόνο τα απαραίτητα αρχεία.

3.4 Διαχείριση και Επιλογή Δεδομένων

Μετά τη λήψη τού συνόλου δεδομένων, επόμενο βήμα είναι η διαχείρισή του και η μετατροπή του σε μια μορφή που μπορεί να δοθεί στα επόμενα στάδια τού πειράματος που περιγράψαμε στην ενότητα 1.2. Επειδή θέλουμε να εφαρμόσουμε συνολικά τα πειράματα και όχι ανά ασθενή, θα χρησιμοποιήσουμε μόνο τα 18 κοινά κανάλια που εμφανίζονται σε όλες τις περιπτώσεις. Αυτά είναι τα FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2,

FZ-CZ και CZ-PZ. Για τον ίδιο λόγο, από τα πειράματα εξαιρούμε τον ασθενή chb12 του οποίου τα κανάλια που καταγράφονται αλλάζουν τρεις φορές ανάμεσα στα αρχεία του.

Τη διαχείριση του συνόλου δεδομένων αναλαμβάνει η συνάρτηση read_and_store_data του αρχείου EdfManipulation.ipynb. Τα αρχεία edf διαβάζονται μέσω της βιβλιοθήκης pyedflib και αποθηκεύονται σε έναν πίνακα numpy κρατώντας μόνο τα 18 κοινά κανάλια. Το διάβασμα των seizures αρχείων δεν υποστηρίζεται από την ίδια βιβλιοθήκη, οπότε καταφύγαμε σε μια λύση που προτάθηκε στο Mathworks , μετατρέποντας τον κώδικα από Matlab σε Python. Η συνάρτηση read_and_store_data διαβάζει για κάθε ασθενή ένα αρχείο edf και το αντίστοιχο seizures αρχείο και δημιουργεί έναν πίνακα με τα δεδομένα των 18 καναλιών. Σε αυτά προσθέτει μια ακόμα στήλη με όνομα ‘seizure’ που συμβολίζει με 0 ή 1 την απουσία ή την ύπαρξη επιληπτικής κρίσης αντίστοιχα στο συγκεκριμένο δείγμα. Επαναλαμβάνει την ίδια διαδικασία για όλα τα αρχεία όλων των ασθενών δημιουργώντας τελικά έναν numpy πίνακα με τα δείγματα όλων των ασθενών χωρίς να μας ενδιαφέρει η πληροφορία από ποιον ασθενή προέρχονται. Τέλος, μετατρέπουμε τον πίνακα σε pandas dataframe που είναι πιο βολικό στη διαχείρισή του στα επόμενα στάδια.

4. ΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΚΑΘΑΡΙΣΜΟΣ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Στο παρόν κεφάλαιο θα περιγραφεί η διαδικασία μετατροπής τού ακατέργαστου μεγάλου συνόλου δεδομένων CHB MIT σε ένα ισορροπημένο μικρότερο σύνολο χαρακτηριστικών. Θα αναλύσουμε τις μεθόδους εξαγωγής χαρακτηριστικών, τη z-score κανονικοποίηση που πραγματοποιήθηκε ανά χαρακτηριστικό, την επιλογή των κυριότερων χαρακτηριστικών μέσω PCA, τη δειγματοληψία δεδομένων από τη μη επιληπτική κλάση με 3 μεθόδους (Cluster Centroids, Near Miss, Random Undersampler), την κατασκευή συνθετικών δεδομένων μέσω ADASYN και SMOTE στην επιληπτική κλάση και τέλος το διαχωρισμό του συνόλου δεδομένων σε δεδομένα εκπαίδευσης και αξιολόγησης. Σκοπός είναι η κατασκευή ενός μικρότερου συνόλου δεδομένων με τις δύο κλάσεις να εκπροσωπούνται ισότιμα και χωρίς να χαθεί χρήσιμη πληροφορία. Συγκεκριμένα, σκοπός είναι η ορθότερη ανάδειξη της πληροφορίας που παίρνουμε από τους αισθητήρες που θα οδηγήσει τους αλγόριθμους ταξινόμησης σε καλύτερη απόδοση και ακριβέστερα αποτελέσματα.

4.1 Μέθοδοι Εξαγωγής Χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών προσφέρει μείωση διαστάσεων και πιο πολύπλοκους χώρους υψηλότερης τάξης, οι οποίοι μπορούν να αυξήσουν τη διακριτική ισχύ τού αλγορίθμου ταξινόμησης που χρησιμοποιείται για την πρόβλεψη. Για να εξετάσουμε το σήμα πολύπλευρα και να εκθέσουμε κρυμμένες πτυχές του που θα βοηθήσουν στην σωστή ταξινόμηση των επιμέρους δειγμάτων του υπολογίζουμε χαρακτηριστικά στο πεδίο του χρόνου και του φάσματος. Στο πείραμα με τα δύο μόνο κανάλια, αυτό του αριστερού ημισφαιρίου και εκείνο του δεξιού, εξάγονται και χαρακτηριστικά συσχέτισης των καναλιών. Στον Πίνακα 2 παρουσιάζονται συνοπτικά όλα τα χαρακτηριστικά που χρησιμοποιήθηκαν χωρισμένα στο πεδίο του χρόνου και της συχνότητας μαζί με το χαρακτηριστικό συσχέτισης καναλιών.

Η ανάλυση κινούμενου παραθύρου εκτελείται συνήθως για να χωρίσει τα ανεπεξέργαστα EEG δεδομένα σε τμήματα μικρότερης διάρκειας που στη συνέχεια χρησιμοποιούνται για εξαγωγή χαρακτηριστικών. Για τα πειράματά μας επιλέξαμε ένα μη επικαλυπτόμενο παράθυρο των 2 sec. Με το ρυθμό δειγματοληψίας τού σήματος να είναι στα 256 Hz, 512 δείγματα θα μας δώσουν ένα διάνυσμα χαρακτηριστικών. Στο config αρχείο η παράμετρος time_window ορίζει το μέγεθος τού παραθύρου σε δευτερόλεπτα και μπορεί ο χρήστης να πειραματιστεί. Ωστόσο, εμείς για τις μετρήσεις και τα πειράματά μας θα αρκεστούμε στο παράθυρο των 2 sec.

Πίνακας 2: Σύνολο εξαχθέντων χαρακτηριστικών από κάθε παράθυρο

Time Domain	Statistical moments	Mean, Variance, Skewness, Kurtosis, Median
	Standard deviation	Square root of variance
	Root Mean Square	Mean square signal root
	Zero crossings	Number of sign changes
	Peak-to-peak voltage	Difference between highest/lowest amplitude
	Entropy	Sample Entropy

Spectral Domain	Power Spectral Density	Energy at: • Delta (0.5-4 Hz) • Theta (4-8 Hz) • Alpha (8-12 Hz) • Beta (12-30 Hz) • Gamma (30-100 Hz)
Correlation	Maximum cross-correlation	Maximum dependence between a pair of EEG channels

4.1.1 Χαρακτηριστικά στο πεδίο του χρόνου

Τα χαρακτηριστικά που υπολογίστηκαν στο πεδίο του χρόνου είναι:

1. *Αριθμητική μέση τιμή (Mean)*

Η αριθμητική μέση τιμή ή μέσος όρος είναι ένα στατιστικό μέτρο θέσης που προκύπτει από το πηλίκο της διαίρεσης του αθροίσματος των τιμών μιας μεταβλητής δια του συνολικού πλήθους τους, δηλαδή του συνόλου των συχνοτήτων τους. Υπολογίζεται μέσω του τύπου:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \text{όπου} \begin{cases} \mu: \mu\acute{\epsilon}\sigma\eta\ \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma \\ x_i: \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma\ \tau\eta\ i-o\sigma\tau\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ n: \pi\lambda\acute{\eta}\theta o\varsigma\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho o\upsilon \end{cases}$$

Πρόκειται για ένα μέτρο για τον υπολογισμό της τιμής του οποίου αξιοποιούνται όλες οι τιμές του δείγματος. Παρόλα αυτά έχει την ιδιότητα να επηρεάζεται δυσανάλογα από τις πολύ μεγάλες ή τις πολύ μικρές παρατηρήσεις του δείγματος. Ο υπολογισμός της αριθμητικής μέσης τιμής έγινε μέσω της βιβλιοθήκης SciPy.

2. *Διακύμανση (Variance)*

Το τετράγωνο της τυπικής απόκλισης ονομάζεται διακύμανση και συμβολίζεται με σ^2. Η διακύμανση είναι η αναμενόμενη τιμή της τετραγωνικής απόκλισης της τυχαίας μεταβλητής από τη μέση τιμή και άτυπα μετρά πόσο μακριά ένα σύνολο αριθμών απλώνεται από τη μέση τιμή του. Υπολογίζεται μέσω του τύπου:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}, \text{ όπου} \begin{cases} \mu: \mu\acute{\epsilon}\sigma\eta\ \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma \\ x_i: \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma\ \tau\eta\ i-o\sigma\tau\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ n: \pi\lambda\acute{\eta}\theta o\varsigma\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho o\upsilon \end{cases}$$

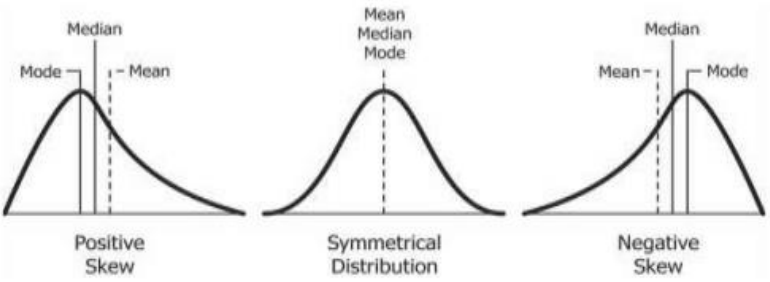
Ο υπολογισμός της διακύμανσης έγινε μέσω της βιβλιοθήκης NumPy.

3. *Ασυμμετρία (Skewness)*

Η ασυμμετρία δείχνει το βαθμό της συμμετρίας γύρω από τον μέσο των δεδομένων, ο οποίος έχει το μειονέκτημα να επηρεάζεται από τις ακραίες τιμές. Υπολογίζεται μέσω του τύπου:

$$G = \frac{\mu_3}{\sigma^3}\ , \text{ όπου } \mu_j = \frac{\sum_{i=1}^{n}(x_i-\mu)^j}{n} \text{ με} \begin{cases} \mu_3: \tau\rho\acute{\iota}\tau\eta\ \kappa\epsilon\nu\tau\rho\iota\kappa\acute{\eta}\ \rho o\pi\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma \\ x_i: \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma\ \tau\eta\ i-o\sigma\tau\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ \sigma: \tau\upsilon\pi\iota\kappa\acute{\eta}\ \alpha\pi\acute{o}\kappa\lambda\epsilon\iota\sigma\eta\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma \end{cases}$$

Όταν μια καμπύλη συχνοτήτων είναι συμμετρική ως προς τον κατακόρυφο άξονα που διέρχεται από την κορυφή της κατανομής, όπως η μεσαία στο Σχήμα 10, τότε η κατανομή είναι συμμετρική. Όταν η καμπύλη συχνοτήτων δεν είναι συμμετρική, δηλαδή, όταν δεξιά και αριστερά του κατακόρυφου άξονα που περνάει από την κορυφή δε βρίσκεται το ίδιο ποσοστό παρατηρήσεων, τότε η κατανομή είναι ασύμμετρη. Υπάρχουν δύο είδη ασυμμετρίας: Θετική ασυμμετρία και αρνητική ασυμμετρία. Μια καμπύλη συχνοτήτων παρουσιάζει θετική ασυμμετρία όταν οι περισσότερες παρατηρήσεις βρίσκονται δεξιά της κορυφής, ενώ παρουσιάζει αρνητική ασυμμετρία όταν οι περισσότερες παρατηρήσεις βρίσκονται αριστερά της κορυφής.



Σχήμα 10: Παραδείγματα συμμετρίας, θετικής ασυμμετρίας και αρνητικής ασυμμετρίας

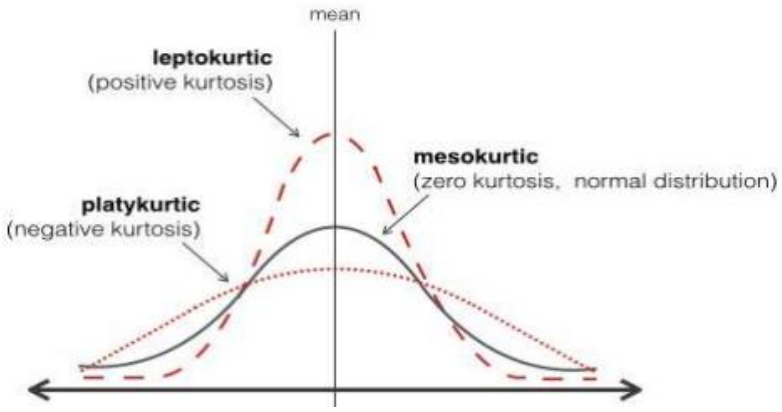
Ο υπολογισμός της ασυμμετρίας έγινε μέσω της βιβλιοθήκης SciPy.

4. *<u>Κύρτωση (Kurtosis)</u>*

Η κύρτωση είναι ένα μέτρο της κατανομής του σήματος γύρω από τη μέση τιμή. Δείχνει την αιχμηρότητα ή την πλάτυνση της κατανομής. Η κανονική κατανομή είναι το μέτρο και για την κυρτότητα μίας κατανομής με δεδομένο ότι πρόκειται για μια μεσόκυρτη κατανομή, αφού οι τιμές της ισοκατανέμονται αριστερά και δεξιά του αριθμητικού μέσου. Αν κάποια κατανομή έχει περισσότερο “οξεία” κορυφή από αυτή της κανονικής κατανομής τότε ονομάζεται λεπτόκυρτη και υποδεικνύει μεγάλη συγκέντρωση τιμών γύρω από τη μέση τιμή. Όταν μια κατανομή έχει περισσότερο “πλατιά” κορυφή τότε ονομάζεται πλατύκυρτη και αυτό δείχνει ότι οι τιμές διασπείρονται πάρα πολύ αριστερά και δεξιά της μέσης τιμής. Ένα παράδειγμα για το πως μοιάζουν αυτοί οι τρεις τύποι κύρτωσης της κατανομής ακολουθεί στο Σχήμα 11. Η ασυμμετρία υπολογίζεται μέσω του τύπου:

$$K = \frac{\mu_4}{\sigma^4}\text{, όπου } \mu_j = \frac{\sum_{i=1}^{n}(x_i-\mu)^j}{n} \text{ με } \begin{cases} \mu_4 \colon \tau\acute{\varepsilon}\tau\alpha\rho\tau\eta\ \kappa\varepsilon\nu\tau\rho\iota\kappa\acute{\eta}\ \rho o\pi\acute{\eta}\ \delta\varepsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma \\ x \colon \tau\iota\mu\acute{\eta}\ \delta\varepsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma\ \tau\eta\ \sigma\upsilon\gamma\kappa\varepsilon\kappa\rho\iota\mu\acute{\varepsilon}\nu\eta\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ \sigma \colon \tau\upsilon\pi\iota\kappa\acute{\eta}\ \alpha\pi\acute{o}\kappa\lambda\varepsilon\iota\sigma\eta\ \delta\varepsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma \end{cases}$$

Ο υπολογισμός της ασυμμετρίας έγινε μέσω της βιβλιοθήκης SciPy.



Σχήμα 11: Παραδείγματα διαφορετικών κυρτώσεων κατανομών

5. *Τυπική Απόκλιση (Standard Deviation)*

Το σημαντικότερο στατιστικό μέτρο διασποράς των τιμών μιας μεταβλητής X γύρω από τον αριθμητικό τους μέσο είναι η τυπική απόκλιση. Η τυπική απόκλιση απαντά στο ερώτημα: «πόσο μακριά από τη μέση τιμή τους βρίσκονται οι παρατηρήσεις;». Έτσι, όταν οι παρατηρήσεις δε διαφέρουν πολύ από τη μέση τιμή τους, η τυπική απόκλιση είναι μικρή, ενώ αντίθετα, η τυπική απόκλιση μεγαλώνει, όσο περισσότερο «διασκορπίζονται» οι παρατηρήσεις γύρω από τη μέση τιμή τους. Δηλαδή, η τυπική απόκλιση μάς δίνει ένα μέτρο της μέσης απόστασης-απόκλισης των παρατηρήσεων από τη μέση τιμή τους. Υπολογίζεται με την τετραγωνική ρίζα του μέσου αριθμητικού των τετραγώνων των αποκλίσεων των τιμών μιας μεταβλητής X από τον αριθμητικό μέσο τους. Η τυπική απόκλιση συμβολίζεται με το σ και υπολογίζεται μέσω του τύπου:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}, \text{ όπου } \begin{cases} \mu: \mu\acute{\epsilon}\sigma\eta\ \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau\sigma\varsigma \\ x_i: \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau\sigma\varsigma\ \tau\eta\ i-\sigma\sigma\tau\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ n: \pi\lambda\acute{\eta}\theta\sigma\varsigma\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho\sigma\upsilon \end{cases}$$

Ο υπολογισμός της τυπικής απόκλισης έγινε μέσω της βιβλιοθήκης NumPy.

6. *Διάμεσος (Median)*

Η διάμεσος είναι ένα ακόμα μέτρο θέσης και εκφράζει την κεντρική θέση τής κατανομής των παρατηρήσεων. Είναι η τιμή για την οποία ισχύει ότι το 50% των παρατηρήσεων είναι μικρότερες από αυτή και το υπόλοιπο 50% των παρατηρήσεων είναι μεγαλύτερες από αυτή. Η τιμή της διαμέσου είναι η κεντρική τιμή του ταξινομημένου δείγματος αν έχει περιττό πλήθος τιμών ή το ημιάθροισμα των δύο κεντρικών τιμών του αν έχει άρτιο πλήθος τιμών και υπολογίζεται από τον ακόλουθο τύπο:

$$\delta = \begin{cases} \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)\Big/2, n\ \acute{\alpha}\rho\tau\iota\sigma\varsigma \\ x_{\left\lceil\frac{n}{2}\right\rceil}, n\ \pi\epsilon\rho\iota\tau\tau\acute{\sigma}\varsigma \end{cases}, \text{ όπου } \begin{cases} \delta: \delta\iota\acute{\alpha}\mu\epsilon\sigma\sigma\varsigma\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau\sigma\varsigma \\ x_i: \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau\sigma\varsigma\ \tau\eta\ i-\sigma\sigma\tau\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ n: \pi\lambda\acute{\eta}\theta\sigma\varsigma\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho\sigma\upsilon \end{cases}$$

Η διάμεσος δεν επηρεάζεται ιδιαιτέρως από ακραίες τιμές. Έτσι, παρόλο που δεν αξιοποιεί όλες τις τιμές τού δείγματος, για την περιγραφή παρατηρήσεων που εμφανίζουν ακραίες τιμές προτιμάται από τη μέση τιμή, η οποία επηρεάζεται πολύ από ακραίες τιμές. Ο υπολογισμός της διαμέσου έγινε μέσω της βιβλιοθήκης NumPy.

7. *Ρυθμός Διάσχισης Μηδενικού Άξονα (Zero Crossing Rate)*

Μία διάσχιση του μηδενικού άξονα ενός δειγματοληπτημένου σήματος προκύπτει κάθε φορά που δύο διαδοχικές τιμές έχουν διαφορετικό πρόσημο. Το ZCR υπολογίζεται ως ο αριθμός των χρονικών διασχίσεων του μηδενικού άξονα διαιρεμένος με τον συνολικό αριθμό των δειγμάτων στο παράθυρο που εξετάζουμε. Ο τύπος υπολογισμού είναι:

$$ZC = \frac{1}{n}\sum_{i=2}^{n}\frac{|sgn(x_i) - sgn(x_{i-1})|}{2}$$

όπου $\begin{cases} ZC: \pi\lambda\acute{\eta}\theta o\varsigma\ \delta\iota\alpha\sigma\chi\acute{\iota}\sigma\epsilon\omega\nu\ \mu\eta\delta\epsilon\nu\iota\kappa o\acute{\upsilon}\ \acute{\alpha}\xi o\nu\alpha \\ x_i: \tau\iota\mu\acute{\eta}\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\tau o\varsigma\ \tau\eta\ i-o\sigma\tau\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta} \\ n: \pi\lambda\acute{\eta}\theta o\varsigma\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho o\upsilon \end{cases}$ και $sgn(x_i) = \begin{cases} 1, & x_i \geq 0 \\ -1, & x_i < 0 \end{cases}$

Η έντονη δραστηριότητα τού σήματος που φανερώνεται και μέσω του αυξημένου ρυθμού διάσχισης του μηδενικού άξονα είναι ένα πιθανό δείγμα μιας υποκείμενης επιληπτικής κρίσης.

8. *Ενεργός τιμή σήματος (Root Mean Square)*

Η ενεργός τιμή του σήματος ορίζεται ως η τετραγωνική ρίζα της μέσης ισχύος. Για διακριτό σήμα υπολογίζεται ως η τετραγωνική ρίζα τού μέσου όρου της τετραγωνικής τιμής του σήματος:

$$RMS = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1} x_i^2}$$

9. *Εύρος μεταβολής (Peak to Peak)*

Το απλούστερο μέτρο διασποράς είναι το εύρος μεταβολής ενός σήματος. Ορίζεται ως η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής και υπολογίζεται ως:

$$E = x_{max} - x_{min},\ \text{όπου}\ \begin{cases} x_{min}: \epsilon\lambda\acute{\alpha}\chi\iota\sigma\tau o\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho o\upsilon \\ x_{max}: \mu\acute{\epsilon}\gamma\iota\sigma\tau o\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\ \pi\alpha\rho\alpha\theta\acute{\upsilon}\rho o\upsilon \end{cases}$$

Το εύρος μεταβολής είναι αρκετά απλό, αλλά δε θεωρείται αξιόπιστο γιατί εξαρτάται μόνο από τις δύο ακραίες τιμές των δεδομένων. Αν η διαφορά των ακραίων τιμών είναι πολύ μεγάλη, τότε και το εύρος θα είναι ανάλογο. Παρόλα αυτά είναι μια ένδειξη έντονης μεταβολής.

10. *Εντροπία Δείγματος (Sample Entropy)*

Το 2000 οι Richman και Moorman ανέπτυξαν μια τροποποιημένη μορφή της προσεγγιστικής εντροπίας (Approximate Entropy), η οποία είχε ως στόχο να περιορίσει μεγάλο κομμάτι των μειονεκτημάτων που παρουσίασε. Η νέα τεχνική ονομάστηκε Εντροπία Δείγματος (Sample Entropy) και οι βασικές διαφορές της σε σχέση με την Προσεγγιστική Εντροπία είναι το γεγονός ότι κατά τον υπολογισμό της ποσότητας των όμοιων διανυσμάτων, δεν περιλαμβάνει την σύγκριση του κάθε

διανύσματος με τον εαυτό του. Όπως είναι κατανοητό, αυτή η προσέγγιση περιορίζει σε σημαντικό βαθμό την προκατάληψη στο τελικό αποτέλεσμα, ενώ παράλληλα η διαδικασία ολοκληρώνεται ταχύτερα μιας και περιορίζονται οι περιττές συγκρίσεις.

Η Εντροπία Δείγματος χρησιμοποιείται για την αξιολόγηση της πολυπλοκότητας των φυσιολογικών χρονοσειρών για τη διάγνωση ασθενειών. Βασίζεται σε μια έννοια παρόμοια με την προσεγγιστική εντροπία, όπου η εντροπία δείγματος συγκρίνει τον συνολικό αριθμό διανυσμάτων παραθύρων μεγέθους m και m+1, αλλά διαφέρει από την προσεγγιστική στο ότι η ομοιότητα όλων των ζευγών διανυσμάτων παραθύρων $u[i]$ και $u[j]$ με ανοχή r υπολογίζονται μέσω του τύπου:

$$\varphi^m(r) = \sum_{j=0, j\neq i}^{N-m} \sum_{i=0}^{N-m} \Theta(r - \|u[i] - u[j]\|_\infty)$$

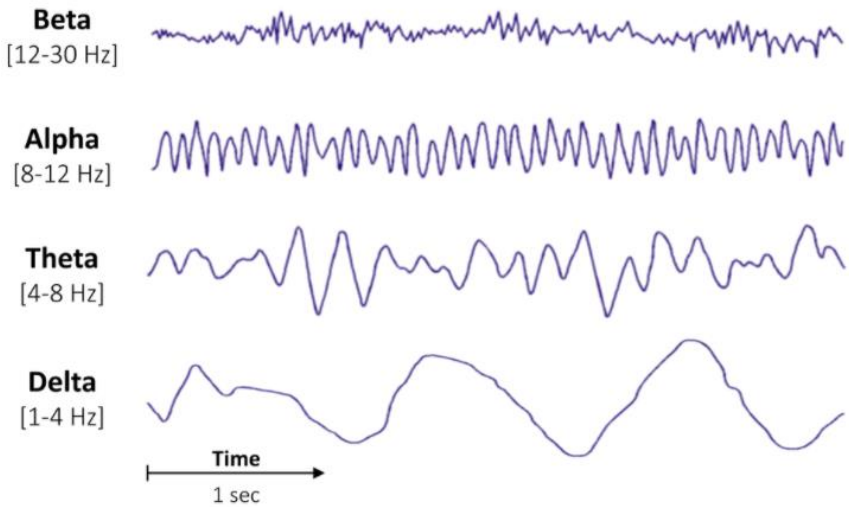
όπου $\Theta(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$ η μοναδιαία συνάρτηση βηματισμού (Heaviside), το σήμα παραθύρου μεγέθους m ορίζεται ως $\mathrm{u[i]} = \left[x[i]\, x[i+1]\, \cdots x[i+m-1]\right]^T$, N το μέγεθος της χρονοσειράς x, m το μέγεθος του παραθύρου και r είναι η ανοχή (tolerance). Τότε η εντροπία δείγματος υπολογίζεται ως εξής:

$$SampEn(x, m, r) = \log \varphi^m(r) - \log \varphi^{m+1}(r)$$

Η εντροπία δείγματος έχει δύο πλεονεκτήματα: ανεξαρτησία μήκους δεδομένων και μια σχετικά απλή εφαρμογή. Ο υπολογισμός της εντροπίας δείγματος έγινε μέσω της βιβλιοθήκης pyEntrp.

4.1.2 Φασματικά χαρακτηριστικά

Μία από τις πιο ευρέως χρησιμοποιούμενες μεθόδους για την ανάλυση δεδομένων EEG είναι η αποσύνθεση του σήματος σε λειτουργικά διακριτές ζώνες συχνοτήτων, όπως οι δέλτα (0.5-4 Hz), θήτα (4-8 Hz), άλφα (8-12 Hz), βήτα (12–30 Hz) και γάμμα (30–100 Hz).



Σχήμα 12: Ζώνες συχνοτήτων σήματος

Ας εμβαθύνουμε λίγο σε αυτές τις ζώνες συχνοτήτων:

- *Δέλτα (0.5-4 Hz)*: Τα πιο αργά και υψηλότερου πλάτους εγκεφαλικά κύματα. Οι δέλτα συχνότητες είναι ισχυρότερες στο δεξί ημισφαίριο τού εγκεφάλου και οι πηγές τους συνήθως εντοπίζονται στο θάλαμο.

- *Θήτα (4-8 Hz)*: Τα κύματα της θήτα ζώνης μπορούν να καταγραφούν από όλο τον φλοιό, υποδεικνύοντας ότι δημιουργούνται από ένα ευρύ δίκτυο που περιλαμβάνει μεσαίες προμετωπιαίες περιοχές, κεντρικούς, βρεγματικούς και μεσαίους χρονικούς φλοιούς. Τα θήτα εγκεφαλικά κύματα συνδέονται γενικά με εγκεφαλικές διεργασίες που βασίζονται στο διανοητικό φόρτο εργασίας ή στη μνήμη εργασίας.
- *Άλφα (8-12 Hz)*: Τα κύματα της άλφα ζώνης ορίζονται ως ρυθμική ταλάντωση εντός του εύρους συχνοτήτων 8-12 Hz. Έχουν αρκετές λειτουργικές συσχετίσεις που αντανακλούν αισθητήριες λειτουργίες, κινητικές λειτουργίες και λειτουργίες μνήμης.
- *Βήτα (12-30 Hz)*: Οι ταλαντώσεις εντός τού εύρους ζώνης 12-40 Hz αναφέρονται συνήθως ως δραστηριότητα ζώνης βήτα. Αυτή η συχνότητα δημιουργείται τόσο στις οπίσθιες όσο και στις μετωπικές περιοχές. Η ενεργή, απασχολημένη ή ανήσυχη σκέψη και η ενεργή συγκέντρωση είναι γενικά γνωστό ότι συσχετίζονται με υψηλότερη ισχύ βήτα.
- *Γάμμα (30-100 Hz)*: Προς το παρόν, οι συχνότητες γάμμα είναι οι μαύρες τρύπες της έρευνας EEG, καθώς δεν είναι ακόμη σαφές πού ακριβώς δημιουργούνται στον εγκέφαλο και τι αντικατοπτρίζουν αυτές οι ταλαντώσεις.

Ο διαχωρισμός σε ζώνες συχνοτήτων συνεπάγεται την αποσύνθεση του EEG σήματος σε συνιστώσες συχνότητας, η οποία επιτυγχάνεται συνήθως μέσω μετασχηματισμών Fourier. Ο σχεδόν πάντα χρησιμοποιούμενος αλγόριθμος για τον υπολογισμό του είναι ο FFT (Fast Fourier Transform), ο οποίος επιστρέφει, για κάθε κάδο συχνοτήτων, ένα μιγαδικό αριθμό από τον οποίο μπορεί εύκολα να εξαχθεί το πλάτος και η φάση του σήματος στη συγκεκριμένη συχνότητα. Στη φασματική ανάλυση είναι συνηθισμένο να λαμβάνουμε το τετράγωνο του FFT για να κάνουμε μια εκτίμηση της φασματικής πυκνότητας ισχύος (ή φάσματος ισχύος ή περιοδογράμματος).

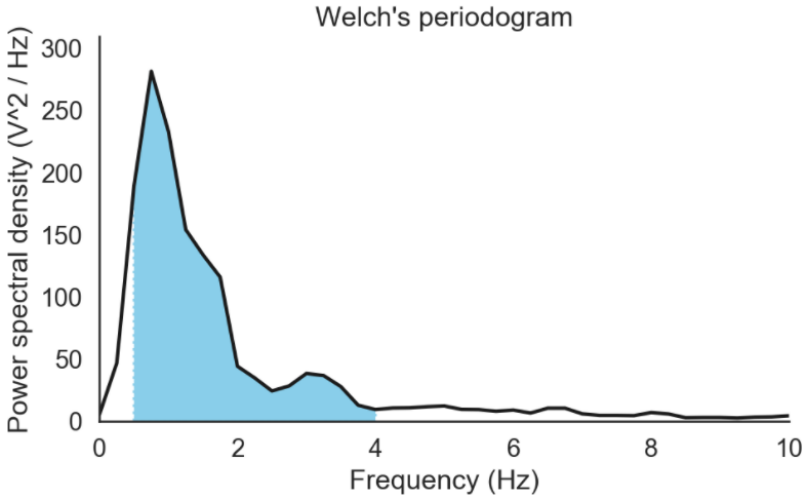
Το PSD είναι ένας κατάλληλος υποψήφιος για επεξεργασία EEG σήματος λόγω του γεγονότος ότι κατανέμει την ισχύ σήματος πάνω στη συχνότητα και εκφράζει την ισχύ των διακυμάνσεων ενέργειας ως συνάρτηση της συχνότητας. Με άλλα λόγια, δείχνει σε ποιες συχνότητες οι διακυμάνσεις είναι ισχυρές και σε ποιες είναι αδύναμες. Το PSD είναι ο μέσος όρος του τετραγώνου τού μετασχηματισμού Fourier, σε μεγάλο χρονικό διάστημα:

$$S_x = \lim_{T\to\infty} E\left\{\frac{1}{2T}\left|\int_{-T}^{T} x(t)e^{-j2\pi ft}dt\right|\right\}$$

Αν και μπορούν να πραγματοποιηθούν πολλές αναλύσεις από την πυκνότητα φασματικής ισχύος, θα επικεντρωθούμε εδώ σε μια πολύ απλή: τη μέση ισχύ ζώνης, η οποία συνίσταται στον υπολογισμό ενός μόνο αριθμού που συνοψίζει τη συμβολή τής δεδομένης ζώνης συχνοτήτων στη συνολική ισχύ τού σήματος. Αυτό είναι ιδιαίτερα χρήσιμο σε μια προσέγγιση μηχανικής μάθησης, όταν συχνά απαιτείται να εξαχθούν ορισμένα χαρακτηριστικά από τα δεδομένα και ενίοτε να υπολογιστεί ένας μόνο αριθμός που θα συνοψίζει μια συγκεκριμένη πτυχή των δεδομένων.

Για να υπολογίσουμε τη μέση ισχύ των προαναφερθεισών ζωνών, πρέπει πρώτα να υπολογίσουμε μια εκτίμηση της φασματικής πυκνότητας ισχύος. Η μέθοδος που χρησιμοποιείται συνήθως είναι η μέθοδος Welch. Η μέθοδος του Welch υπολογίζει μια εκτίμηση του PSD διαιρώντας τα δεδομένα σε επικαλυπτόμενα τμήματα, υπολογίζοντας ένα τροποποιημένο περιοδόγραμμα για κάθε τμήμα και στη συνέχεια το μέσο όρο των περιοδογραμμάτων.

Η μέθοδος του Welch βελτιώνει την ακρίβεια του κλασικού περιοδογράμματος. Ο λόγος είναι απλός: τα δεδομένα EEG ποικίλλουν πάντα στο χρόνο, πράγμα που σημαίνει ότι εάν κοιτάξουμε 30 δευτερόλεπτα δεδομένων EEG, είναι απίθανο το σήμα να μοιάζει με ένα τέλειο άθροισμα καθαρών ημιτονοειδών. Αντίθετα, το φασματικό περιεχόμενο του EEG με την πάροδο του χρόνου συνεχώς τροποποιείται από τη νευρωνική δραστηριότητα. Το πρόβλημα είναι ότι, για να επιστρέψουμε μια πραγματική φασματική εκτίμηση, ένα κλασικό περιοδόγραμμα απαιτεί το φασματικό περιεχόμενο του σήματος να είναι σταθερό κατά τη διάρκεια της εξεταζόμενης χρονικής περιόδου. Επειδή αυτό δε συμβαίνει ποτέ, το περιοδόγραμμα είναι γενικά προκατειλημμένο και περιέχει υπερβολικά μεγάλη διακύμανση. Μέσω του μέσου όρου των περιοδογραμμάτων που λαμβάνονται σε μικρά τμήματα των παραθύρων, η μέθοδος του Welch επιτρέπει τη δραστική μείωση αυτής της διακύμανσης. Αυτό, ωστόσο, οδηγεί στο κόστος μιας ανάλυσης χαμηλότερης συχνότητας.



Σχήμα 13: Παράδειγμα υπολογισμού μέσης ισχύος στη ζώνη δέλτα

Ορίζοντας τις επιθυμητές ζώνες και έχοντας εφαρμόσει τη μέθοδο Welch για να υπολογίσουμε μια εκτίμηση του PSD, αρκεί να υπολογίσουμε το εμβαδόν του περιοδογράμματος μεταξύ των ακραίων συχνοτήτων κάθε μπάντας για να προσδιορίσουμε τη μέση ισχύ ανά ζώνη. Το Σχήμα 13 είναι ένα παράδειγμα υπολογισμού της μέσης ισχύος στη ζώνη δέλτα. Στο σχήμα η ζώνη δέλτα έχει επισημανθεί με γαλάζιο χρώμα και εκτείνεται από 0,5 Hz έως 4 Hz. Δεδομένου ότι δεν υπάρχει έκφραση κλειστής μορφής για να υπολογιστεί το ολοκλήρωμα αυτής της περιοχής, πρέπει να το προσεγγίσουμε. Αυτό επιτυγχάνεται συνήθως χρησιμοποιώντας τον σύνθετο κανόνα του Simpson. Η ιδέα πίσω από αυτήν είναι στην πραγματικότητα πολύ απλή: αποσυνθέτουμε αυτήν την περιοχή σε αρκετές παραβολές και στη συνέχεια αθροίζουμε την περιοχή αυτών των παραβολών.

Για τους υπολογισμούς χρησιμοποιήθηκε η βιβλιοθήκη SciPy. Συγκεκριμένα, για τον υπολογισμό της μεθόδου Welch χρησιμοποιήθηκε η υποενότητα signal και η συνάρτηση welch, ενώ για τον υπολογισμό ολοκληρώματος με τον σύνθετο κανόνα του Simpson χρησιμοποιήθηκε η υποενότητα integrate και η συνάρτηση simps.

4.1.3 Διμερή Χαρακτηριστικά

Το πιο συνηθισμένο διμερές χαρακτηριστικό είναι η αλληλοσυσχέτιση που εκτιμά την εξάρτηση μεταξύ ενός ζεύγους καναλιών EEG, λαμβάνοντας υπόψη τις καθυστερήσεις

χρόνου μετατοπίζοντας ένα από τα δύο σήματα. Η μέγιστη αλληλοσυσχέτιση διατηρείται ως το πιο ενημερωτικό χαρακτηριστικό.

Η απόλυτη τιμή της μέγιστης τιμής της αλληλοσυσχέτισης χρησιμοποιείται ως μέτρο συνδεσιμότητας και υπολογίζεται για κάθε ζεύγος καναλιών EEG για να ποσοτικοποιηθεί η ομοιότητα μεταξύ οποιωνδήποτε δύο σημάτων EEG. Το πλεονέκτημά της έγκειται στο γεγονός ότι λαμβάνει υπόψη τις πιθανές καθυστερήσεις χρόνου που έχουν εισαχθεί μεταξύ δύο χωρικά απομακρυσμένων σημάτων. Η μέγιστη αλληλοσυσχέτιση C_{xy} μεταξύ δύο σημάτων $x(t)$, $y(t)$ δίνεται ως

$$C_{xy} = max_{\tau}(|c_{xy}(\tau)|), \tau \in [-w, w] \; με \; w: μέγεθος \; παραθύρου \; (sec) \; με$$

$$c_{xy}(\tau) = \sum_{n} x[n+\tau] \cdot \overline{y[n]}$$

Υπενθυμίζουμε ότι το συγκεκριμένο χαρακτηριστικό εξήχθη μόνο στο ένα από τα τρία πειράματα για τα οποία θα μιλήσουμε στο Κεφάλαιο 7. Το πείραμα αυτό ήταν με δύο μόνο κανάλια. Στο πείραμα με το ένα αντιπροσωπευτικό κανάλι μέσων τιμών δεν υπήρχε δεύτερο κανάλι για να υπολογίσουμε αλληλοσυσχέτιση. Στο άλλο πείραμα που χρησιμοποιούμε και τα 18 διαθέσιμα κανάλια, θα χρειαζόταν να υπολογίσουμε αυτό το χαρακτηριστικό $\binom{18}{2} = \frac{18!}{16!\,2!} = 153$ φορές που σε συνδυασμό με τα υπόλοιπα $\left(10\frac{time\ features}{channel} + 5\frac{spectral\ features}{channel}\right) \cdot 18\ channels = 270$ χαρακτηριστικά, θα δημιουργούσε ένα δύσκολα διαχειρίσιμο σύνολο 423 χαρακτηριστικών. Ο υπολογισμός των αλληλοσυσχετίσεων μεταξύ των καναλιών έγινε με χρήση της μεθόδου correlate της βιβλιοθήκης NumPy, θέτοντας την παράμετρο mode σε full και κρατώντας τη μέγιστη τιμή του διανύσματος που επιστρέφει.

4.2 Κανονικοποίηση

Στόχος της κανονικοποίησης είναι η αλλαγή στις τιμές των αριθμητικών στηλών τού συνόλου δεδομένων σε μια κοινή κλίμακα, χωρίς να παραμορφωθούν οι διαφορές στο εύρος τιμών. Η κανονικοποίηση των δεδομένων γίνεται ώστε να αντιμετωπιστούν δυσκολίες ορισμένων μεθόδων εξόρυξης καθώς είναι μια κοινή απαίτηση για πολλούς αλγορίθμους μηχανικής μάθησης και σε νευρωνικά δίκτυα. Αυτό συμβαίνει γιατί ενδέχεται να συμπεριφέρονται άσχημα εάν τα μεμονωμένα χαρακτηριστικά έχουν διαφορετικά εύρη τιμών και δε μοιάζουν περισσότερο ή λιγότερο με τυπικά κανονικά κατανεμημένα δεδομένα (π.χ. Gaussian με 0 μέση τιμή και διακύμανση μονάδας).

Για να πετύχουμε τα παραπάνω, εφαρμόζουμε στο dataframe των χαρακτηριστικών τον StandardScaler και τη μέθοδο fit_transform τής υποενότητας Preprocessing της βιβλιοθήκης scikit-learn. Ο Standard Scaler κανονικοποιεί τα χαρακτηριστικά αφαιρώντας το μέσο όρο και κλιμακώνοντας στη μοναδιαία διακύμανση μέσω του τύπου $z = (x - u) \,/\, s$, όπου u η μέση τιμή του δείγματος, s η τυπική απόκλιση και z η νέα τιμή του δείγματος. Αυτή είναι η λεγόμενη z-score κανονικοποίηση. Η fit_transform εφαρμόζει την παραπάνω διαδικασία στα χαρακτηριστικά.

4.3 Μείωση Διαστάσεων με PCA

Όπως προαναφέρθηκε και στην ενότητα 1.2.4 και παρατηρήσαμε και στην ενότητα 4.1, τα χαρακτηριστικά που εξήχθησαν είναι πολυάριθμα και δημιουργούν ένα χώρο πολύ μεγάλης διάστασης. Σκοπός μας είναι να μειώσουμε τη διάσταση τού συνόλου χαρακτηριστικών ώστε να είναι πιο εύκολα διαχειρίσιμο και η επεξεργασία του να απαιτεί

λιγότερο χρόνο. Παράλληλα, όμως, μας ενδιαφέρει να μη χαθεί η πληροφορία που περιέχεται μέσα σε αυτό. Η πιο δημοφιλής μέθοδος που ικανοποιεί τις δύο απαιτήσεις μας είναι η PCA.

Η μέθοδος PCA (Ανάλυση Κύριων Συνιστωσών) αποτελεί μία γραμμική μέθοδο συμπίεσης δεδομένων, η οποία συνίσταται από τον επαναπροσδιορισμό των συντεταγμένων ενός συνόλου δεδομένων σε ένα άλλο σύστημα συντεταγμένων, το οποίο θα είναι καταλληλότερο στην επικείμενη ανάλυση δεδομένων. Αυτές οι νέες συντεταγμένες είναι το αποτέλεσμα ενός γραμμικού συνδυασμού προερχόμενου από τις αρχικές μεταβλητές και εκπροσωπούνται σε ορθογώνιο άξονα, ενώ τα επικείμενα σημεία διατηρούν μια φθίνουσα σειρά όσο αφορά στη τιμή της διακύμανσής τους. Για το λόγο αυτό, το πρώτο κύριο συστατικό (principal component) διατηρεί περισσότερες πληροφορίες δεδομένων σε σύγκριση με το δεύτερο και τα επόμενα, τα οποία δεν διατηρούν πληροφορίες οι οποίες έχουν εισέλθει νωρίτερα.

Οι διευθύνσεις στις οποίες πρέπει να γίνει η προβολή των δεδομένων στην PCA για να έχουμε μέγιστη διακύμανση, είναι αυτές που ορίζονται από τα ιδιοδιανύσματα τού πίνακα συνδιακύμανσης των δεδομένων, καθώς με τον τρόπο αυτό οι μεταβλητές τού νέου συστήματος συντεταγμένων είναι ασυσχέτιστες. Κάθε αντίστοιχη ιδιοτιμή είναι η διακύμανση σε κάθε νέα διεύθυνση προβολής (η διακύμανση κάθε νέας μεταβλητής).

Η PCA είναι ένας μετασχηματισμός δεδομένων (των οποίων οι μεταβλητές είναι ενδεχομένως συσχετισμένες), με τον οποίο οι νέες μεταβλητές είναι ορθογώνιες μεταξύ τους, αποκτούν μέγιστη διακύμανση και είναι ασυσχέτιστες. Το αποτέλεσμα της PCA είναι τριπλό:

1. *Μεγιστοποίηση διακύμανσης*: Το εύρος των τιμών των δεδομένων αυξάνεται σε έναν από τους νέους άξονες (και ελαττώνεται στους υπόλοιπους). Ο μετασχηματισμός της PCA γίνεται με προβολή των σημείων σε μια διεύθυνση στην οποία η διακύμανσή τους μεγιστοποιείται (μέγιστο εύρος). Ταυτόχρονα, το άθροισμα των αποστάσεων των σημείων από αυτή την ευθεία είναι ελάχιστο. Η διεύθυνση μέγιστης διακύμανσης ονομάζεται πρωτεύων άξονας 1. Οι κάθετοι σε αυτόν είναι οι υπόλοιποι πρωτεύοντες άξονες (2, 3 κ.λπ. κατά σειρά φθίνουσας διακύμανσης), που ορίζουν ένα νέο σύστημα συντεταγμένων, με νέες μεταβλητές που ονομάζονται πρωτεύοντα συστατικά (principal components) και είναι ασυσχέτιστες.
2. *Ελαχιστοποίηση συσχέτισης*: Κατά το μετασχηματισμό της PCA, η μεγιστοποίηση της διακύμανσης οδηγεί ταυτόχρονα τις νέες μεταβλητές (τα πρωτεύοντα συστατικά) να είναι μεταξύ τους ασυσχέτιστες, δηλαδή να έχουν μηδενική συνδιακύμανση, και κατά συνέπεια είναι πιθανό να είναι και ανεξάρτητες μεταξύ τους. Πρακτικά, αυτό σημαίνει ότι η μεταβολή της μιας μεταβλητής δεν οδηγεί σε μεταβολή της άλλης. Ο μετασχηματισμός της PCA επιτυγχάνεται με περιστροφή των αρχικών δεδομένων ώστε το νέο σύστημα συντεταγμένων να γίνει αυτό των πρωτευόντων αξόνων. Έτσι, η μία μεταβλητή αποκτά μεγάλη διακύμανση (πρωτεύοντας άξονας 1), ενώ γύρω από αυτήν (στις υπόλοιπες μεταβλητές, δηλ. στον πρωτεύοντα άξονα 2 και τυχόν υπόλοιπους) η διακύμανση είναι μικρή.
3. *Ελάττωση διαστάσεων*: Η PCA χρησιμοποιείται συνήθως για ελάττωση του πλήθους των διαστάσεων των δεδομένων (features reduction) μέσω της επιλογής των πρωτευόντων συστατικών με τις μεγαλύτερες διακυμάνσεις. Με τον τρόπο αυτό πραγματοποιείται προβολή των δεδομένων σε χώρο λιγότερων διαστάσεων μετά το μετασχηματισμό, με διατήρηση, όμως, του μεγαλύτερου μέρους της χρήσιμης πληροφορίας (ασυσχέτιστες μεταβλητές μεγάλης διακύμανσης).

Αποτέλεσμα αυτού είναι η καλύτερη κατανόηση και οπτικοποίηση και η ευκολότερη επεξεργασία τους.

Για την υλοποίηση της μεθόδου PCA στα χαρακτηριστικά μας έγινε χρήση του αλγορίθμου PCA και της μεθόδου fit_transform τής υποενότητας Decomposition της βιβλιοθήκης scikit-learn. Ο αλγόριθμος PCA, μέσω της παραμέτρου n_components, μας δίνει τη δυνατότητα να θέσουμε το επιθυμητό ποσοστό τής διακύμανσης στα δεδομένα που συνεπάγεται αντίστοιχη διατήρηση πληροφορίας. Αυτό το ορίζουμε στο config αρχείο μέσω της παραμέτρου pca_tolerance και για τα πειράματά μας τέθηκε ίσο με 0.9.

4.4 Εξισορρόπηση συνόλου δεδομένων

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για την ταξινόμηση σχεδιάστηκαν γύρω από την υπόθεση ίσου αριθμού δειγμάτων για κάθε κλάση. Κάθε άλλο σενάριο ενδέχεται να τους δημιουργήσει προβλήματα, εκτός ίσως αν η διαφορά είναι ελάχιστη. Παρόλα αυτά, καθόλου ασυνήθιστο δεν είναι να συναντούμε ανισορροπία στα δεδομένα. Υπάρχουν προβλήματα όπου μια ανισορροπία δεν είναι απλώς συνηθισμένη, αλλά αναμενόμενη. Για παράδειγμα, σύνολα δεδομένων όπως αυτά που χαρακτηρίζουν δόλιες συναλλαγές είναι μη ισορροπημένα. Η συντριπτική πλειοψηφία των συναλλαγών θα είναι στην κατηγορία «Μη απάτη» και μια πολύ μικρή μειονότητα στην κατηγορία «Απάτη». Όταν το σύνολο δεν είναι ισορροπημένο οι κλάσεις χωρίζονται στις εξής δύο κατηγορίες:

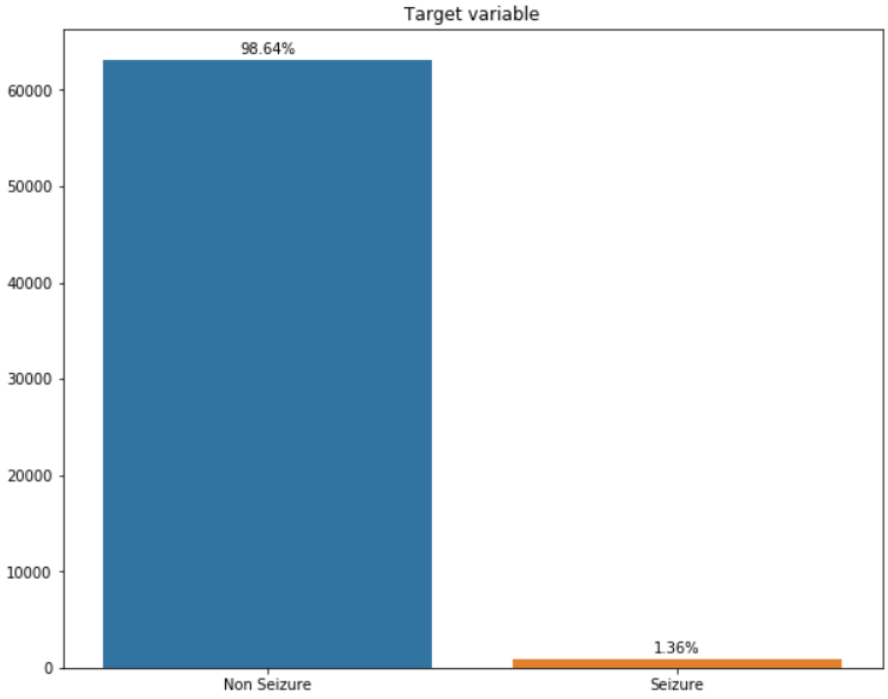
- *Επικρατούσα κλάση (Majority class)*: Περισσότερα από τα μισά δείγματα ανήκουν σε αυτήν την κλάση, συνήθως η αρνητική ή φυσιολογική περίπτωση.
- *Μειονεκτούσα κλάση (Minority class)*: Λιγότερα από τα μισά παραδείγματα ανήκουν σε αυτήν την κλάση, συνήθως η θετική ή μη φυσιολογική περίπτωση.

Όπως προείπαμε στην ενότητα 1.2.5 και βλέπουμε και στο Σχήμα 14, υπάρχει μια συντριπτική πλειοψηφία της μη επιληπτικής-φυσιολογικής κατηγορίας δειγμάτων (98.64%) έναντι της επιληπτικής-μη φυσιολογικής (1.36%). Υπενθυμίζουμε ότι από το αρχικό σύνολο δεδομένων κρατήσαμε μόνο τα 18 κανάλια που είναι κοινά σε όλους τους ασθενείς και επιλέξαμε μόνο τις καταγραφές που περιείχαν επιληπτικές κρίσεις μέσα (βλ. ενότητες 1.2.1 και 3.4). Παρόλα αυτά η αναλογία παραμένει απαγορευτική για να εφαρμόσουμε τους αλγορίθμους ταξινόμησης. Για να μετατρέψουμε το σύνολό μας σε ένα ισορροπημένο σύνολο δεδομένων υπάρχουν δύο τρόποι:

- *Υποδειγματοληψία επικρατούσας κλάσης*: Αφαίρεση δειγμάτων από την πλειοψηφία
- *Υπερδειγματοληψία μειονεκτούσας κλάσης*: Προσθήκη δειγμάτων στη μειονεκτούσα κλάση, δημιουργία συνθετικών δεδομένων

Η επιλογή μας είναι να χρησιμοποιήσουμε συνδυασμό των δύο παραπάνω τρόπων και όχι έναν από αυτούς. Η εξήγηση είναι απλή και οφείλεται στη χαοτική διαφορά των δύο κλάσεων. Αν χρησιμοποιούσαμε μόνο την υποδειγματοληψία επικρατούσας κλάσης, θα καταλήγαμε με ένα πολύ μικρό σύνολο δεδομένων που δε θα μας δώσει αντιπροσωπευτικά αποτελέσματα. Αντίστοιχα, αν χρησιμοποιούσαμε μόνο την υπερδειγματοληψία μειονεκτούσας κλάσης θα έπρεπε να παράξουμε τόσο πολλά δείγματα, που πολλά θα ήταν όμοια και δε θα μας έδιναν νέες πληροφορίες, ενώ το σύνολο δεδομένων θα μεγάλωνε αρκετά. Οπότε, επιλέγουμε πρώτα να υποδειγματοληπτήσουμε την επικρατούσα κλάση κατά ένα ποσοστό και μετά να δημιουργήσουμε όσα συνθετικά δεδομένα χρειάζονται στη μειονεκτούσα τάξη, ώστε να

επέλθει ισορροπία. Το ποσοστό υποδειγματοληψίας ορίζεται στο config αρχείο ως undersampling_rate και για τα πειράματα μας τίθεται ίσο με 0.2.



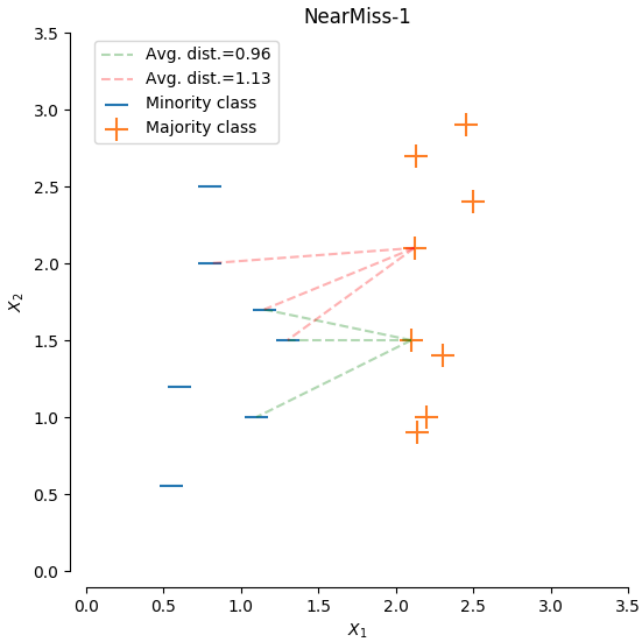
Σχήμα 14: Κατανομή δειγμάτων συνόλου δεδομένων σε κλάσεις

4.4.1 Υποδειγματοληψία επικρατούσας κλάσης

Οι μέθοδοι υποδειγματοληψίας λειτουργούν με την πλειοψηφία. Σε αυτές τις μεθόδους εξαλείφουμε δείγματα τής επικρατούσας κλάσης. Μειώνουμε τον αριθμό των δειγμάτων από την επικρατούσα κλάση για να ισορροπήσει το σύνολο δεδομένων. Αυτό οδηγεί σε σοβαρή απώλεια πληροφοριών. Αυτή η τεχνική εφαρμόζεται όταν το σύνολο δεδομένων είναι τεράστιο.

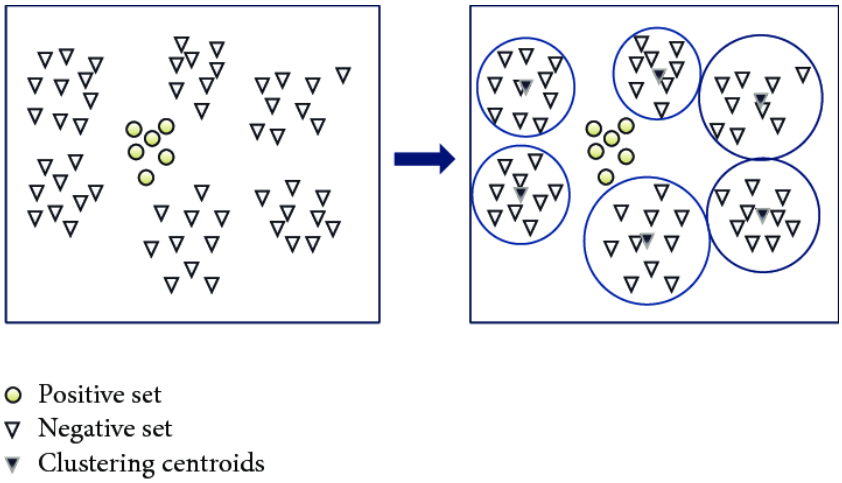
Για την εφαρμογή της υποδειγματοληψίας έγινε χρήση της υποενότητας under-sampling της βιβλιοθήκης imblearn. Στα πλαίσια τής υλοποίησης μας δοκιμάσαμε τρεις μεθόδους:

1. *Near Miss 1*: Δειγματοληψία μόνο των σημείων τής επικρατούσας κλάσης που είναι απαραίτητα για τη διάκριση της από τις άλλες. Στην τεχνική δειγματοληψίας NearMiss-1 επιλέγουμε δείγματα από την επικρατούσα κλάση για τα οποία η μέση απόσταση των N πλησιέστερων δειγμάτων τής μειονεκτούσας κλάσης είναι η μικρότερη. Στο config αρχείο δηλώνουμε την παράμετρο undersampling_neighbors που ορίζει το μέγεθος της γειτονιάς για να υπολογιστεί η μέση απόσταση από τα δείγματα της μειονεκτούσας κλάσης και για τα πειράματα μας τη θέσαμε ίση με 3.



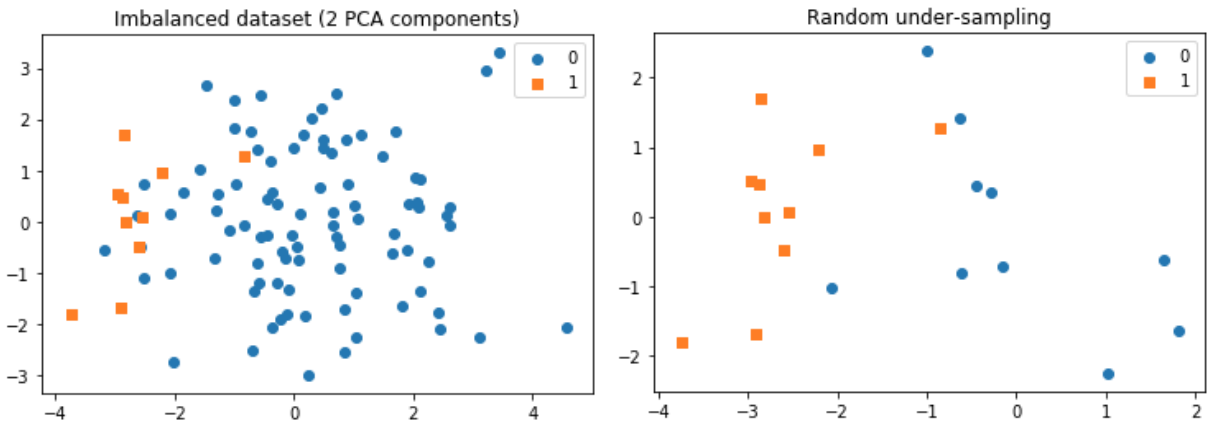
Σχήμα 15: Τεχνική δειγματοληψίας NearMiss-1

2. *Cluster Centroids*: Το σύνολο δεδομένων της επικρατούσας κλάσης χωρίζεται μέσω του αλγορίθμου KMeans σε μικρότερες ομάδες Ν δειγμάτων, τις συστάδες (clusters). Κάθε συστάδα υπολογίζει το κεντροειδές (centroid) της, το οποίο στη συνέχεια γίνεται το νέο δείγμα της κλάσης.



Σχήμα 16: Τεχνική δειγματοληψίας Cluster Centroids

3. *Random Undersampler*: Στη μέθοδο τυχαίας υποδειγματοληψίας εξισορροπούμε την κατανομή των κλάσεων επιλέγοντας και εξαλείφοντας τυχαία τις παρατηρήσεις από την τάξη πλειοψηφίας για να κάνουμε το σύνολο δεδομένων ισορροπημένο.



Σχήμα 17: Παράδειγμα τυχαίας υποδειγματοληψίας

Για τα πειράματά μας χρησιμοποιήσαμε τη μέθοδο Cluster Centroids. Ο χρήστης μπορεί να χρησιμοποιήσει κάποια άλλη αλλάζοντας από το config αρχείο την παράμετρο undersampling_method.

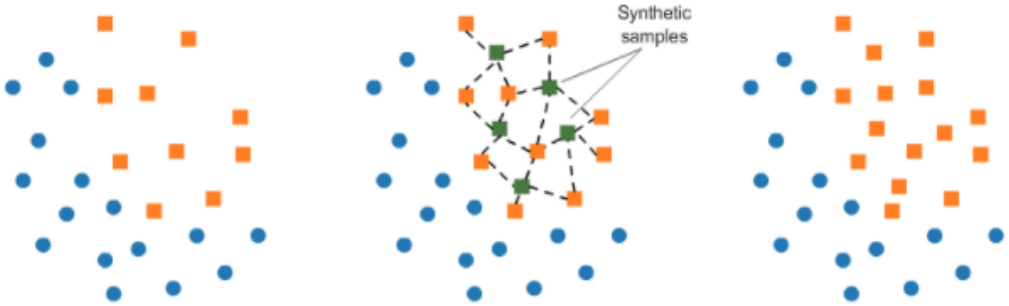
4.4.2 Υπερδειγματοληψίας μειονεκτούσας κλάσης

Οι μέθοδοι υπερδειγματοληψίας λειτουργούν με τη μειονεκτούσα κλάση. Σε αυτές τις μεθόδους δημιουργούμε συνθετικά δεδομένα βασιζόμενοι στα υπάρχοντα δεδομένα τής επικρατούσας κλάσης. Η σύνθεση δεδομένων ενδέχεται να οδηγήσει σε υπερπροσαρμογή λόγω πολλών παρόμοιων σημείων.

Για την εφαρμογή της υποδειγματοληψίας έγινε χρήση της υποενότητας over-sampling της βιβλιοθήκης imblearn. Στο config αρχείο δηλώνουμε την παράμετρο oversampling_neighbors, που ορίζει το μέγεθος της γειτονιάς με βάση την οποία θα δημιουργήσουμε τα συνθετικά δεδομένα, και για τα πειράματα μας τη θέσαμε ίση με 11. Στα πλαίσια τής κατασκευής συνθετικών δεδομένων χρησιμοποιήσαμε τις ακόλουθες δύο μεθόδους:

- *SMOTE*: ισχυρή και ευρέως χρησιμοποιούμενη μέθοδος. Σύμφωνα με αυτήν την τεχνική, τα συνθετικά δεδομένα δημιουργούνται βάσει του χώρου των χαρακτηριστικών με bootstrapping και αλγόριθμο k εγγύτερων γειτόνων (kNN). Λειτουργεί ως εξής:
    1. Πρώτα απ' όλα λαμβάνουμε τη διαφορά μεταξύ του υπό εξέταση δείγματος και του πλησιέστερου γείτονά του.
    2. Στη συνέχεια πολλαπλασιάζουμε αυτήν τη διαφορά με έναν τυχαίο αριθμό μεταξύ 0 και 1.
    3. Μετά προσθέτουμε αυτόν τον αριθμό στο υπό εξέταση διάνυσμα χαρακτηριστικών.
    4. Έτσι επιλέγουμε ένα τυχαίο σημείο κατά μήκος του γραμμικού τμήματος μεταξύ δύο συγκεκριμένων χαρακτηριστικών.

  Συνεπώς, το SMOTE δημιουργεί νέες παρατηρήσεις με παρεμβολή μεταξύ των υπαρχουσών παρατηρήσεων στο σύνολο δεδομένων, όπως φαίνεται και στο παράδειγμα του σχήματος 18.



Σχήμα 18: Παράδειγμα σύνθεσης δειγμάτων SMOTE

- *ADASYN*: Αυτή η τεχνική λειτουργεί με παρόμοιο τρόπο με το SMOTE. Όμως ο αριθμός των παραγόμενων δειγμάτων είναι ανάλογος με τον αριθμό των κοντινών δειγμάτων που δεν ανήκουν στην ίδια τάξη. Έτσι, επικεντρώνεται στα σημεία που διαφέρουν σημαντικά κατά τη δημιουργία των νέων δειγμάτων εκπαίδευσης.

Για τα πειράματά μας χρησιμοποιήσαμε τη μέθοδο ADASYN. Ο χρήστης μπορεί να χρησιμοποιήσει τη SMOTE αλλάζοντας από το config αρχείο την παράμετρο undersampling_method.

4.5 Διαχωρισμός δεδομένων εκπαίδευσης και δεδομένων αξιολόγησης

Οι αλγόριθμοι απαιτούν ένα σύνολο δεδομένων εκπαίδευσης και ένα σύνολο αξιολόγησης. Θα δοκιμάσουμε και θα συγκρίνουμε τόσο τον εξαρχής διαχωρισμό τού συνόλου στα δύο όσο και τη διασταυρούμενη επικύρωση και θα αξιολογήσουμε αν η δεύτερη μας δίνει καλύτερα αποτελέσματα από την πρώτη. Για το θεωρητικό υπόβαθρο των δύο μεθόδων έγινε αναφορά στην ενότητα 1.2.6 και δεν κρίνεται σκόπιμο να επανέλθουμε. Ενδιαφέρον παρουσιάζουν μόνο οι λεπτομέρειες υλοποίησης και οι βιβλιοθήκες που χρησιμοποιήθηκαν. Στο config αρχείο χρησιμοποιείται η παράμετρος test_ratio, που προσδιορίζει το μέγεθος του συνόλου αξιολόγησης, για να ορίσουμε την αναλογία των δύο συνόλων. Για τα πειράματά μας η τιμή της παραμέτρου ήταν 0.3, δηλαδή 70% δεδομένα εκπαίδευσης και 30% δεδομένα αξιολόγησης, με δεδομένο ότι το σύνολο εκπαίδευσης πρέπει να έχει σημαντικά περισσότερα δεδομένα από το σύνολο αξιολόγησης.

4.5.1 Διαχωρισμός συνόλου δεδομένων

Για την υλοποίηση του διαχωρισμού του συνόλου δεδομένων έγινε χρήση της μεθόδου train_test_split της υποενότητας model_selection της βιβλιοθήκης scikit-learn. Αφήσαμε true το όρισμα shuffle, ώστε τα δεδομένα να ανακατεύονται και, κατά συνέπεια, η μεθοδολογία μας να είναι όσο το δυνατόν πιο ανεξάρτητη από τους ασθενείς. Στόχος ήταν να μην έχουμε συνεχόμενες καταγραφές του ίδιου ασθενούς και η επόμενη τιμή να μην αναμένεται με βάση την προηγούμενη. Ιδιαίτερα, δε θα θέλαμε να υπάρχουν συνεχόμενα δεδομένα από τον ίδιο ασθενή γιατί, όπως είδαμε και στο Κεφάλαιο 2, διαφορετικοί ασθενείς παρουσιάζουν διαφορετικά συμπτώματα στην κατάσταση κρίσης και τη μεσοκρισική κατάσταση, γεγονός που κάνει το έργο τού ταξινομητή δυσκολότερο.

4.5.2 Διασταυρούμενη επικύρωση

Για την υλοποίηση του διαχωρισμού του συνόλου δεδομένων έγινε χρήση της μεθόδου KFold της υποενότητας model_selection της βιβλιοθήκης scikit-learn. Στο config αρχείο χρησιμοποιείται η παράμετρος k_fold για να ορίσουμε τις κ επαναλήψεις, που για τα πειράματά μας ήταν 5. Για τους ίδιους λόγους με το διαχωρισμό συνόλου δεδομένων, θέσαμε το όρισμα shuffle αληθές.

5. ΑΛΓΟΡΙΘΜΟΙ ΤΑΞΙΝΟΜΗΣΗΣ

Η ταξινόμηση ή κατηγοριοποίηση εξετάζει τα γνωρίσματα ενός στιγμιοτύπου και το αντιστοιχεί σε μία προκαθορισμένη κλάση. Ένα μοντέλο ταξινόμησης κατασκευάζεται λαμβάνοντας ως σύνολο εκπαίδευσης ένα πλήθος ταξινομημένων δεδομένων και χρησιμοποιώντας κατάλληλους αλγορίθμους καταφέρνει να αποδώσει σωστά την κατηγορία σε άγνωστα δεδομένα. Αφού, λοιπόν, περιγράψαμε στο προηγούμενο κεφάλαιο όλη τη διαδικασία παραγωγής ενός κατάλληλου συνόλου χαρακτηριστικών, σε αυτό το κεφάλαιο θα ασχοληθούμε με τους αλγορίθμους ταξινόμησης που χρησιμοποιήθηκαν, αλλά και τη λογική με την οποία επιλέχθηκαν οι παράμετροι τους.

5.1 Βελτιστοποίηση υπερπαραμέτρων

Οι περισσότεροι από τους αλγορίθμους μάθησης διαθέτουν ένα σύνολο παραμέτρων που χωρίζονται σε δύο τύπους. Ο πρώτος τύπος είναι οι παράμετροι που μαθαίνονται από τα δεδομένα κατά τη διάρκεια εκπαίδευσης του αλγορίθμου, όπως τα βάρη στην παλινδρόμηση. Ο δεύτερος τύπος αφορά το σύνολο εκείνο των παραμέτρων, των λεγόμενων υπερπαραμέτρων (hyperparameters), που πρέπει να καθοριστούν πριν την εκπαίδευση του αλγορίθμου. Η επιλογή των κατάλληλων τιμών για τις δεύτερες μπορεί να επηρεάσει σε σημαντικό βαθμό την απόδοση και τη συμπεριφορά τού τελικού μοντέλου και απαιτούν προσεκτική βελτιστοποίηση.

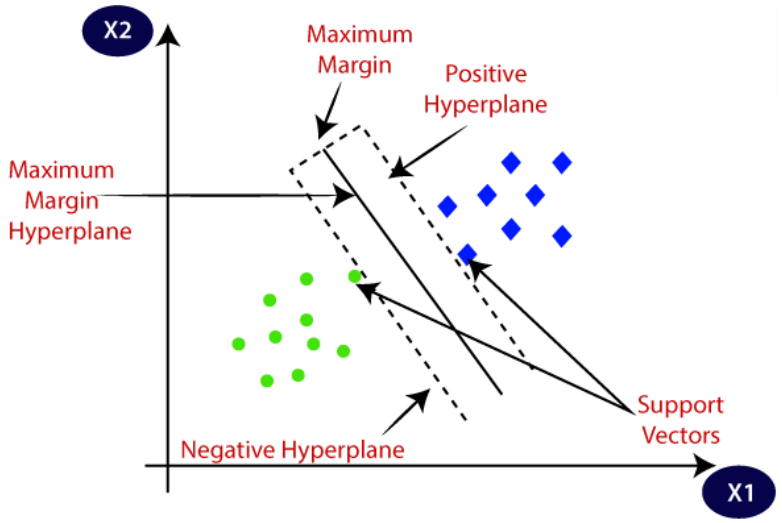
Προκειμένου να επιτύχουμε την επιλογή ενός κατάλληλου συνόλου υπερπαραμέτρων δεν καταφύγαμε απλά στη λύση της χρήσης των προεπιλεγμένων τιμών, αλλά προχωρήσαμε σε βελτιστοποίηση υπερπαραμέτρων (hyperparameter tuning) και συγκεκριμένα την αναζήτηση πλέγματος (grid search). Η αναζήτηση πλέγματος αποτελεί αλγόριθμο εξαντλητικής αναζήτησης και μία από τις πιο διαδομένες τεχνικές στην βελτιστοποίηση υπερπαραμέτρων. Συνήθως, η αναζήτηση πλέγματος συνδυάζεται με την διασταυρούμενη επικύρωση και αξιολογείται με μια μετρική απόδοσης. Στα προβλήματα κατηγοριοποίησης, η μετρική που συνήθως χρησιμοποιείται είναι η ακρίβεια. Δεδομένου, λοιπόν, ενός συνόλου υπερπαραμέτρων η αναζήτηση πλέγματος αξιολογεί την απόδοση του μοντέλου για κάθε συνδυασμό τιμών. Βρίσκει το μοντέλο με την καλύτερη απόδοση και επιστρέφει τον συνδυασμό των υπερπαραμέτρων που πέτυχαν την υψηλότερη βαθμολογία.

Η υλοποίηση της αναζήτησης πλέγματος έγινε μέσω της μεθόδου GridSearchCV της υποενότητας model_selection της βιβλιοθήκης sklearn. Ως μετρική χρησιμοποιούμε την ακρίβεια, την οποία συνδυάζουμε με διασταυρούμενη επικύρωση 5 επαναλήψεων. Το πλέγμα των παραμέτρων προκύπτει από τις δημοφιλέστερες παραλλαγές της βιβλιογραφίας. Έτσι, όλες οι τιμές που χρησιμοποιήθηκαν στις μεθόδους που θα ακολουθήσουν στο τρέχον κεφάλαιο προκύπτουν από το αποτέλεσμα αυτής της διαδικασίας.

5.2 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

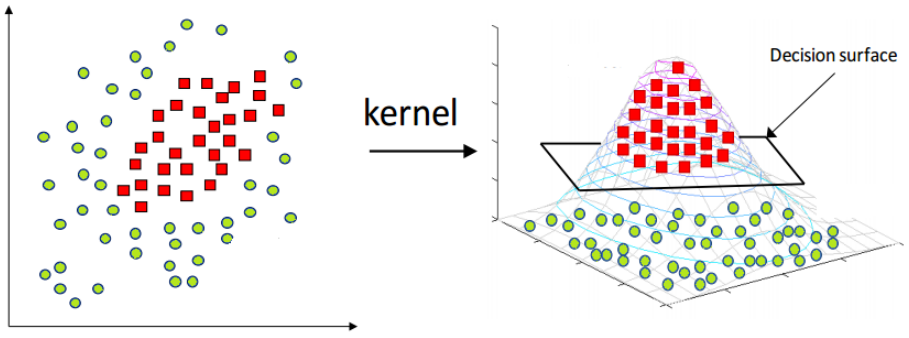
Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) είναι μια μέθοδος μηχανικής μάθησης για δυαδικά προβλήματα ταξινόμησης. Στόχος του είναι η εύρεση ενός υπερεπιπέδου σε έναν N-διάστατο χώρο (N: πλήθος χαρακτηριστικών) που ταξινομεί με σαφήνεια τα σημεία δεδομένων. Με τον όρο υπερεπίπεδο αναφερόμαστε σε σύνορα αποφάσεων που βοηθούν στην ταξινόμηση των σημείων δεδομένων. Δεν υπάρχει μόνο ένα υπερεπίπεδο που να διαχωρίζει τα σημεία των δύο κλάσεων. Στόχος του αλγορίθμου είναι η εύρεση τού υπερεπιπέδου που έχει τη μέγιστη απόσταση

(maximum margin) από τα σημεία και των δύο κλάσεων. Για την κατασκευή του βέλτιστου υπερεπιπέδου σημαντικό ρόλο παίζουν τα διανύσματα υποστήριξης (support vectors). Τα διανύσματα υποστήριξης είναι αυτά τα πρότυπα που στο διανυσματικό χώρο βρίσκονται πλησιέστερα στην επιφάνεια διαχωρισμού, είναι δηλαδή πιο δύσκολο να ταξινομηθούν σωστά.



Σχήμα 19: Παράδειγμα SVM για γραμμικά διαχωρίσιμα στον δισδιάστατο χώρο

Στο Σχήμα 19 βλέπουμε ένα απλό παράδειγμα γραμμικά διαχωρίσιμων δεδομένων στο δισδιάστατο χώρο. Συνήθως, όμως, δεν είναι τόσο απλά τα πράγματα και τα δεδομένα δε μπορούν να διαχωριστούν απευθείας γραμμικά. Ο SVM λύνει το πρόβλημα αυτό μετασχηματίζοντας τα δεδομένα σε έναν καινούργιο χώρο χαρακτηριστικών υψηλότερης διάστασης, όπου τα πρότυπα είναι γραμμικά διαχωρίσιμα, όπως φαίνεται και στο Σχήμα 20. Η απεικόνιση αυτή σε έναν άλλο χώρο λέγεται πυρήνας (kernel) της SVM.



Σχήμα 20: Παράδειγμα SVM για μη γραμμικά διαχωρίσιμα στον δισδιάστατο χώρο

Για λόγους πληρότητας, θα περιγράψουμε, χωρίς πολλές λεπτομέρειες, το μαθηματικό μοντέλο της μεθόδου SVM. Στόχος του ταξινομητή είναι η σχεδίαση ενός υπερεπιπέδου $g(x) = w^T x + w_0 = 0$, το οποίο διαχωρίζει όλα τα στιγμιότυπα του σετ εκπαίδευσης σε δύο κλάσεις, αφήνοντας το μεγαλύτερο περιθώριο μεταξύ τους. Η απόσταση ενός σημείου από ένα υπερεπίπεδο ισούται με $z = \frac{|g(x)|}{\|w\|}$. Για να ισούται το $g(x)$ στα πλησιέστερα σημεία των δύο κλάσεων με ± 1, πρέπει να μεγιστοποιηθεί η απόσταση $\frac{2}{\|w\|}$

με τις παραμέτρους w, w_0 να ικανοποιούν τις εξισώσεις: $\begin{cases} w^T x + w_0 \geq 1, \forall x \in \omega_1 \\ w^T x + w_0 \leq -1, \forall x \in \omega_2 \end{cases}$. Τελικά, οι παράμετροι w, w_0 του υπερεπιπέδου υπολογίζονται έτσι ώστε να ελαχιστοποιείται η ποσότητα $\frac{\|w\|^2}{2}$ ικανοποιώντας τους περιορισμούς $y_i \cdot (w^T x + w_0) \geq 1, i = 1,2, \dots, N$

Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή SVC της υποενότητας svm της βιβλιοθήκης scikit-learn. Οι σημαντικότερες παράμετροι, στις οποίες εφαρμόσαμε αναζήτηση πλέγματος για να βρούμε τις κατάλληλες τιμές, είναι:

- kernel: Η λειτουργία του πυρήνα είναι να μετατρέψει τα δεδομένα εισόδου στην απαιτούμενη μορφή, όπως ήδη είδαμε. Πειραματιστήκαμε με τρεις πυρήνες, το γραμμικό (linear), τον πολυωνυμικό (poly) και τη συνάρτηση ακτινικής βάσης (rbf). Τελικά, καταλήξαμε στον πολυωνυμικό σε όλα τα πειράματα. Στο πείραμα του μέσου καναλιού η rbf δίνει καλύτερα αποτελέσματα ως προς την ακρίβεια, αλλά, αν λάβουμε υπόψη και τις άλλες μετρικές, δεν ήταν μια συνολικά καλύτερη ταξινόμηση, αφού παρούσιαζε πολλές ψευδώς αρνητικές τιμές.
- degree: Βαθμός της πολυωνυμικής συνάρτησης πυρήνα (poly). Δοκιμάστηκαν οι τιμές 2,3 και 4 και σε όλα τα πειράματα χρησιμοποιήσαμε την προεπιλεγμένη τιμή 3. Στο πείραμα μέσου καναλιού, ο βαθμός 4 δίνει καλύτερα αποτελέσματα σε όλες τις μετρικές, αλλά ο χρόνος που διαρκεί μια εκτέλεσή του είναι απαγορευτικός. Αντίστοιχα και στο πείραμα των καναλιών αριστερού και δεξιού ημισφαιρίου ο βαθμός 4 έδινε καλύτερες μετρήσεις και χωρίς να αυξάνεται δραματικά ο χρόνος εκτέλεσης, αλλά ελλοχεύει ο κίνδυνος υπερπροσαρμογής του ταξινομητή στα δεδομένα εκπαίδευσης. Τέλος, στο πείραμα με όλα τα δεδομένα δεν εντοπίζεται σημαντική διαφορά σε χρόνους και μετρήσεις αξιολόγησης, οπότε επελέγη ο βαθμός 3 για λόγους συνέπειας με τα άλλα δύο πειράματα.
- C: Παράμετρος κανονικοποίησης. Η ισχύς της κανονικοποίησης είναι αντιστρόφως ανάλογη με το C που πρέπει να είναι αυστηρά θετικό. Η ποινή είναι μια τετράγωνη ποινή L2. Οι τιμές που δοκιμάσαμε ήταν οι [0.1, 1, 10, 100, 1000] και επιλέξαμε την προεπιλεγμένη τιμή C=1.
- gamma: Η παράμετρος αυτή καθορίζει το μέγεθος της επιρροής ενός μόνο σημείου εκπαίδευσης. Αυτό σημαίνει ότι η υψηλότερη τιμή θα λαμβάνει υπόψη μόνο σημεία κοντά στο υπερεπίπεδο και οι χαμηλότερες τιμές θα λαμβάνουν υπόψη σημεία σε μεγαλύτερη απόσταση. Οι πιθανές τιμές που εξετάστηκαν ήταν [0.0001, 0.001, 0.01, 0.1,1,10] μαζί με τις τιμές scale και auto. Η τιμή που χρησιμοποιήσαμε και στα τρία πειράματα είναι η auto.

5.3 Κ Εγγύτεροι Γείτονες (KNN)

Ο αλγόριθμος Κ Εγγύτερων Γειτόνων (KNN) είναι ένας από τους απλούστερους, αλλά ταυτόχρονα και δημοφιλέστερους αλγόριθμους μηχανικής μάθησης που χρησιμοποιούνται σε προβλήματα ταξινόμησης. Ο KNN είναι ένας μη παραμετρικός και οκνηρός (lazy) αλγόριθμος. Μη παραμετρικός σημαίνει ότι δεν υπάρχει καμία υπόθεση για την κατανομή των δεδομένων και η δομή του μοντέλου καθορίζεται από το σύνολο δεδομένων. Αυτό είναι ιδιαίτερα χρήσιμο γιατί τα πραγματικά σύνολα δεδομένων δεν ακολουθούν μαθηματικές παραδοχές και θεωρητικά μοντέλα. Οκνηρός σημαίνει ότι δε χρειάζεται ιδιαίτερη εκπαίδευση και όλα τα δεδομένα εκπαίδευσης χρησιμοποιούνται στη φάση της ταξινόμησης.

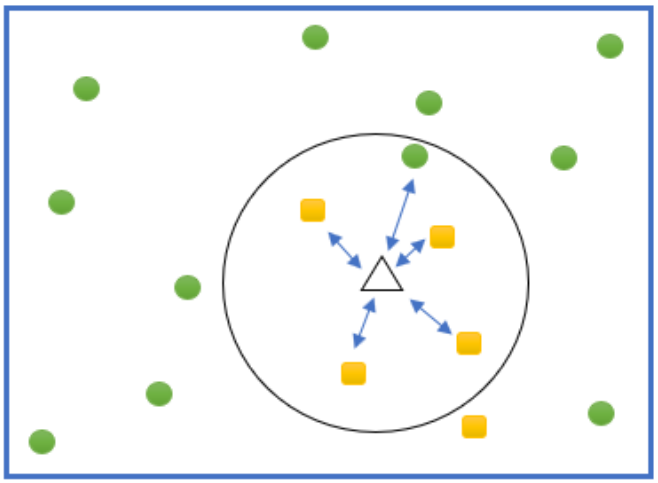
Ο kNN βασίζεται στην υπόθεση ότι παρόμοια δεδομένα υπάρχουν σε κοντινή απόσταση. Ο αλγόριθμος χρησιμοποιεί τα αποθηκευμένα δεδομένα για να βρει ένα συγκεκριμένο

πλήθος των πιο όμοιων προτύπων εκπαίδευσης (k κοντινότεροι γείτονες), σύμφωνα με μια μετρική απόστασης. Το νέο δεδομένο δοκιμής ανατίθεται στην κατηγορία που είναι επικρατέστερη μεταξύ των κοντινότερων γειτόνων του, δηλαδή η κατάταξη του δεδομένου δοκιμής καθορίζεται από τα δεδομένα εκπαίδευσης που επικρατούν σε αριθμό σε εκείνη την περιοχή. Στην περίπτωση που τα k κοντινότερα πρότυπα δεν είναι όλα της ίδιας τάξης, τότε χρησιμοποιείται η ψήφος πλειοψηφίας ή ψήφος σταθμισμένων αποστάσεων. Το k σε δυαδικά προβλήματα ταξινόμησης επιλέγεται να είναι κάποιος περιττός αριθμός, ώστε να μην υπάρχει περίπτωση ισοβαθμίας, και εξαρτάται εν πολλοίς από τα δεδομένα. Η επιλογή μικρού k κάνει τον αλγόριθμο ευαίσθητο στο θόρυβο, ενώ η επιλογή ενός πολύ μεγάλου k, διευρύνει τα όρια των κλάσεων και έτσι η γειτονία μπορεί να περιέχει σημεία από άλλες κλάσεις. Οι συνηθέστερες μετρικές απόστασης είναι η Ευκλείδεια, η Manhattan, η Mahalanobis και η Chebychev και ο τρόπος υπολογισμού τους φαίνεται στον Πίνακα 3.

Πίνακας 3: Μετρικές αποστάσεων αλγορίθμου KNN

Euclidean Distance $d(x,y) = \sqrt{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^2}$	**Manhattan Distance** $d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Mahalanobis Distance $d(x,y) = \sqrt{(x-y)\Sigma^{-1}(x-y)^T}$ Σ^{-1}: Πίνακας Συνδιασποράς	**Chebychev Distance** $d(x,y) = \max_{i=1,2,\ldots,n} \lvert x_i - y_i \rvert$

Στο Σχήμα 21 βλέπουμε ένα από παράδειγμα εφαρμογής του kNN σε ένα δυαδικό πρόβλημα ταξινόμησης. Οι δύο κλάσεις είναι τα κίτρινα τετράγωνα και οι πράσινοι κύκλοι. Χρησιμοποιούμε kNN με k=5 και θέλουμε να ταξινομήσουμε το τρίγωνο σε μία από τις δύο κλάσεις. Στα 5 κοντινότερα σημεία του, τα 4 είναι κίτρινα τετράγωνα και το 1 είναι πράσινος κύκλος. Άρα, λόγω πλειοψηφίας, συμπεραίνουμε ότι το τρίγωνο ανήκει στην κλάση των κίτρινων τετραγώνων.



Σχήμα 21: Παράδειγμα αλγορίθμου KNN για k=5

Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή KNeighborsClassifier της υποενότητας neighbors της βιβλιοθήκης scikit-learn. Οι σημαντικότερες παράμετροι, στις οποίες εφαρμόσαμε αναζήτηση πλέγματος για να βρούμε τις κατάλληλες τιμές, είναι:

- n_neighbors: Ο αριθμός των γειτόνων, k όπως τον αναφέραμε νωρίτερα. Δοκιμάστηκαν οι τιμές [3,5,7,11,19,31] και τα καλύτερα αποτελέσματα ήταν για k=3.
- metric: Η μέθοδος μέτρησης της απόστασης. Οι πιθανές τιμές φαίνονται στον Πίνακα 3 και αυτή που χρησιμοποιήθηκε στα πειράματά μας είναι η απόσταση Manhattan.
- weights: Συνάρτηση βάρους που χρησιμοποιείται στην πρόβλεψη με πιθανές τιμές τις uniform και distance. Η uniform δίνει ομοιόμορφα βάρη σε όλα τα σημεία κάθε γειτονιάς. Η τιμή distance προβλέπει ότι οι πιο κοντινοί γείτονες έχουν μεγαλύτερη επιρροή από πιο μακρινούς. Μεγαλύτερη ακρίβεια πετύχαμε με τη δεύτερη.

5.4 Απλοϊκός Bayes

Οι πιθανοτικές μέθοδοι είναι από τις πιο θεμελιώδεις μεθόδους στην κατηγοριοποίηση δεδομένων. Οι πιθανοτικοί αλγόριθμοι κατηγοριοποίησης χρησιμοποιούν στατιστικά συμπεράσματα προκειμένου να επιτύχουν την εύρεση της καλύτερης κλάσης δεδομένου ενός στιγμιότυπου. Κεντρική ιδέα του ταξινομητή Bayes είναι ότι η κατηγοριοποίηση ενός νέου στιγμιοτύπου βασίζεται άμεσα στην κατανομή των πιθανοτήτων που έχουν τα υπόλοιπα στιγμιότυπα του συνόλου δεδομένων αναφορικά με τη κλάση που ανήκουν. Ουσιαστικά αναζητείται η κλάση του στιγμιοτύπου ψάχνοντας τη μεγαλύτερη πιθανότητα να ταξινομηθεί με τα συγκεκριμένα γνωρίσματα σε μια από τις κλάσεις. Οι ταξινομητές Bayes βασίζονται στο θεώρημα του Bayes το οποίο διατυπώνεται συνοπτικά ως εξής:

Αν X το στιγμιότυπο προς ταξινόμηση και α οι δυνατές τιμές της κλάσης τότε

$$P(a|X) = \frac{P(X|a)\cdot P(a)}{P(x)}, \text{όπου } \begin{cases} P(a): \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \tau\alpha\xi\iota\nu\acute{o}\mu\eta\sigma\eta\varsigma\ \sigma\tau\eta\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\ \alpha \\ P(x): \epsilon\kappa\ \tau\omega\nu\ \pi\rho o\tau\acute{\epsilon}\rho\omega\nu (a\ priori) \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \tau o\upsilon\ x \\ P(X|a): \epsilon\kappa\ \tau\omega\nu\ \upsilon\sigma\tau\acute{\epsilon}\rho\omega\nu\ (a\ posteriori) \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \tau o\upsilon\ X \end{cases}$$

Για να προβλέψει την κλάση ενός στιγμιοτύπου ένας ταξινομητής Bayes, επιλέγει την κλάση που μεγιστοποιεί την εκ των υστέρων πιθανότητα.

Ο απλοϊκός Bayes βασίζεται στο παραπάνω θεώρημα με την επιπρόσθετη υπόθεση ότι τα γνωρίσματα είναι ανεξάρτητα μεταξύ τους. Έστω ένα στιγμιότυπο X που περιγράφεται από τα γνωρίσματα: $< x_1, x_2, \dots, x_n >$ και αναζητούμε την κλάση στην οποία ανήκει. Οι κατηγορίες της κλάσης σε ένα δυαδικό πρόβλημα συμβολίζονται με $< k_1, k_2 >$. Οπότε, με βάση το θεώρημα του Bayes, αρκεί να υπολογίσουμε για κάθε κλάση την πιθανότητα ένα στιγμιότυπο X να ανήκει στην κατηγορία αυτή:

$$P(k_d|X) = \frac{P(k_d)\cdot P(X|k_d)}{P(X)}, d = 1,2\ \kappa\alpha\iota\ X =< x_1, x_2, \dots, x_n >$$

Το στιγμιότυπο ανήκει στην κατηγορία που έχει μεγαλύτερη πιθανότητα $P(k_d|X)$. Αναλογικά, αρκεί η μεγιστοποίηση του $P(k_d)\cdot P(X|k_d)$. Λόγω της ανεξαρτησίας των γνωρισμάτων έχουμε:

$$P(X|k_d) = P(x_1, x_2, \dots, x_n|k_d) = P(x_1|k_d)\cdot P(x_2|k_d)\cdots P(x_n|k_d) = \prod_{i=1}^{n} P(x_i|k_d)$$

Συνεπώς ο απλοϊκός Bayes ψάχνει ποια από τις δύο κλάσεις μεγιστοποιεί το γινόμενο $\prod_{i=1}^{n} P(x_i|k_d) \cdot P(k_d)$ και ταξινομεί σε αυτή το στιγμιότυπο Χ.

Μια τυπική υπόθεση που γίνεται είναι ότι τα δεδομένα ακολουθούν κανονική κατανομή. Το σύνολο δεδομένων μας έχει ήδη κανονικοποιηθεί, όπως είδαμε και στην ενότητα 4.2.

Οπότε: $P(x_i|k_d) = \frac{1}{\sqrt{2\pi\sigma_{k_d}^2}} \cdot e^{-\frac{\left(x_i-\mu_{k_d}\right)^2}{2\sigma_{k_d}^2}}$, $όπου \begin{cases} \mu_{k_d}: μέση\ τιμή\ κλάσης\ k_d \\ \sigma_{k_d}: τυπική\ απόκλιση\ κλάσης\ k_d \end{cases}$

Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή GaussianNB της υποενότητας naive_bayes της βιβλιοθήκης scikit-learn που δεν έχει υπερπαραμέτρους προς βελτιστοποίηση.

5.5 Δέντρα Απόφασης

Ένα δέντρο απόφασης (decision tree) είναι ένας ταξινομητής που εκφράζεται ως αναδρομική ιεραρχική διχοτόμηση του χώρου γεγονότων. Η γενική προσέγγιση τού αλγορίθμου είναι ότι προσπαθεί αναδρομικά να χωρίσει τα δεδομένα εκπαίδευσης, έτσι ώστε να μεγιστοποιήσει τη διάκριση μεταξύ των διαφορετικών κλάσεων σε διαφορετικούς κόμβους. Όταν μεγιστοποιηθεί το επίπεδο της κλίσης μεταξύ των διαφορετικών κλάσεων τότε μεγιστοποιείται και η διάκριση μεταξύ αυτών. Η ποσοτικοποίηση της κλίσης επιτυγχάνεται είτε με τον δείκτη gini είτε με την εντροπία. Ο τρόπος υπολογισμού τους για έναν κόμβο N για k διαφορετικές κλάσεις φαίνεται στον Πίνακα 4.

Πίνακας 4: Συναρτήσεις μέτρησης ποιότητας διαχωρισμού

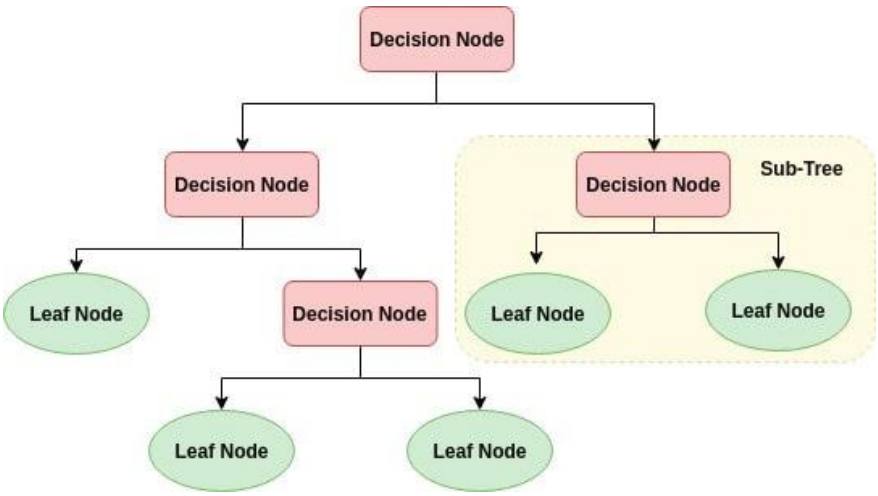
Gini	$G(N) = 1 - \sum_{i=1}^{k} p_i^2$
Entropy	$E(N) = -\sum_{i=1}^{k} p_i \cdot \log(p_i)$

Το δέντρο αποφάσεων αποτελείται από κόμβους που σχηματίζουν ένα δέντρο με ρίζα, το οποίο σημαίνει ότι είναι ένα δέντρο με κατεύθυνση ένα κόμβο που ονομάζεται ρίζα, ο οποίος δεν έχει καμία εισερχόμενη άκρη. Όλοι οι άλλοι κόμβοι έχουν ακριβώς μία εισερχόμενη ακμή. Ένας κόμβος με εξερχόμενη άκρη αναφέρεται ως ένας κόμβος εσωτερικός ή εξέτασης. Όλοι οι άλλοι κόμβοι ονομάζονται φύλλα (επίσης γνωστοί ως τερματικοί κόμβοι ή κόμβοι απόφασης).

Στο δέντρο απόφασης, κάθε εσωτερικός κόμβος χωρίζει το χώρο του γεγονότος σε δύο ή περισσότερους υποχώρους σύμφωνα με μία διακριτή συνάρτηση των αξιών των χαρακτηριστικών εισόδου. Στην πιο απλή και πιο συχνή περίπτωση, κάθε δοκιμή θεωρείται ένα μοναδικό χαρακτηριστικό, έτσι ώστε ο χώρος του γεγονότος να κατανέμεται σύμφωνα με την αξία των χαρακτηριστικών. Στην περίπτωση των αριθμητικών χαρακτηριστικών, η κατάσταση αναφέρεται σε ένα εύρος.

Δύο ή περισσότερα κλαδιά μπορούν να αναπτυχθούν από κάθε εσωτερικό κόμβο (όχι από κόμβο φύλλο). Κάθε κόμβος αντιστοιχεί με ένα συγκεκριμένο χαρακτηριστικό και τα κλαδιά αντιστοιχούν σε ένα εύρος τιμών. Αυτές οι περιοχές τιμών πρέπει να δώσουν ένα μέρος του συνόλου των αξιών του δοσμένου χαρακτηριστικού.

Τα γεγονότα ταξινομούνται με βάση την πλοήγησή τους από τη ρίζα του δέντρου σε ένα φύλλο, σύμφωνα με τα αποτελέσματα των δοκιμών κατά μήκος της διαδρομής. Το σκελετό ενός δέντρου μπορούμε να τον παρατηρήσουμε στο Σχήμα 22.



Σχήμα 22: Σχεδιάγραμμα δέντρου απόφασης

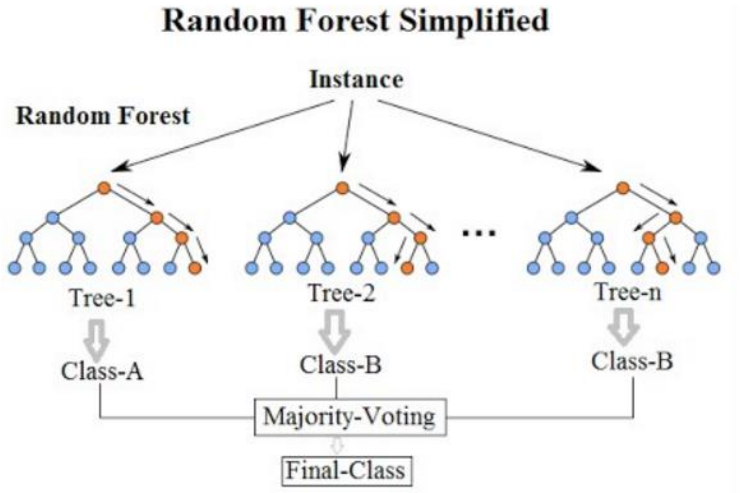
Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή DecisionTreeClassifier τής υποενότητας tree της βιβλιοθήκης scikit-learn. Οι σημαντικότερες παράμετροι, στις οποίες εφαρμόσαμε αναζήτηση πλέγματος για να βρούμε τις κατάλληλες τιμές, είναι:

- criterion: Η συνάρτηση μέτρησης της ποιότητας ενός διαχωρισμού. Οι δύο πιθανές συναρτήσεις είναι η gini και η entropy και ο τρόπος υπολογισμού τους είναι διαθέσιμος στον Πίνακα 4. Στο πείραμα των καναλιών αριστερού και δεξιού ημισφαιρίου χρησιμοποιήσαμε την πρώτη, ενώ στις άλλες δύο τη δεύτερη.
- min_samples_split: Ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για τη διάσπαση ενός εσωτερικού κόμβου. Δοκιμάσαμε όλους τους άρτιους αριθμούς από το 2 μέχρι το 100. Για το πείραμα μέσου καναλιού και αυτό των καναλιών αριστερού και δεξιού ημισφαιρίου ο αριθμός αυτός τέθηκε ίσος με 2, ενώ σε εκείνο με όλα τα κανάλια 4.
- max_depth: Το μέγιστο βάθος τού δέντρου. Εάν δεν υπάρχει, τότε οι κόμβοι επεκτείνονται έως ότου όλα τα φύλλα είναι καθαρά ή έως ότου όλα τα φύλλα περιέχουν λιγότερα δείγματα από min_samples_split. Δοκιμάσαμε όλους τους άρτιους αριθμούς από το 2 μέχρι το 100. Για το πείραμα τού μέσου καναλιού το μέγιστο βάθος είναι 82, για εκείνο των καναλιών αριστερού και δεξιού ημισφαιρίου είναι 100, ενώ για το πείραμα με όλα τα κανάλια το μέγιστο βάθος είναι 70.
- splitter: Η στρατηγική που χρησιμοποιείται για την επιλογή του διαχωρισμού σε κάθε κόμβο. Οι υποστηριζόμενες στρατηγικές είναι οι «best» για να επιλεγεί η καλύτερη διαίρεση και «random» για να επιλεγεί η καλύτερη τυχαία διάσπαση. Στο πείραμα των καναλιών αριστερού και δεξιού ημισφαιρίου χρησιμοποιήσαμε την «best», ενώ στις άλλες δύο τη «random».

5.6 Τυχαίο Δάσος

Ο ταξινομητής τυχαίου δάσους (random forest) είναι μια συνάθροιση πολλαπλών δέντρων αποφάσεων, τα οποία είδαμε στην ενότητα 5.5. Το τυχαίο δάσος συνδυάζει δεκάδες ή εκατοντάδες δέντρα αποφάσεων, εκπαιδεύει το καθένα από αυτά σε ένα

ελαφρώς διαφορετικό σύνολο παρατηρήσεων και χωρίζει κόμβους σε κάθε δέντρο λαμβάνοντας υπόψη έναν περιορισμένο αριθμό χαρακτηριστικών. Κάθε δέντρο μπορεί να θεωρηθεί ως ένας μεμονωμένος ταξινομητής και η έξοδος της ταξινόμησης του δάσους δίνεται πλειοψηφικά από όλα τα δέντρα απόφασης. Μια βασική μορφή ενός τυχαίου δάσους παρουσιάζεται στο σχήμα 23.



Σχήμα 23: Σχεδιάγραμμα τυχαίου δάσους

Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή RandomForestClassifier τής υποενότητας ensemble της βιβλιοθήκης scikit-learn. Οι σημαντικότερες παράμετροι, στις οποίες εφαρμόσαμε αναζήτηση πλέγματος για να βρούμε τις κατάλληλες τιμές, είναι:

- criterion: Η συνάρτηση μέτρησης της ποιότητας ενός διαχωρισμού. Οι δύο πιθανές συναρτήσεις είναι η gini και η entropy και ο τρόπος υπολογισμού τους είναι διαθέσιμος στον Πίνακα 4. Σε όλα τα πειράματα χρησιμοποιήσαμε τη gini.
- min_samples_split: Ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για τη διάσπαση ενός εσωτερικού κόμβου. Δοκιμάσαμε όλους τους άρτιους αριθμούς από το 2 μέχρι το 100 και καταλήξαμε την τιμή 2.
- max_depth: Το μέγιστο βάθος τού δέντρου. Εάν δεν υπάρχει, τότε οι κόμβοι επεκτείνονται έως ότου όλα τα φύλλα είναι καθαρά ή έως ότου όλα τα φύλλα περιέχουν λιγότερα δείγματα από min_samples_split. Δοκιμάσαμε όλους τους άρτιους αριθμούς από το 2 μέχρι το 100. Για το πείραμα τού μέσου καναλιού το μέγιστο βάθος είναι 60, για εκείνο των καναλιών αριστερού και δεξιού ημισφαιρίου είναι 86, ενώ για το πείραμα με όλα τα κανάλια το μέγιστο βάθος είναι 52.
- n_estimators: Ο αριθμός των δέντρων στο δάσος. Δοκιμάσαμε όλα τα πολλαπλάσια του 10 από 10 μέχρι 100. Για το πείραμα τού μέσου καναλιού το χρησιμοποιήσαμε 100 δέντρα, για εκείνο των καναλιών αριστερού και δεξιού ημισφαιρίου 100, ενώ για το πείραμα με όλα τα κανάλια 90 δέντρα.

5.7 Ανάλυση Γραμμικής Διάκρισης (LDA)

Η Γραμμική Διακριτική Ανάλυση (LDA: Linear Discriminant Analysis ή Fisher's LDA) είναι μια μέθοδος μετασχηματισμού δεδομένων που ανήκουν σε κατηγορίες (κλάσεις). Σκοπός της είναι η μείωση διάστασης των δεδομένων και ο καλύτερος διαχωρισμός τους.

Η κεντρική ιδέα της LDA είναι ο μετασχηματισμός των δεδομένων με τέτοιο τρόπο ώστε να μεγιστοποιηθεί η απόσταση μεταξύ των κλάσεων, δηλαδή οι κλάσεις να είναι απομακρυσμένες, και ταυτόχρονα να ελαχιστοποιηθεί η διασπορά εντός των κλάσεων, δηλαδή τα δεδομένα κάθε κλάσης να είναι συγκεντρωμένα γύρω από τη μέση τιμή τους. Η LDA υποθέτει κανονική κατανομή των δεδομένων με ίσες μήτρες συνδιασποράς και για τις δύο κλάσεις.

Για την επίτευξη της καλύτερης διαχωρισιμότητας, τα δεδομένα κατά το μετασχηματισμό τους προβάλλονται σε χώρο μικρότερης διάστασης από τον αρχικό, με αποτέλεσμα να ελαττώνεται το πλήθος των διαστάσεών τους (δηλαδή των μεταβλητών από τις οποίες αποτελούνται), έτσι ώστε στη συνέχεια η ταξινόμηση νέων δεδομένων να είναι ευχερέστερη και ακριβέστερη.

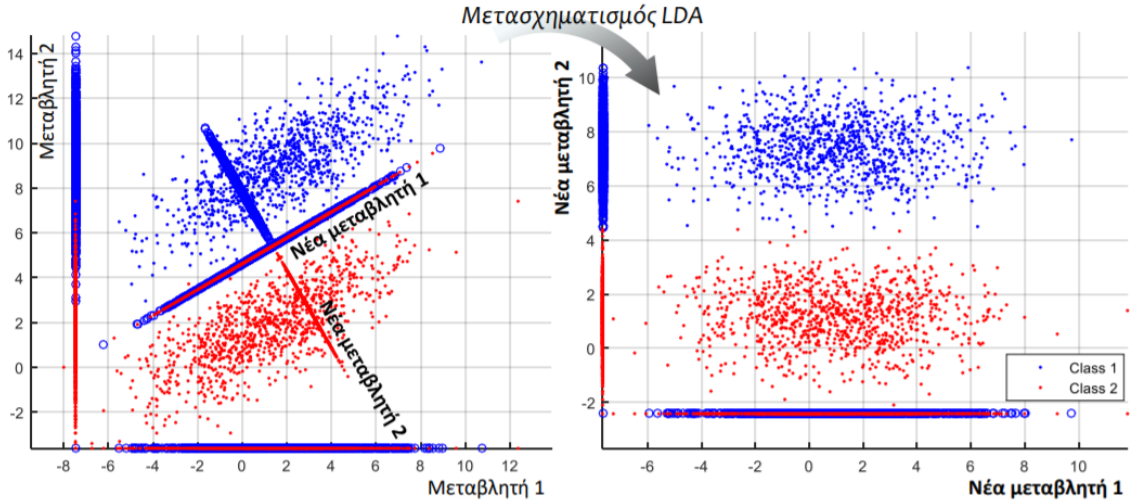
Συχνά, τα δεδομένα δύο ή περισσότερων κλάσεων δεν είναι γραμμικώς διαχωρίσιμα, ενώ μπορεί να αποτελούνται από πολλές μεταβλητές, από τις οποίες να μην είναι εύκολο να βρεθούν εκείνες που διαχωρίζουν καλύτερα τις κατηγορίες. Ο σκοπός της LDA είναι να επιτύχουμε καλό διαχωρισμό μεταξύ των κλάσεων, με ταυτόχρονη ελάττωση διαστάσεων.

Τα κριτήρια καλού διαχωρισμού μεταξύ των κλάσεων (κριτήριο Fisher) είναι:

1. Μεγιστοποίηση αποστάσεων μεταξύ των κλάσεων (απομακρυσμένες μέσες τιμές)
2. Ελαχιστοποίηση διασποράς μεταξύ των κλάσεων (συγκέντρωση γύρω από μέση τιμή)

Συγκεκριμένα, το κριτήριο Fisher είναι ο λόγος των δύο παραπάνω μεγεθών, ο οποίος, όταν μεγιστοποιείται, θεωρούμε ότι οι κλάσεις είναι καλύτερα διαχωρισμένες, δηλαδή η επικάλυψη τους είναι η μικρότερη δυνατή. Συνεπώς, πρέπει να βρεθεί κατάλληλη διεύθυνση προβολής των δεδομένων έτσι ώστε να μεγιστοποιηθεί αυτό το κλάσμα.

Με προβολή των δεδομένων στην κατάλληλη διεύθυνση, ο λόγος της απόστασης μεταξύ των κλάσεων προς τη διασπορά εντός των κλάσεων γίνεται μέγιστος, όπως στο Σχήμα 24. Συγκεκριμένα, παρατηρούμε ότι, μετά το μετασχηματισμό κατά LDA, στον άξονα της Νέας μεταβλητής 2 ο διαχωρισμός των κλάσεων είναι βέλτιστος.



Σχήμα 24: Παράδειγμα μετασχηματισμού LDA

Εξετάζοντας ελαφρώς το μαθηματικό μοντέλο, έστω 2 κλάσεις X_1 και X_2 σε πίνακες διαστάσεων $M \times N_k$ (M:μεταβλητές, N_k:παρατηρήσεις). Το κριτήριο Fisher διαχωρισμού

των κλάσεων είναι $J = \frac{S_b}{S_w}$, όπου $S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ η διασπορά μεταξύ των κλάσεων, $S_w = S_1 + S_2 = (X_1 - \mu_1)(X_1 - \mu_1)^T + (X_2 - \mu_2)(X_2 - \mu_2)^T$ η διασπορά εντός των κλάσεων, όπου μ_1, μ_2 οι μέσες τιμές των κλάσεων και S_1, S_2 οι πίνακες διασπορών των δύο κλάσεων. Ο σκοπός της LDA είναι ο μετασχηματισμός των δεδομένων ώστε να μεγιστοποιηθεί το κριτήριο Fisher J, κάτι που συμβαίνει όταν μηδενιστεί η παράγωγός του. Μετά τον κατάλληλο μετασχηματισμό μέσω πίνακα V, αποδεικνύεται ότι:

$$J(V) = \frac{V^T S_b V}{V^T S_w V}$$

Ο μηδενισμός της παραγώγου του $J(V)$ οδηγεί στην εξίσωση ιδιοτιμών, που λύνουμε ως προς V: $SV = \lambda V$ με $S = S_w^{-1} S_b$ και $\lambda = J(V)$. Η μέγιστη ιδιοτιμή λ^*, δηλαδή η μέγιστη τιμή του κριτηρίου Fisher J, καθώς και το αντίστοιχο ιδιοδιάνυσμα v^* που ορίζει τη διεύθυνση προβολής για το βέλτιστο διαχωρισμό των κλάσεων, μπορούν να βρεθούν και με απ' ευθείας λύση της εξίσωσης $S_w^{-1} S_b V = \lambda V$, ως εξής:

$$\lambda^* = (\mu_2 - \mu_1)^T S_w^{-1} (\mu_2 - \mu_1), \;\; v^* = \mathrm{c} \cdot S_w^{-1} (\mu_2 - \mu_1) \; \kappa\alpha\iota \; \hat{v}^* = \frac{v^*}{\sqrt{v^{*T} v^*}}$$

Οι κλάσεις μετά το μετασχηματισμό είναι:

$$\begin{cases} Y_1 = V^T X_1 \; (\text{ή } Y_1^* = v^{*T} X_1) \\ Y_2 = V^T X_2 \; (\text{ή } Y_2^* = v^{*T} X_2) \end{cases}$$

Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή LinearDiscriminantAnalysis της υποενότητας discriminant analysis της βιβλιοθήκης scikit-learn. Οι σημαντικότερες παράμετροι, στις οποίες εφαρμόσαμε αναζήτηση πλέγματος για να βρούμε τις κατάλληλες τιμές, είναι:

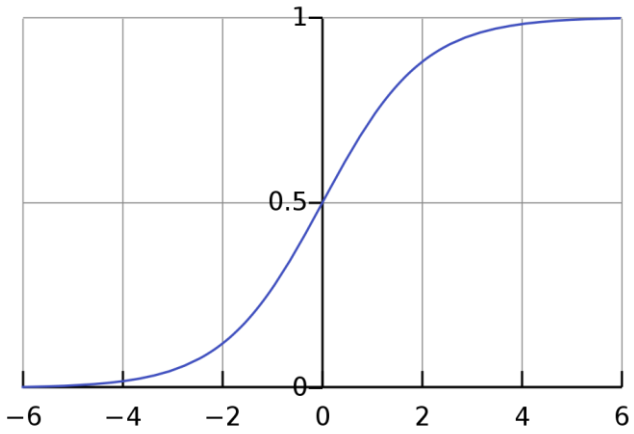
- solver: Οι πιθανές μέθοδοι είναι οι SVD, ελαχίστων τετραγώνων (lsqr) και αποσύνθεσης ιδιοτιμών (eigen). Και στα τρία πειράματα χρησιμοποιήθηκε ο επιλυτής ελαχίστων τετραγώνων.
- shrinkage: Παράμετρος συρρίκνωσης. Οι πιθανές τιμές αυτής της παραμέτρου είναι none, όπου δε γίνεται συρρίκνωση, auto, όπου γίνεται αυτόματα μέσω Ledoit-Wolf και η σταθερή παράμετρος συρρίκνωσης μεταξύ 0 και 1. Οι σταθερές τιμές που δοκιμάστηκαν ήταν [0, 0.2, 0.4, 0.8, 1], μαζί με τις auto και none. Για το πείραμα των καναλιών αριστερού και δεξιού ημισφαιρίου χρησιμοποιήσαμε την τιμή auto, ενώ για τα άλλα δύο η τιμή 0.
- tol: Απόλυτο κατώτατο όριο ώστε μια μοναδική τιμή του X να θεωρείται σημαντική, χρησιμοποιείται για την εκτίμηση της τάξης του X. Οι διαστάσεις των οποίων οι μοναδικές τιμές δεν είναι σημαντικές απορρίπτονται. Αυτή η τιμή είναι σημαντική μόνο αν στην παράμετρο solver έχουμε χρησιμοποιήσει svd. Εδώ δε μας είναι χρήσιμη και κρατήσαμε την προκαθορισμένη.

5.8 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση (Logistic Regression) είναι ένας αλγόριθμος ταξινόμησης Μηχανικής Μάθησης που χρησιμοποιείται για την πρόβλεψη της πιθανότητας μιας κατηγορικής εξαρτημένης μεταβλητής. Στη λογιστική παλινδρόμηση, η εξαρτημένη μεταβλητή είναι μια δυαδική μεταβλητή που περιέχει δεδομένα κωδικοποιημένα ως 1 (ναι, επιτυχία κ.λπ.) ή 0 (όχι, αποτυχία κ.λπ.).

Στη γλώσσα της στατιστικής, η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος προσαρμόζοντας τα δεδομένα τής μελέτης στην εξίσωση της λογιστικής καμπύλης, όπως στο Σχήμα 25.

Η δίτιμη λογιστική παλινδρόμηση έχει τη μορφή: $f(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$, z: $μεταβλητή\ εισόδου$ και η έξοδος της περιορίζεται στο (0,1). Η μεταβλητή z (λογιστική) εκφράζει το μέτρο της ολικής συνεισφοράς όλων των συμμετεχουσών ανεξάρτητων μεταβλητών στο μοντέλο και ορίζεται ως $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, όπου β_0 το ύψος της κλίσης της γραμμής παλινδρόμησης και $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$ οι συντελεστές παλινδρόμησης, καθένας εκ των οποίων εκφράζει το μέγεθος συνεισφοράς τής αντίστοιχης μεταβλητής. Θετική τιμή του συντελεστή δηλώνει ότι η επεξηγηματική μεταβλητή αυξάνει την πιθανότητα της επιτυχημένης έκβασης (να συμβεί δηλαδή το γεγονός), αρνητική τιμή σημαίνει ότι η μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης. Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επηρεάζει πολύ ισχυρά την πιθανότητα να συμβεί το γεγονός ή μη, ενώ χαμηλή τιμή δηλώνει μικρή επίδραση της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.



Σχήμα 25: Πρότυπη λογιστική σιγμοειδής συνάρτηση

Συνοψίζοντας, η λογιστική παλινδρόμηση χρησιμεύει στην περιγραφή της σχέσης που αναπτύσσεται μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών και μιας δυαδικής μεταβλητής απόκρισης εκφρασμένης ως πιθανότητα δυνάμενη να πάρει μία από δύο τιμές, π.χ. επιληπτικό δείγμα (1) – μη επιληπτικό δείγμα (0).

Οι πιθανότητες που συγκλίνουν υπέρ της εμφάνισης ενός γεγονότος εκφράζονται ως λόγος ζεύγους ακέραιων τιμών (odds), όπου ο αριθμητής προσδιορίζει την πιθανότητα που έχει το προσδοκώμενο γεγονός να συμβεί και ο παρονομαστής την πιθανότητα να μη συμβεί. Η σχέση αυτή μπορεί να ενσωματωθεί στο μοντέλο της παλινδρόμησης σε λογαριθμική μορφή ως εξής:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Οι συντελεστές της παλινδρόμησης υπολογίζονται με τη βοήθεια της εκτίμησης της μέγιστης πιθανοφάνειας (Maximum Likelihood Estimate – MLE), ως:

$L = \prod_{i=1}^{k} f(x_i\theta)$ ή $L = \prod_{i=1}^{k} \ln(f(x_i\theta))$, όπου θ: μια παράμετρος της μεταβλητής η οποία μπορεί να μεταβάλλεται ελεύθερα. Η προβλεπόμενη τιμή για κάθε παρατήρηση θα ισούται με $\hat{l} = \frac{\ln(l)}{k}$.

Η λογιστική παλινδρόμηση κανονικοποιημένη με $l1$ λύνει το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{\beta,c} \|\beta\|_1 + C \sum_{i=1}^{k} \log(e^{-y_i(X_i^T\beta + c)} + 1)$$

Για την υλοποίηση χρησιμοποιήσαμε τον ταξινομητή LogisticRegression της υποενότητας linear_model της βιβλιοθήκης scikit-learn. Οι σημαντικότερες παράμετροι, στις οποίες εφαρμόσαμε αναζήτηση πλέγματος για να βρούμε τις κατάλληλες τιμές, είναι:

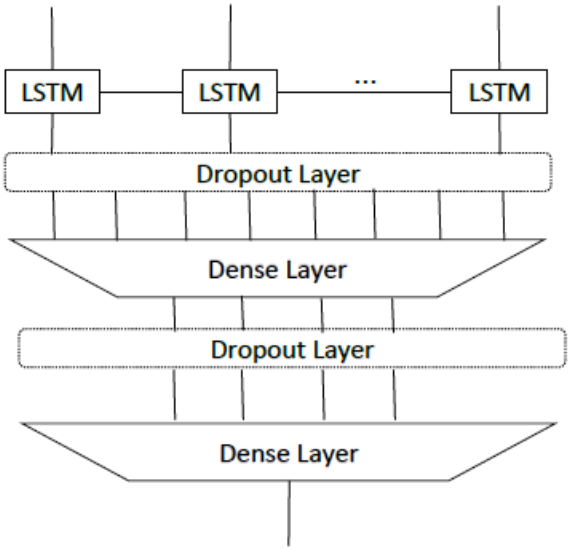
- penalty: Χρησιμοποιείται για τον καθορισμό της νόρμας που χρησιμοποιείται στην ποινή. Πιθανές τιμές οι l1, l2, elasticnet και none. Οι μέθοδοι επίλυσης newton-cg, sag και lbfgs υποστηρίζουν μόνο l2 ποινές. Το elasticnet υποστηρίζεται μόνο από τη μέθοδο saga. Εάν επιλεγεί το “none” (δεν υποστηρίζεται από τη μέθοδο liblinear), δεν εφαρμόζεται κανονικοποίηση. Για όλα τα πειράματα βέλτιστη νόρμα ήταν η l2.
- C: Το αντίστροφο της ισχύος κανονικοποίησης. Πρέπει να είναι θετικό. Οι μικρότερες τιμές του C προσδιορίζουν ισχυρότερη κανονικοποίηση. Οι τιμές που δοκιμάστηκαν ήταν [0.01, 0.1, 1, 10] και, για όλα τα πειράματα, βέλτιστα αποτελέσματα προέκυψαν για C=0.1.
- max_iter: Μέγιστος αριθμός επαναλήψεων για τη σύγκλιση των επιλυτών. Οι τιμές που δοκιμάστηκαν ήταν [200, 400, 600, 800, 1000]. Για όλα τα πειράματα προέκυψαν για 200 επαναλήψεις.
- solver: Ο αλγόριθμος που χρησιμοποιήθηκε στο πρόβλημα βελτιστοποίησης. Δοκιμάστηκαν οι lbfgs και liblinear και χρησιμοποιήθηκε τελικά ο lbfgs.

5.9 Νευρωνικά δίκτυα με LSTM

Μια από τις πιο διαδεδομένες τεχνικές μηχανικής μάθησης είναι τα νευρωνικά δίκτυα. Τα Τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks - ANN) αποτελούν μια αρχιτεκτονική δομή σχεδιασμένη ώστε να μοντελοποιεί τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος εκτελεί μια συγκεκριμένη λειτουργία. Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από κόμβους (νευρώνες) για να κάνουν τις απαραίτητες συνδέσεις (συνάψεις) και επεξεργασίες, ώστε να αναγνωρίσουν ένα συγκεκριμένο πρότυπο, όπως κάνουν και οι βιολογικοί νευρώνες. Προσλαμβάνουν γνώση από το περιβάλλον και μαθαίνουν μέσα από την εμπειρία, όπως και ο άνθρωπος. Τα νευρωνικά δίκτυα αποτελούν ισχυρά εργαλεία για τη διαδικασία μοντελοποίησης, ειδικά σε περιπτώσεις που η σχέση μεταξύ των υποκείμενων δεδομένων δεν είναι γνωστή, και έχουν την ικανότητα να αναγνωρίζουν και να θυμούνται συσχετισμένα πρότυπα ανάμεσα σε ένα σύνολο δεδομένων εισόδου και συγκεκριμένες αντίστοιχες τιμές. Μετά τη διαδικασία της εκπαίδευσης, τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για να προβλέψουν το αποτέλεσμα ενός νέου ανεξάρτητου συνόλου δεδομένων.

Η ιδέα για το σχεδιασμό του νευρωνικού δικτύου προέρχεται από το [3], όπου οι συγγραφείς χρησιμοποιούν το ίδιο σύνολο δεδομένων με την παρούσα εργασία και εφαρμόζουν ένα πείραμα πρόβλεψης της επιληπτικής κρίσης. Το δίκτυο μας, που παρουσιάζεται στο Σχήμα 26, είναι ελαφρώς τροποποιημένο σε σχέση με το

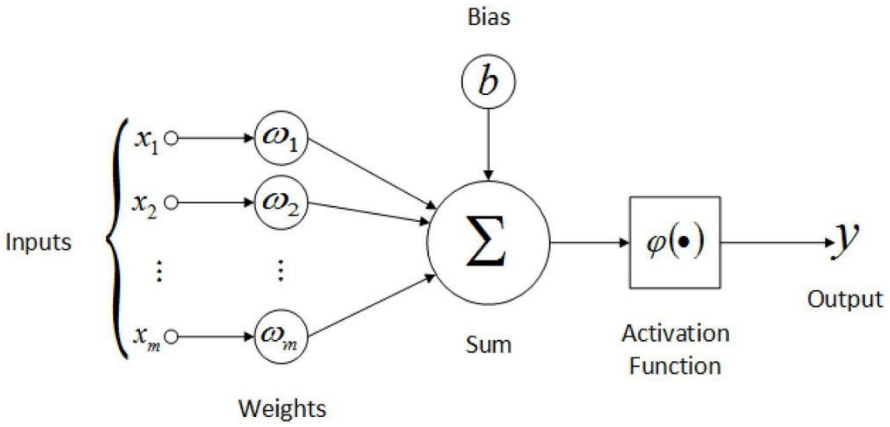
προτεινόμενο και προσαρμοσμένο στο δικό μας πείραμα που είναι η ανίχνευση τής επιληπτικής κρίσης, όχι η πρόβλεψη αυτής. Πρόκειται για ένα νευρωνικό δίκτυο με ένα επίπεδο εισόδου και δύο πυκνά επίπεδα (dense layers). Μεταξύ των τριών αυτών επιπέδων παρεμβάλλονται δύο dropout επίπεδα για την αποφυγή της υπερπροσαρμογής του δικτύου στα δεδομένα. Επειδή ο τομέας των νευρωνικών δικτύων είναι αρκετά ευρύς και τα νευρωνικά δίκτυα δεν είναι το κύριο κομμάτι ενασχόλησης τής εργασίας, θα αναφερθούμε μόνο στις έννοιες που κρίνονται απαραίτητες για την κατανόηση του δικτύου του σχήματος 26, αποφεύγοντας την εμβάθυνση.



Σχήμα 26: Νευρωνικό δίκτυο με LSTM

Σε αυτό το σημείο θα σταθούμε για λίγο στην εσωτερική δομή ενός ANN. Ο κάθε νευρώνας του ANN είναι μια σύναψη των εισόδων του. Όταν μιλάμε για σύναψη αναφερόμαστε στην άθροιση των γινομένων των x_i εισόδων, με $i = 1, \dots, m$ με συντελεστές βάρους w_i. Ένας τυπικός τεχνητός νευρώνας φαίνεται στο Σχήμα 27. Τα τρία βασικά στοιχεία του είναι:

1. Ένα σύνολο συνάψεων με τα αντίστοιχα βάρη. Ένα σήμα εισόδου x_i στην είσοδο της σύναψης i που συνδέεται με το νευρώνα k, πολλαπλασιάζεται επί το συναπτικό βάρος w_i.
2. Έναν αθροιστή (adder) για την άθροιση των σημάτων εισόδου, σταθμισμένων από τα αντίστοιχα συναπτικά βάρη του νευρώνα. Ουσιαστικά πρόκειται για έναν γραμμικό συνδυαστή (linear combiner).
3. Μια συνάρτηση ενεργοποίησης (activation function) για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα.



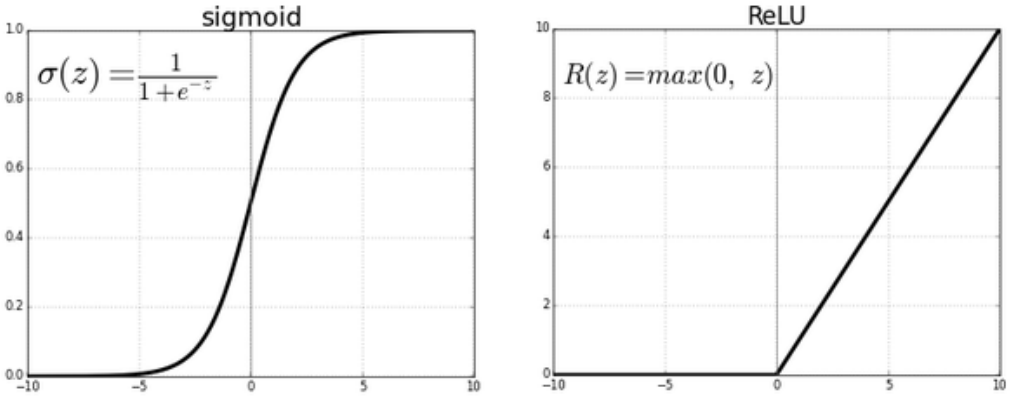
Σχήμα 27: Τυπική μορφή νευρώνα ANN

Οι είσοδοι μπορεί να είναι κάποια εξωτερικά σήματα ή οι έξοδοι από άλλους νευρώνες. Επίσης, πρέπει να αναφέρουμε την ύπαρξη μιας επιπλέον εισόδου b_k, που είναι γνωστή ως προκατάληψη (bias), και χρησιμοποιείται προκειμένου να αυξήσει ή να μειώσει το αποτέλεσμα της σύναψης ενός νευρώνα ανάλογα με το αν αυτό είναι θετικό ή αρνητικό. Στη συνέχεια, το άθροισμα των εισόδων με τα αντίστοιχα βάρη μετασχηματίζεται από μια συνάρτηση ενεργοποίησης φ και προκύπτει η έξοδος σύμφωνα με τη σχέση:

$$y_k = \varphi\left(\sum_{i=1}^{m} w_i x_i + b_k\right)$$

Υπάρχουν διάφορες συναρτήσεις ενεργοποίησης. Αυτές που χρησιμοποιήσαμε εμείς ήταν:

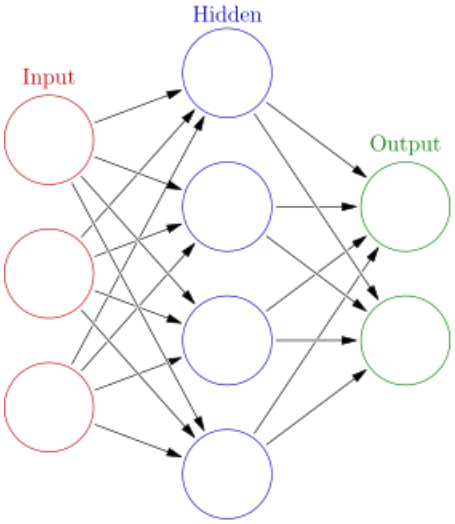
- ReLU: Η συνάρτηση ενεργοποίησης ReLU, η γραφική παράσταση της οποίας φαίνεται στο Σχήμα 28, υπολογίζεται από τον τύπο $\varphi(x) = \max(0, x)$.
- Sigmoid: Η σιγμοειδής συνάρτηση ενεργοποίησης, η γραφική παράσταση της οποίας φαίνεται στο Σχήμα 28, υπολογίζεται από τον τύπο $\varphi(x) = \frac{1}{1+e^{-x}}$.



Σχήμα 28: Συναρτήσεις ενεργοποίησης Sigmoid και ReLU

Οι νευρώνες σε ένα ANN είναι οργανωμένοι σε επίπεδα (layers). Τα τρία βασικά επίπεδα των νευρωνικών δικτύων παρατίθενται παρακάτω: Τα εξωτερικά σήματα εφαρμόζονται στους νευρώνες του επιπέδου εισόδου (input layer). Οι έξοδοι των νευρώνων του επιπέδου εισόδου μεταφέρουν τις πληροφορίες τους στους νευρώνες των ενδιάμεσων ή κρυμμένων επιπέδων (hidden layers), οι οποίοι δεν έχουν άμεση σχέση με το

περιβάλλον. Τέλος, οι νευρώνες του επιπέδου εξόδου (output layer) ενημερώνουν το χρήστη για την έξοδο του νευρωνικού δικτύου.



Σχήμα 29: Δομή πλήρως συνδεδεμένου νευρωνικού δικτύου ενός επιπέδου

Ως προς τον τρόπο σύνδεσης των νευρώνων τα νευρωνικά δίκτυα κατατάσσονται σε δύο μεγάλες κατηγορίες: τα δίκτυα εμπρόσθιας τροφοδότησης (feed-forward) και τα δίκτυα ανατροφοδότησης (Recurrent Neural Network - RNN). Στα δίκτυα εμπρόσθιας τροφοδότησης τα σήματα μεταφέρονται προς μία μόνο κατεύθυνση, από το επίπεδο εισόδου προς το επίπεδο εξόδου, και η έξοδος ενός νευρώνα δεν μπορεί να αποτελεί είσοδο σε νευρώνα του ίδιου ή προηγούμενου επιπέδου. Αντίθετα, στην περίπτωση των ANN ανατροφοδότησης, οι αναδράσεις επιτρέπονται και, συνεπώς, τα σήματα μπορούν να μεταφέρονται και προς τις δύο κατευθύνσεις. Το ιδιαίτερο χαρακτηριστικό τους είναι ότι έχουν δυναμική συμπεριφορά, δηλαδή η κατάσταση τους αλλάζει συνεχόμενα μέχρι να φτάσει ένα σημείο ισορροπίας.

Ας εστιάσουμε στα δίκτυα ανατροφοδότησης. Μια σύνδεση ανατροφοδότησης συνδέει έναν νευρώνα ενός επιπέδου είτε με ένα προηγούμενο επίπεδο είτε με τον ίδιο τον νευρώνα ή με νευρώνα του ίδιου επιπέδου. Οι περισσότερες αρχιτεκτονικές ανατροφοδοτούμενων νευρωνικών δικτύων διατηρούν κατάσταση στις συνδέσεις ανατροφοδότησης, αντίθετα με τα δίκτυα εμπρόσθιας τροφοδότησης. Η κατάσταση ενός ανατροφοδοτούμενου νευρωνικού δικτύου λειτουργεί ως ένα είδος βραχυπρόθεσμης μνήμης για το νευρωνικό δίκτυο και, κατά συνέπεια, δεν θα παράγει πάντα την ίδια έξοδο για μια δεδομένη είσοδο. Οι ανατροφοδοτούμενες συνδέσεις δεν μπορούν ποτέ να στοχεύσουν τους νευρώνες εισόδου ή τους νευρώνες προκατάληψης.

Κατά τη διάρκεια της εκπαίδευσης του RNN, καθώς οι πληροφορίες περνούν ξανά και ξανά, προκαλούνται πολύ μεγάλες ενημερώσεις στα βάρη του νευρικού δικτύου που οδηγούν σε ένα ασταθές δίκτυο. Τα LSTM (Long Short Term Memory) δίκτυα είναι μια παραλλαγή του RNN μοντέλου που λύνουν αυτό το πρόβλημα, γιατί χρησιμοποιούν πύλες για τον έλεγχο της διαδικασίας απομνημόνευσης. Παράλληλα, τα LSTM, όπως μαρτυρά και το όνομά τους, είναι ικανά να μαθαίνουν μακροχρόνιες εξαρτήσεις, κάτι που στην πράξη έχει αποδειχθεί ότι δυσκολεύει τα RNN. Μια μονάδα LSTM, όπως μπορούμε να δούμε και στο Σχήμα 30, αποτελείται από τρεις πύλες:

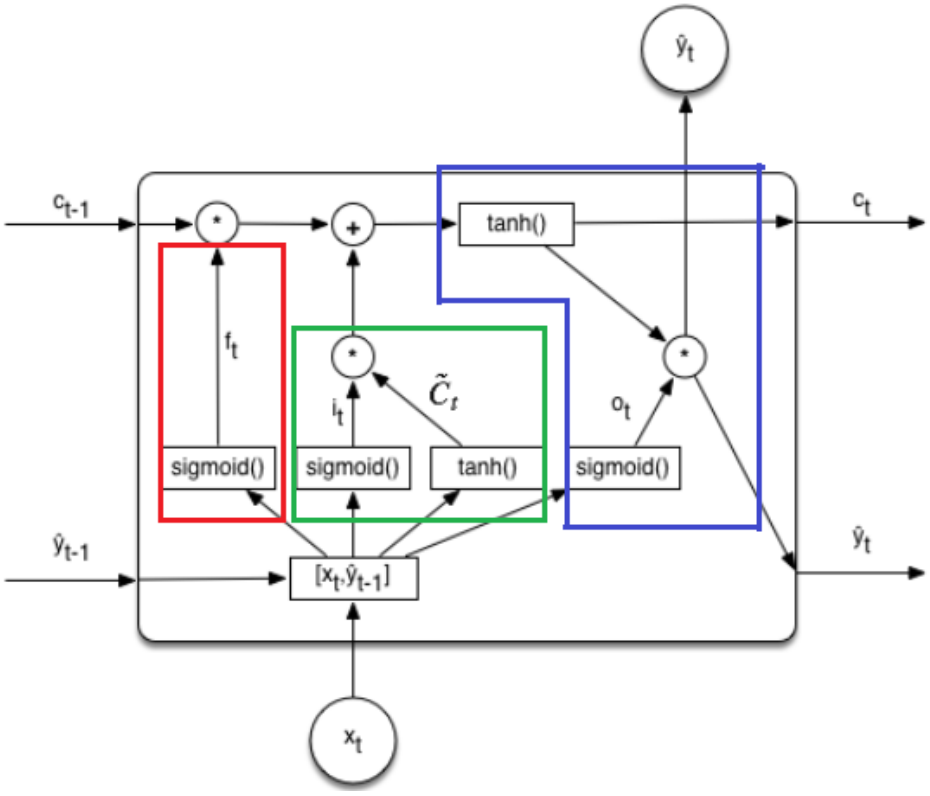
- *Forget Gate (f_t):* Ελέγχει πότε και τι πρέπει να ξεχαστεί.
- *Input Gate (i_t):* Ελέγχει πότε και τι πρέπει να απομνημονευθεί.

- *Output Gate (o_t):* Ελέγχει πότε μια τιμή που απομνημονεύθηκε επιτρέπεται να περάσει από τη μονάδα.

Στα επιμέρους κομμάτια του σχήματος 30, ορίζουμε ως x_t την τρέχουσα είσοδο, με c_{t-1} και c_t τη μνήμη από την προηγούμενη και την τρέχουσα LSTM μονάδα αντίστοιχα και με $\hat{y}_{t-1}$ και $\hat{y}_t$ την έξοδο από την προηγούμενη και την τρέχουσα LSTM μονάδα αντίστοιχα. Κάθε γραμμή φέρει ένα ολόκληρο διάνυσμα, από την έξοδο ενός κόμβου έως τις εισόδους άλλων. Με το + συμβολίζεται η άθροιση πληροφορίας ενώ με * η κλιμάκωση πληροφορίας. Μαθηματικά, αρχικά υπολογίζουμε τη forget gate. Αυτή η πύλη καθορίζει εάν η μονάδα LSTM πρέπει να ξεχάσει τη βραχυπρόθεσμη μνήμη της. Η τιμή b είναι η προκατάληψη (bias), όπως είδαμε και στο νευρώνα του σχήματος 27, μόνο που το LSTM έχει προκατάληψη για κάθε μία από τις 3 πύλες, τις b_f, b_i και b_O αντίστοιχα. Οι ακόλουθες εξισώσεις θα μας βοηθήσουν να κατανοήσουμε πως λειτουργεί εσωτερικά μια LSTM μονάδα:

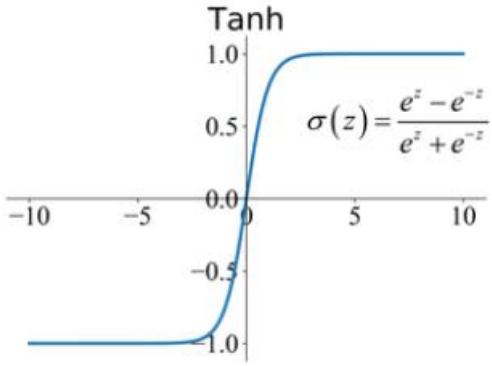
- $f_t = S\left(W_f \cdot [\hat{y}_{t-1}, x_t] + b_f\right)$
- $i_t = S(W_i \cdot [\hat{y}_{t-1}, x_t] + b_i)$
- $\tilde{C}_t = tanh(W_C \cdot [\hat{y}_{t-1}, x_t] + b_C)$
- $C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$
- $o_t = S(W_o \cdot [\hat{y}_{t-1}, x_t] + b_o)$
- $\hat{y}_t = o_t \cdot \tanh(C_t)$

Με S συμβολίσαμε τη σιγμοειδή συνάρτηση που είδαμε και στο Σχήμα 28, ενώ με tanh την υπερβολική εφαπτομένη, της οποίας η γραφική παράσταση φαίνεται στο Σχήμα 31, που υπολογίζεται ως $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.



Σχήμα 30: Παράδειγμα LSTM μονάδας

Το σιγμοειδές επίπεδο της forget πύλης παίρνει την είσοδο x_t και την προηγούμενη έξοδο $\hat{y}_{t-1}$ και αποφασίζει ποια μέρη από την παλιά έξοδο πρέπει να αφαιρεθούν, παράγοντας έξοδο μηδέν για αυτά (Σχήμα 30 – κόκκινο πλαίσιο). Το σιγμοειδές επίπεδο της input πύλης αποφασίζει ποιες από τις νέες πληροφορίες πρέπει να ενημερωθούν ή να αγνοηθούν. Ένα επίπεδο tanh δημιουργεί ένα διάνυσμα όλων των πιθανών τιμών από τη νέα είσοδο. Αυτά τα δύο πολλαπλασιάζονται για να ενημερώσουν το νέο κελί. Αυτή η νέα μνήμη προστίθεται στη συνέχεια στην παλιά μνήμη C_{t-1} για να δώσει C_t (Σχήμα 30 – πράσινο πλαίσιο). Το σιγμοειδές επίπεδο της output πύλης αποφασίζει ποια μέρη της κατάστασης της μονάδας πρόκειται να δοθεί ως έξοδος. Στη συνέχεια, περνάμε την κατάσταση μέσα από ένα tanh επίπεδο που δημιουργεί όλες τις πιθανές τιμές και πολλαπλασιάζουμε με την έξοδο της σιγμοειδούς πύλης, έτσι ώστε να εξάγουμε μόνο τα μέρη που αποφασίσαμε (Σχήμα 30 – μπλε πλαίσιο).



Σχήμα 31: Υπερβολική εφαπτομένη (tanh)

Αφού περιγράψαμε το LSTM μοντέλο, ας αναφερθούμε στα άλλα δύο ξεχωριστά επίπεδα που χρησιμοποιούμε, το πυκνό (dense) επίπεδο και το dropout επίπεδο. Το μεν πρώτο είναι το κλασικό πυκνά συνδεδεμένο επίπεδο, όπως το κρυφό επίπεδο του σχήματος 29. Πυκνά συνδεδεμένο σημαίνει ότι όλοι οι νευρώνες τού προηγούμενου επιπέδου μεταφέρουν την έξοδό τους ως είσοδο σε όλους τους κόμβους του πυκνού επιπέδου και οι έξοδοι των νευρώνων αυτού του επιπέδου είναι συνδεδεμένες με όλους τους νευρώνες του επόμενου επιπέδου. Το πυκνό επίπεδο εκτελεί την ακόλουθη πράξη output = activation(dot(input, kernel) + bias), όπου kernel τα δεδομένα βάρους. Στο πρώτο πυκνό επίπεδο χρησιμοποιήσαμε τη συνάρτηση ενεργοποίησης ReLU και στο δεύτερο τη σιγμοειδή (Σχήμα 28). Το δε dropout επίπεδο χρησιμοποιείται για την αποφυγή του προβλήματος τής υπερπροσαρμογής στα δεδομένα, αφαιρώντας ένα ποσοστό των εισόδων. Στο πρώτο από τα δύο dropout επίπεδα ορίσαμε το ποσοστό στο 0.2 και στο δεύτερο στο 0.1.

Για τη διαμόρφωση της διαδικασίας μάθησης του μοντέλου χρησιμοποιήθηκε συνάρτηση κόστους μέσου τετραγωνικού σφάλματος (mean squared error) για την εύρεση σφάλματος ή απόκλισης στη μαθησιακή διαδικασία, βελτιστοποιητής (optimizer) βαρών εισόδου Adam και ακρίβεια ως μετρική για την αξιολόγηση τού μοντέλου. Η εκπαίδευση του μοντέλου γίνεται σε ομάδες (batch) των 10 δειγμάτων και για 100 επαναλήψεις (εποχές). Μια επανάληψη ή εποχή (epoch) είναι η περίοδος με την οποία ανανεώνονται τα βάρη. Στην πραγματικότητα δεν είναι απαραίτητο ότι θα πραγματοποιηθούν όλες οι επαναλήψεις, αφού γίνεται χρήση της τεχνικής Early Stopping, που μας επιτρέπει να παρακολουθούμε μια μετρική και να σταματάμε την εκπαίδευση όταν αυτή σταματήσει να βελτιώνεται. Έτσι, αφού στόχος μας στην εκπαίδευση είναι η ελαχιστοποίηση τής

απώλειας, αν αυτή δε βελτιωθεί σημαντικά μετά από 5 επαναλήψεις, επαναφέρουμε τα βέλτιστα βάρη και σταματάμε την εκπαίδευση.

Στο config αρχείο έχουμε ορίσει κάποιες παραμέτρους του νευρωνικού δικτύου τού σχήματος 26. Οι τιμές αυτές επιτρέπουν τον πειραματισμό τού χρήστη με το δίκτυο αν και έχει προηγηθεί μια διαδικασία επιλογής των βέλτιστων παραμέτρων για το δίκτυο και τα δεδομένα μας. Οι παράμετροι που είναι διαθέσιμες στο config αρχείο:

- batch: Οι ομάδες δειγμάτων που χρησιμοποιούνται για την εκπαίδευση. Προεπιλεγμένη τιμή 10.
- epochs: Οι επαναλήψεις ενημέρωσης των τιμών των βαρών του νευρωνικού δικτύου. Προεπιλεγμένη τιμή 100.
- dropout_percentage: Το ποσοστό μείωσης των εισόδων στα dropout επίπεδα. Συγκεκριμένα στο πρώτο επίπεδο το ποσοστό ισούται με $\mathrm{dropout\_percentage}$, ενώ στο δεύτερο με $\mathrm{dropout\_percentage}/2$.
- loss_function: Συνάρτηση κόστους της διαδικασία εκμάθησης. Προεπιλεγμένη τιμή mean_squared_error (μέσο τετραγωνικό σφάλμα).
- metric: Η μετρική που θέλουμε να βελτιστοποιήσουμε στη διαδικασία εκμάθησης. Προεπιλεγμένη τιμή accuracy (ακρίβεια).

6. ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΣΗΣ

Αφού περιγράψαμε στο προηγούμενο κεφάλαιο τους αλγορίθμους ταξινόμησης, είναι σημαντικό να παρέχουμε στο χρήστη μια ενδεικτική εικόνα για την αξιοπιστία του μοντέλου ή ακόμη και σε μας, ώστε να μπορέσουμε να βελτιώσουμε το σύστημα. Υπάρχουν δεκάδες μετρικές αξιολόγησης μιας ταξινόμησης, αλλά εμείς επιλέξαμε τις δημοφιλέστερες με βάση τη βιβλιογραφία μας. Στα πλαίσια της υλοποίησης αξιοποιήθηκαν οι μέθοδοι accuracy_score, cohen_kappa_score, confusion_matrix, f1_score και matthews_corrcoef της υποενότητας metrics της βιβλιοθήκης scikit-learn. Στις περιπτώσεις της διασταυρούμενης επικύρωσης υπολογίζουμε τις μετρικές ανά επανάληψη και δίνουμε ως τελικό αποτέλεσμα το μέσο όρο τους.

6.1 Πίνακας Σύγχυσης

Ο πίνακα σύγχυσης (confusion matrix) είναι ένας πίνακας με γραμμές και στήλες τις κλάσεις τού εκάστοτε προβλήματος ταξινόμησης. Κάθε στήλη του δείχνει σε ποια κλάση προέβλεψε ο αλγόριθμος να ανήκουν τα στοιχεία μιας κλάσης, ενώ κάθε γραμμή δείχνει σε ποια κλάση όντως ανήκαν τα στοιχεία που αλγόριθμος προέβλεψε ότι ανήκουν σε μια κλάση. Στην περίπτωση που έχουμε δύο κατηγορίες, εν προκειμένω την επιληπτική και τη μη επιληπτική, στη στήλη απόφασης μπορούμε να χωρίσουμε τα αποτελέσματα σε τέσσερις κατηγορίες, όπως φαίνονται στον Πίνακα 5.

Πίνακας 5: Πίνακας Σύγχυσης

		Predicted Class	
		Seizure	Non-Seizure
True Class	Seizure	TP (True Positives)	FN (False Negatives)
	Non-Seizure	FP (False Positives)	TN (True Negatives)

Οι μεταβλητές που περιέχει ο πίνακας σύγχυσης περιγράφουν τα εξής:

- *TP (True Positive)*: το πλήθος των στοιχείων τα οποία ο ταξινομητής προβλέπει σωστά ως επιληπτικά.
- *TN (True Negative)*: το πλήθος των στοιχείων τα οποία ο ταξινομητής προβλέπει σωστά ως μη επιληπτικά.
- *FP (False Positive)*: το πλήθος των στοιχείων τα οποία ο ταξινομητής προβλέπει ως επιληπτικά, αλλά στην πραγματικότητα ανήκουν στη μη επιληπτική κατηγορία.
- *FN (False Negative)*: το πλήθος των στοιχείων τα οποία ο ταξινομητής προβλέπει ως μη επιληπτικά, αλλά στην πραγματικότητα ανήκουν στην επιληπτική κατηγορία.

Με βάση αυτές τις μεταβλητές που προκύπτουν από τον πίνακα σύγχυσης θα υπολογίσουμε τις μετρικές μας στη συνέχεια.

6.2 Ακρίβεια

Η ακρίβεια ορίζεται ως το ποσοστό των παρατηρήσεων που ταξινομήθηκαν σωστά ως προς το σύνολο των παρατηρήσεων. Οι πιθανές τιμές που μπορεί να πάρει η ακρίβεια ανήκουν στο διάστημα $[0,1]$.

$$ACC = \frac{\#correct\ predictions}{\#samples} = \frac{TP + TN}{TP + TN + FP + FN}$$

Η ακρίβεια, όμως, δεν είναι πάντα μια αξιόπιστη μετρική για την πραγματική απόδοση ενός ταξινομητή. Για παράδειγμα σε ένα πρόβλημα ταξινόμησης, όπου έχουμε 95 στιγμιότυπα στη μία κλάση και 5 στην άλλη, αν ταξινομήσουμε όλα τα δείγματα στην πρώτη κλάση, έχουμε ακρίβεια 95%. Παρόλα αυτά έχουμε ένα πολύ κακό μοντέλο καθώς το μοντέλο έχει 100% επιτυχία στη μία κλάση, αλλά 0% στην άλλη. Βέβαια, στα πειράματά μας έχει προβλεφθεί η αντιμετώπιση αυτού του προβλήματος με την εξισορρόπηση των κλάσεων, αλλά, εν γένει, για την αξιολόγηση και σύγκριση των μοντέλων χρειάζονται και άλλες μετρικές.

6.3 Ευαισθησία/Ανάκληση

Η ευαισθησία ή θετική ανάκληση μετράει την αναλογία των θετικών αποτελεσμάτων που ανιχνεύει ο ταξινομητής, τα οποία είναι αληθώς θετικά, από το σύνολο των θετικών δειγμάτων. Με απλά λόγια είναι η πιθανότητα σωστής ταυτοποίησης μιας επιληπτικής κρίσης και, ως πιθανότητα, οι δυνατές τιμές που μπορεί να πάρει ανήκουν στο διάστημα $[0,1]$.

$$SNV = \frac{TP}{TP + FN}$$

6.4 Ειδικότητα

Η ειδικότητα μετράει την αναλογία των αρνητικών αποτελεσμάτων που ανιχνεύει ο ταξινομητής, τα οποία είναι αληθώς αρνητικά, από το σύνολο των αρνητικών δειγμάτων. Ουσιαστικά είναι ο ρυθμός σωστής ταξινόμησης μη επιληπτικών καταστάσεων ή υγιών δειγμάτων στο σύνολο των μη επιληπτικών. Οι πιθανές τιμές που μπορεί να πάρει η ειδικότητα είναι εντός του διαστήματος $[0,1]$.

$$SPC = \frac{TN}{TN + FP}$$

6.5 Αξιοπιστία

Η αξιοπιστία είναι το ποσοστό των παρατηρήσεων που ταξινομήθηκαν στην επιληπτική κλάση και ανήκουν σε αυτήν προς το ποσοστό των παρατηρήσεων που ταξινομήθηκαν συνολικά σε αυτή την κλάση. Οι πιθανές τιμές που μπορεί να πάρει η αξιοπιστία είναι όποιες και οι τρεις προηγούμενες μετρικές που είδαμε.

$$PPV = \frac{TP}{TP + FP}$$

6.6 Βαθμολογία F1

Η βαθμολογία F1 μπορεί να ερμηνευθεί ως σταθμισμένος μέσος όρος της αξιοπιστίας και της ανάκλησης. Η βαθμολογία F1 φτάνει την καλύτερη τιμή της στο 1 και τη χειρότερη στο

0. Η σχετική συνεισφορά της αξιοπιστίας και της ανάκλησης στη βαθμολογία F1 είναι ίση. Ουσιαστικά είναι ο αρμονικός μέσος μεταξύ αξιοπιστίας και ανάκλησης και δηλώνει πόσο ακριβής και στιβαρός είναι ο ταξινομητής.

$$F1 = \frac{2 \cdot PPV \cdot SNV}{PPV + SNV} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

6.7 Συντελεστής Συσχέτισης Matthews

Ο συντελεστής συσχέτισης Matthew's (Matthew's Correlation Coefficient, MCC) χρησιμοποιείται στη μηχανική μάθηση ως μέτρο αξιολόγησης της δυαδικής ταξινόμησης. Ο συντελεστής λαμβάνει υπόψη αληθείς και ψευδείς, θετικές και αρνητικές μετρήσεις, ενώ γενικά θεωρείται ως ένα ισορροπημένο μέτρο ταξινόμησης, ακόμη και αν οι κατηγορίες αποτελούνται από πολύ διαφορετικά μεγέθη. Ο MCC είναι ισοδύναμος του γνωστού συντελεστή συσχέτισης του Pearson, όταν εφαρμοστεί σε δίτιμα δεδομένα, και λαμβάνει τιμές συσχέτισης στο $[-1,1]$, όπου συντελεστής ίσος με 1 αποτελεί μια τέλεια πρόβλεψη, ίσος με 0 κατά μέσο όρο τυχαία πρόβλεψη και ίσος με -1 αντιπροσωπεύει αντίστροφη πρόβλεψη. Το μεγάλο πλεονέκτημα του συντελεστή συσχέτισης είναι ότι συνδυάζει όλες τις τιμές του πίνακα σύγχυσης σε μία αριθμητική τιμή.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

6.8 Συντελεστής κ του Cohen

Ο συντελεστής κ του Cohen είναι ένα στατιστικό μέτρο που χρησιμοποιείται για τον έλεγχο συμφωνίας μεταξύ δυο κατηγορικών μεταβλητών και παίρνει τιμές στο $[-1,1]$. Ο συντελεστής, όμως, είναι σχεδιασμένος με τέτοιο τρόπο ώστε να λαμβάνει υπόψη και τις περιπτώσεις όπου η συμφωνία μεταξύ των μεταβλητών επιτυγχάνεται κατά τύχη. Η εξίσωση υπολογισμού του κ είναι η ακόλουθη:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Στον παραπάνω τύπο, p_0 είναι η σχετική παρατηρούμενη συμφωνία μεταξύ των κατηγορικών μεταβλητών και p_e είναι η υποθετική πιθανότητα τυχαίας συμφωνίας, χρησιμοποιώντας τα δεδομένα των παρατηρήσεων για τον υπολογισμό των πιθανοτήτων της κάθε μεταβλητής. Εάν οι μεταβλητές είναι σε πλήρη συμφωνία, τότε $\kappa = 1$. Αν δεν υπάρχει καμία συμφωνία μεταξύ των μεταβλητών, εκτός αυτής που θα αναμενόταν τυχαία (όπως ορίζεται από το p_e), τότε $\kappa \leq 0$.

Στο πέρασμα των χρόνων διάφορες κατευθυντήριες γραμμές σχετικά με την τιμή τού συντελεστή κ εμφανίστηκαν στη βιβλιογραφία. Αρχικά, ήταν οι Landis και Koch που χαρακτήρισαν τις τιμές <0 ως ένδειξη μη συμφωνίας, τις 0-0.2 ως ασήμαντη συμφωνία, τις 0.21-0.4 ως ισχνή συμφωνία, τις 0.41–0.6 ως μέτρια συμφωνία, τις 0.61–0.8 ως σημαντική συμφωνία και τις 0.81-1 ως σχεδόν τέλεια συμφωνία. Παρόλο αυτά οι κατευθυντήριες αυτές γραμμές ποτέ δεν έγιναν καθολικά αποδεκτές, καθώς οι Landis και Koch δεν είχαν στοιχεία που αποδεικνύουν τους ισχυρισμούς τους και κατέληξαν σε αυτές μετά από προσωπικές εκτιμήσεις και παρατηρήσεις. Έχει σημειωθεί, άλλωστε, πολλές φορές πως οι παραπάνω κατευθυντήριες γραμμές μπορούν να φανούν περισσότερο αποπροσανατολιστικές παρά χρήσιμες σε μια ανάλυση. Τέλος, ανάλογη γραμμή

ακολούθησε και ο Fleiss αργότερα χαρακτηρίζοντας τα κ μεγαλύτερα του 0.75 ως εξαιρετική, τα 0.40-0.75 από μέτρια ως καλή και τα μικρότερα του 0.4 ως αδύναμη συμφωνία αντίστοιχα.

7. ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο κεφάλαιο αυτό θα αναλύσουμε τα τρία πειράματα που πραγματοποιήθηκαν, τον τρόπο που υπολογίστηκαν τα κανάλια στα δύο πειράματα που δε χρησιμοποιήθηκαν όλα και θα παρατεθούν οι μετρήσεις, που είναι η πεμπτουσία όλης αυτής της διαδικασίας. Παράλληλα, θα εξεταστεί το κατά πόσο παρατηρήθηκε αισθητή διαφορά και το κατά πόσο βελτιώθηκαν ή όχι τα αποτελέσματα των πειραμάτων από τη διασταυρούμενη επικύρωση. Τέλος, θα δούμε ποιες μέθοδοι έχουν καλύτερα αποτελέσματα και αν παρουσιάζουν την αναμενόμενη βελτίωση εφόσον λαμβάνουν ως είσοδο δεδομένα από περισσότερα κανάλια.

7.1 Πείραμα μέσου καναλιού

Σε αυτό το πείραμα χρησιμοποιήθηκε μόνο ένα κανάλι, το οποίο προήλθε ως ο μέσος όρος των μετρήσεων των 18 καναλιών τα οποία κρατήσαμε. Πρώτα υπολογίζεται το κανάλι και στη συνέχεια πραγματοποιείται η εξαγωγή χαρακτηριστικών σε αυτό. Η ιδέα της εκτέλεσης αυτού του πειράματος προήλθε από το [11] και ο τύπος υπολογισμού του μέσου καναλιού - αντιπροσώπου είναι ο εξής:

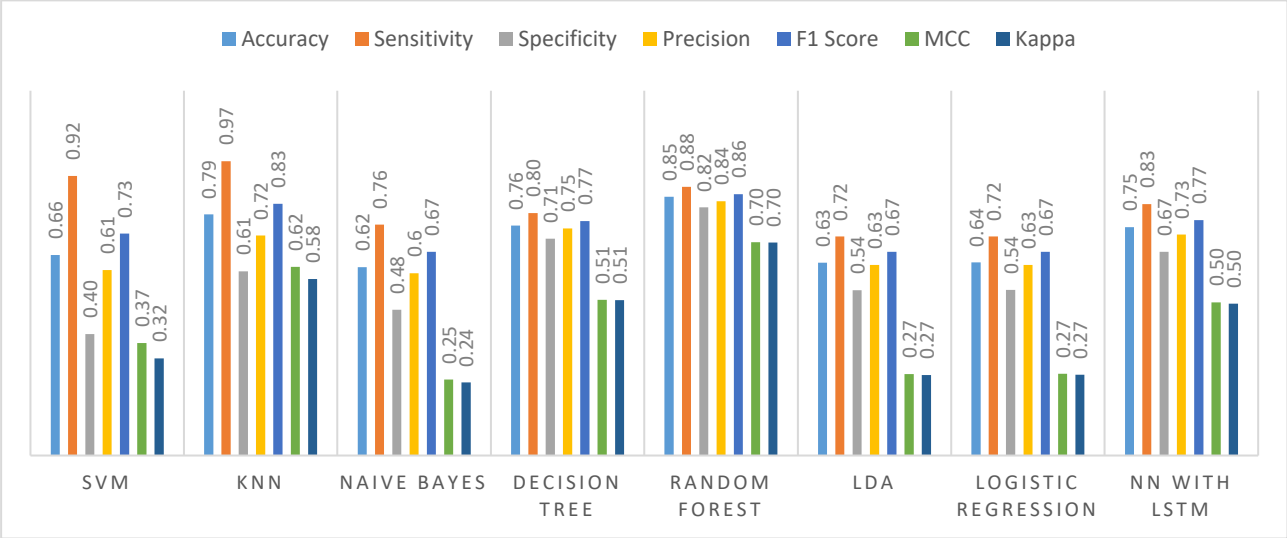
$$Averaged_{EEG} = \frac{1}{c}\sum_{i=1}^{c} x_i$$

όπου c: το πλήθος των καναλιών (εδώ τα 18 κοινά σε όλους τους ασθενείς) και x_i: το διάνυσμα τιμών του i-οστού καναλιού

Στον Πίνακα 6 και το Σχήμα 32 παρατίθενται όλες οι μετρήσεις των μεθόδων ταξινόμησης που εφαρμόστηκαν σε αυτό το πείραμα, αφού πρώτα είχε χωριστεί το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης σε αναλογία 70%-30%.

Πίνακας 6: Μετρικές ταξινόμησης πειράματος μέσου καναλιού με διαχωρισμό συνόλου δεδομένων

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1 SCORE	MCC	KAPPA
SVM	0.66	0.92	0.40	0.61	0.73	0.37	0.32
KNN	0.79	0.97	0.61	0.72	0.83	0.62	0.58
NAIVE BAYES	0.62	0.76	0.48	0.6	0.67	0.25	0.24
DECISION TREES	0.76	0.80	0.71	0.75	0.77	0.51	0.51
RANDOM FOREST	0.85	0.88	0.82	0.84	0.86	0.70	0.70
LDA	0.63	0.72	0.54	0.63	0.67	0.27	0.27
LOGISTIC REGRESSION	0.64	0.72	0.54	0.63	0.67	0.27	0.27
NN WITH LSTM	0.75	0.83	0.67	0.73	0.77	0.50	0.50

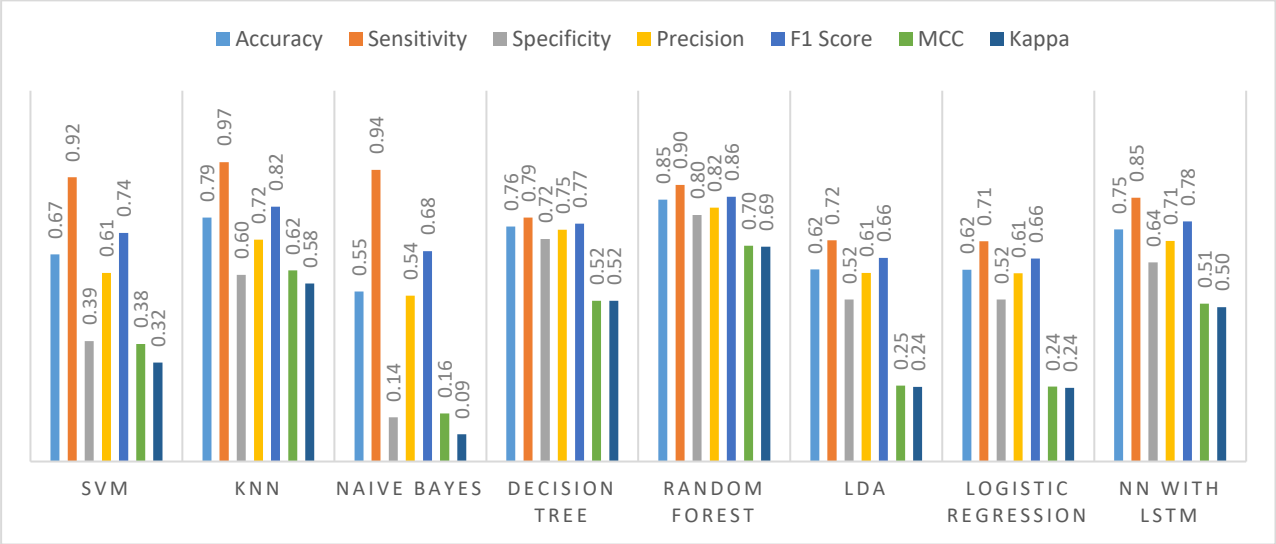


Σχήμα 32: Μετρικές ταξινόμησης πειράματος μέσου καναλιού με διαχωρισμό συνόλου δεδομένων

Στον Πίνακα 7 και το Σχήμα 33 παρατίθενται όλες οι μετρήσεις των μεθόδων ταξινόμησης που εφαρμόστηκαν σε αυτό το πείραμα εφαρμόζοντας διασταυρούμενη επικύρωση 5 επαναλήψεων.

Πίνακας 7: Μετρικές ταξινόμησης πειράματος μέσου καναλιού με διασταυρούμενη επικύρωση

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1 SCORE	MCC	KAPPA
SVM	0.67	0.92	0.39	0.61	0.74	0.38	0.32
KNN	0.79	0.97	0.60	0.72	0.82	0.62	0.58
NAIVE BAYES	0.55	0.94	0.14	0.54	0.68	0.16	0.09
DECISION TREE	0.76	0.79	0.72	0.75	0.77	0.52	0.52
RANDOM FOREST	0.85	0.90	0.80	0.82	0.86	0.70	0.69
LDA	0.62	0.72	0.52	0.61	0.66	0.25	0.24
LOGISTIC REGRESSION	0.62	0.71	0.52	0.61	0.66	0.24	0.24
NN WITH LSTM	0.75	0.85	0.64	0.71	0.78	0.51	0.50



Σχήμα 33: Μετρικές ταξινόμησης πειράματος μέσου καναλιού με διασταυρούμενη επικύρωση

Συγκρίνοντας τις δύο βασικές μετρικές της ακρίβειας και της βαθμολογίας F1 παρατηρούμε ότι οι διαφορές ανάμεσα στη χρήση ή μη της διασταυρούμενης επικύρωσης είναι της τάξης του 1-1,5%. Εξαιρείται ο απλοϊκός Bayes που η χρήση διασταυρούμενης επικύρωσης είχε σημαντικά χειρότερα αποτελέσματα. Παρόλα αυτά καμία από τις δύο περιπτώσεις δεν είναι σταθερά καλύτερη από την άλλη. Συνεπώς, δε μπορεί να ειπωθεί ότι στο πείραμα μέσου καναλιού η διασταυρούμενη επικύρωση είχε ιδιαίτερο αντίκτυπο στα αποτελέσματα.

Σχολιάζοντας τα αποτελέσματα, οι καλύτερες μετρήσεις δίνονται από τη μέθοδο του τυχαίου δάσους που ανιχνεύει σωστά τα επιληπτικά και τα μη επιληπτικά δείγματα σε ποσοστό πάνω του 80%, η ακρίβεια του στο 85% είναι η μεγαλύτερη από όλες τις μεθόδους, ενώ από τις μετρικές MCC και κ του Cohen αντιλαμβανόμαστε ότι το πρόβλημα δυαδικής ταξινόμησης αντιμετωπίζεται ικανοποιητικά από αυτόν τον ταξινομητή. Σε ένα δεύτερο επίπεδο θα τοποθετούσαμε τους ταξινομητές kNN, δέντρου απόφασης και νευρωνικών δικτύων με LSTM που έχουν παρεμφερή αποτελέσματα και δείχνουν να αναγνωρίζουν καλύτερα τα επιληπτικά σε σχέση με τα μη επιληπτικά δείγματα. Στο τελευταίο επίπεδο θα τοποθετούσαμε τους ταξινομητές SVM, απλοϊκό Bayes, LDA και Λογιστικής Παλινδρόμησης που έχουν παρόμοιες μετρήσεις που είναι χαμηλές.

7.2 Πείραμα καναλιών αριστερού και δεξιού ημισφαιρίου

Σύμφωνα με την αναφορά [12], για το συγκεκριμένο σύνολο δεδομένων τα κανάλια (F3-C3) και (C3-P3) και τα κανάλια (F4-C4) και (C4-P4) αντιπροσωπεύουν καλύτερα την πληροφορία από το αριστερό και το δεξιό ημισφαίριο αντίστοιχα, όπως φαίνεται και στην Εικόνα 3. Συγκεκριμένα, στο [12] αναφέρεται ότι η θέση P3 και P4 τού εγκεφάλου είναι πιο επιρρεπής σε επιληπτικές κρίσεις και είναι πάνω από την περιοχή τής εγκεφαλοαγγειακής λεκάνης απορροής που είναι υψηλού κινδύνου εγκεφαλικού τραυματισμού. Ακόμα, σύμφωνα με τις δικές τους πηγές, έχει παρατηρηθεί ότι το κανάλι C3-C4 έδινε την υψηλότερη ακρίβεια στους ταξινομητές. Συνεπώς, επιλέξαμε τα προαναφερθέντα δύο ζευγάρια καναλιών ώστε να φτιάξουμε δύο κανάλια αντιπροσωπευτικά του αριστερού και του δεξιού ημισφαιρίου, τα οποία υπολογίζονται από τους ακόλουθους τύπους:

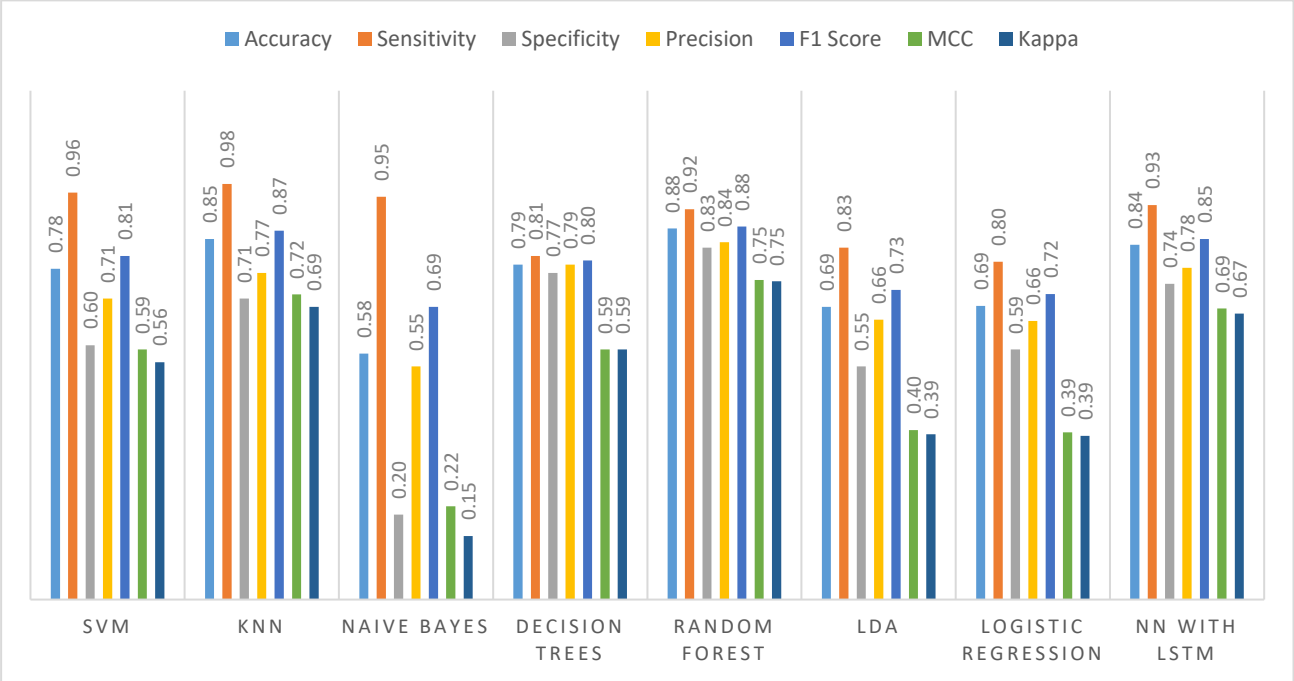
$$\begin{cases} LeftHemispheric_{EEG} = \dfrac{x['F3 - C3'] + x['C3 - P3']}{2} \\ RightHemispheric_{EEG} = \dfrac{x['F4 - C4'] + x['C4 - P4']}{2} \end{cases}$$

όπου $x[i]$: το διάνυσμα τιμών στο κανάλι i και $LeftHemispheric_{EEG}$, $RightHemispheric_{EEG}$: τα διανύσματα τιμών στα δύο κανάλια που δημιουργήσαμε για αυτό το πείραμα.

Στον Πίνακα 8 και το Σχήμα 34 παρατίθενται όλες οι μετρήσεις των μεθόδων ταξινόμησης που εφαρμόστηκαν σε αυτό το πείραμα, αφού πρώτα είχε χωριστεί το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης σε αναλογία 70%-30%.

Πίνακας 8: Μετρικές ταξινόμησης πειράματος καναλιών αριστερού και δεξιού ημισφαιρίου με διαχωρισμό συνόλου δεδομένων

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1 SCORE	MCC	KAPPA
SVM	0.78	0.96	0.60	0.71	0.81	0.59	0.56
KNN	0.85	0.98	0.71	0.77	0.87	0.72	0.69
NAIVE BAYES	0.58	0.95	0.20	0.55	0.69	0.22	0.15
DECISION TREE	0.79	0.81	0.77	0.79	0.80	0.59	0.59
RANDOM FOREST	0.88	0.92	0.83	0.84	0.88	0.75	0.75
LDA	0.69	0.83	0.55	0.66	0.73	0.40	0.39
LOGISTIC REGRESSION	0.69	0.80	0.59	0.66	0.72	0.39	0.39
NN WITH LSTM	0.84	0.93	0.74	0.78	0.85	0.69	0.67

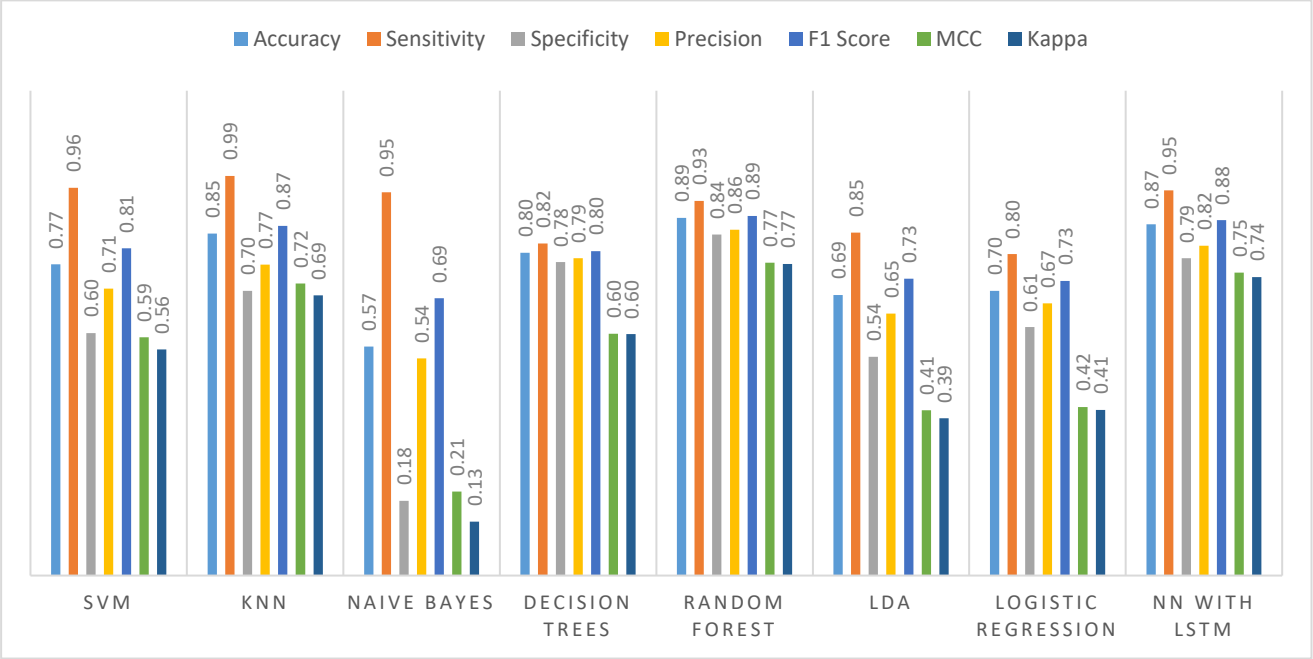


Σχήμα 34: Μετρικές ταξινόμησης πειράματος καναλιών αριστερού και δεξιού ημισφαιρίου με διαχωρισμό συνόλου δεδομένων

Στον Πίνακα 9 και το Σχήμα 35 παρατίθενται όλες οι μετρήσεις των μεθόδων ταξινόμησης που εφαρμόστηκαν σε αυτό το πείραμα εφαρμόζοντας διασταυρούμενη επικύρωση 5 επαναλήψεων.

Πίνακας 9: Μετρικές ταξινόμησης πειράματος καναλιών αριστερού και δεξιού ημισφαιρίου με διασταυρούμενη επικύρωση

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1 SCORE	MCC	KAPPA
SVM	0.77	0.96	0.60	0.71	0.81	0.59	0.56
KNN	0.85	0.99	0.70	0.77	0.87	0.72	0.69
NAIVE BAYES	0.57	0.95	0.18	0.54	0.69	0.21	0.13
DECISION TREE	0.80	0.82	0.78	0.79	0.80	0.60	0.60
RANDOM FOREST	0.89	0.93	0.84	0.86	0.89	0.77	0.77
LDA	0.69	0.85	0.54	0.65	0.73	0.41	0.39
LOGISTIC REGRESSION	0.70	0.80	0.61	0.67	0.73	0.42	0.41
NN WITH LSTM	0.87	0.95	0.79	0.82	0.88	0.75	0.74



Σχήμα 35: Μετρικές ταξινόμησης πειράματος καναλιών αριστερού και δεξιού ημισφαιρίου με διασταυρούμενη επικύρωση

Σχετικά με τη διασταυρούμενη επικύρωση, με εξαίρεση τον απλοϊκό Bayes που παίρνουμε ελαφρώς μικρότερη τιμή σε ακρίβεια, σε όλες τις υπόλοιπες μεθόδους λαμβάνουμε τουλάχιστον τις ίδιες και συχνά και καλύτερες μετρήσεις σε ακρίβεια και βαθμολογία F1 με την εφαρμογή της. Αποκορύφωμα αποτελούν οι μετρήσεις του νευρωνικού δικτύου με LSTM, όπου έχουμε βελτίωση 3-6% σε όλες τις μετρικές ταξινόμησης με χρήση διασταυρούμενης επικύρωσης. Συνεπώς, σε αυτό το πείραμα μπορεί να ειπωθεί ότι η διασταυρούμενη επικύρωση ευνόησε τις μετρήσεις μας.

Όσον αφορά τα αποτελέσματα, και πάλι η ταξινόμηση τυχαίου δάσους δίνει τα καλύτερα αποτελέσματα, αλλά εδώ, στο πρώτο επίπεδο ταξινομητών, μπορούμε να εντάξουμε και τους kNN και νευρωνικό δίκτυο με LSTM, που έχουν εξίσου καλά αποτελέσματα. Σε ένα δεύτερο επίπεδο θα τοποθετούσαμε τον SVM και τον ταξινομητή δέντρου απόφασης που έχουν ακρίβεια 5-9% μικρότερη από τους αλγόριθμους τού πρώτου επιπέδου. Προφανώς το δέντρο απόφασης δε θα μπορούσε να είναι στο πρώτο επίπεδο, αφού εξ ορισμού το τυχαίο δάσος, που αποτελείται από πολλά δέντρα απόφασης, θα έχει καλύτερα αποτελέσματα από ένα μόνο δέντρο απόφασης. Στο τρίτο επίπεδο ταξινομητών για το πείραμα αυτό τοποθετούμε τους ταξινομητές LDA και λογιστικής παλινδρόμησης που παρουσιάζουν ακρίβεια 69% και παρόμοιες τιμές σε όλες τις μετρικές. Στο τελευταίο επίπεδο τοποθετείται ο απλοϊκός Bayes που έχει αισθητά μικρότερες τιμές σε ακρίβεια και βαθμολογία F1 σε σχέση με τους άλλους ταξινομητές και ταξινομεί περισσότερα δείγματα από όσα πρέπει στα επιληπτικά.

Τέλος, με την προσθήκη καναλιών παρουσιάστηκε η αναμενόμενη και σημαντική βελτίωση που είχε αναφερθεί στο εισαγωγικό κομμάτι αυτού του κεφαλαίου σε όλους τους ταξινομητές πλην του απλοϊκού Bayes, που ούτως ή άλλως παρουσιάζει ασταθή και αναξιόπιστη συμπεριφορά σε όλα τα πειράματα.

7.3 Πείραμα με όλα τα κανάλια

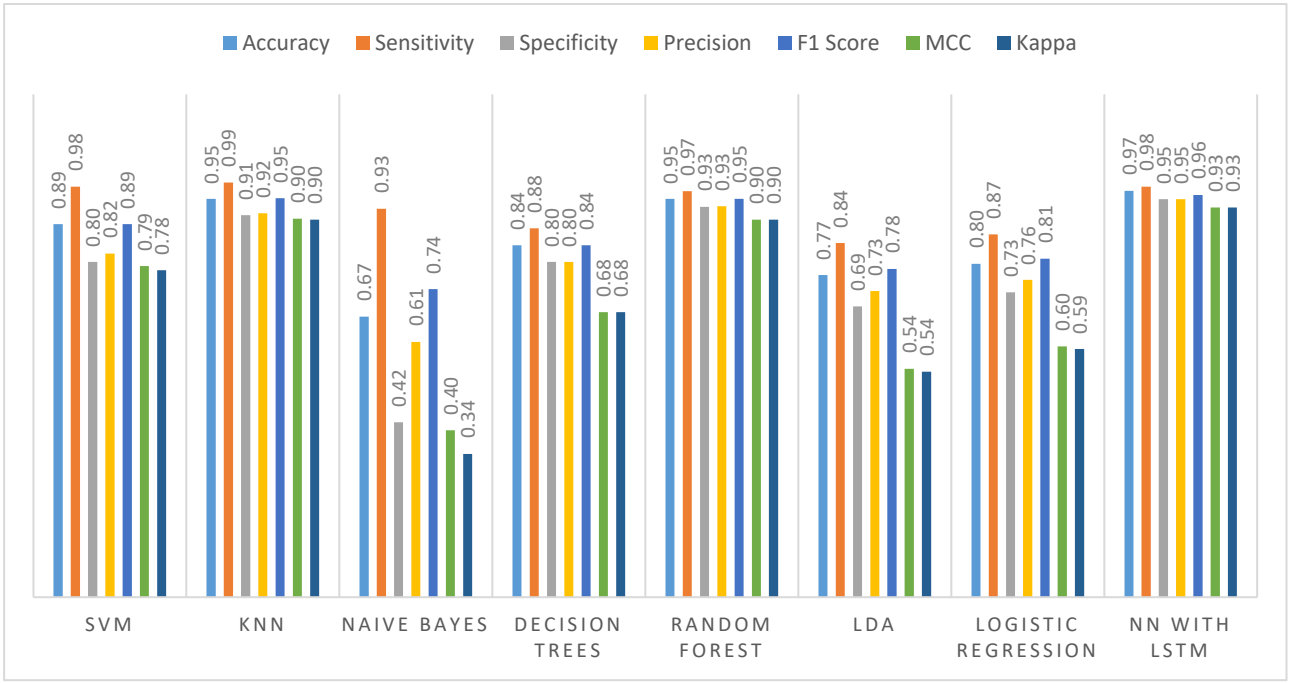
Στο τελευταίο πείραμα θα χρησιμοποιήσουμε όλα τα διαθέσιμα κανάλια. Πρόκειται για ένα πείραμα με πολύ περισσότερα δεδομένα από τα προηγούμενα, γεγονός που το κάνει

σαφώς πιο πολύπλοκο υπολογιστικά, αλλά ευκολότερο να διαχωριστεί από έναν καλά σχεδιασμένο ταξινομητή.

Στον Πίνακα 10 και το Σχήμα 36 παρατίθενται όλες οι μετρήσεις των μεθόδων ταξινόμησης που εφαρμόστηκαν σε αυτό το πείραμα, αφού πρώτα είχε χωριστεί το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης σε αναλογία 70%-30%.

Πίνακας 10: Μετρικές ταξινόμησης πειράματος με όλα τα κανάλια με διαχωρισμό συνόλου δεδομένων

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1 SCORE	MCC	KAPPA
SVM	0.89	0.98	0.80	0.82	0.89	0.79	0.78
KNN	0.95	0.99	0.91	0.92	0.95	0.90	0.90
NAIVE BAYES	0.67	0.93	0.42	0.61	0.74	0.40	0.34
DECISION TREES	0.84	0.88	0.80	0.80	0.84	0.68	0.68
RANDOM FOREST	0.95	0.97	0.93	0.93	0.95	0.90	0.90
LDA	0.77	0.84	0.69	0.73	0.78	0.54	0.54
LOGISTIC REGRESSION	0.80	0.87	0.73	0.76	0.81	0.60	0.59
NN WITH LSTM	0.97	0.98	0.95	0.95	0.96	0.93	0.93

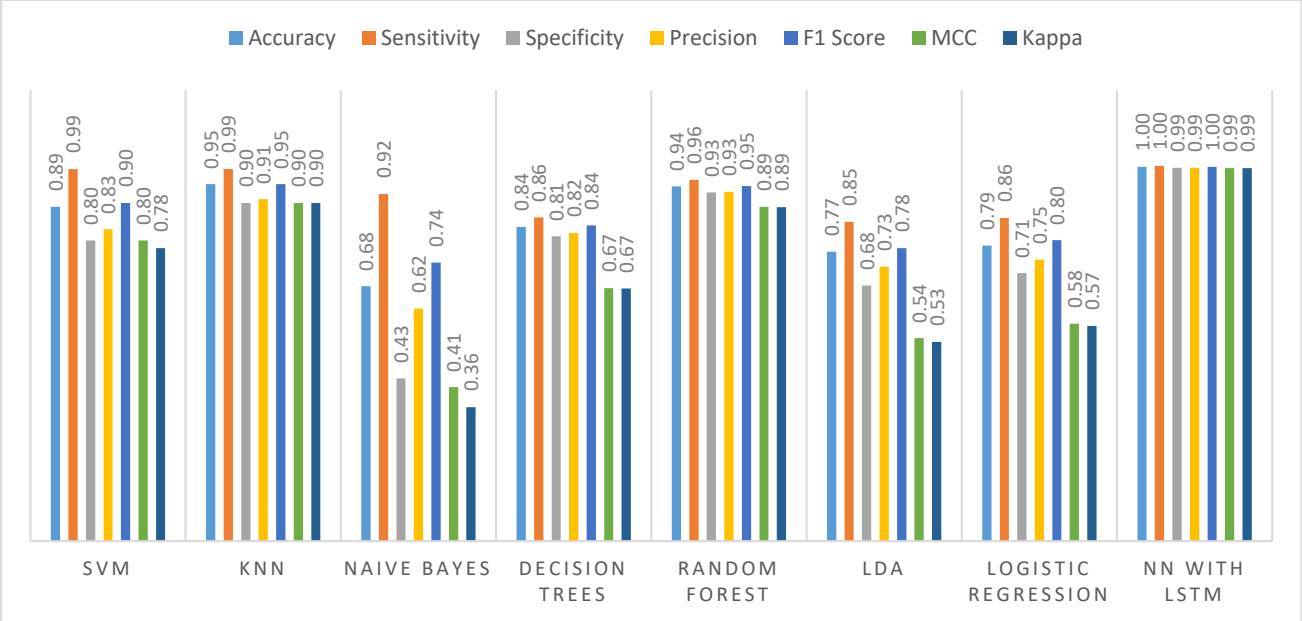


Σχήμα 36: Μετρικές ταξινόμησης πειράματος με όλα τα κανάλια με διαχωρισμό συνόλου δεδομένων

Στον Πίνακα 11 και το Σχήμα 37 παρατίθενται όλες οι μετρήσεις των μεθόδων ταξινόμησης που εφαρμόστηκαν σε αυτό το πείραμα εφαρμόζοντας διασταυρούμενη επικύρωση 5 επαναλήψεων.

Πίνακας 11: Μετρικές ταξινόμησης πειράματος με όλα τα κανάλια με διασταυρούμενη επικύρωση

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1 SCORE	MCC	KAPPA
SVM	0.89	0.99	0.80	0.83	0.90	0.80	0.78
KNN	0.95	0.99	0.90	0.91	0.95	0.90	0.90
NAIVE BAYES	0.68	0.92	0.43	0.62	0.74	0.41	0.36
DECISION TREES	0.84	0.86	0.81	0.82	0.84	0.67	0.67
RANDOM FOREST	0.94	0.96	0.93	0.93	0.95	0.89	0.89
LDA	0.77	0.85	0.68	0.73	0.78	0.54	0.53
LOGISTIC REGRESSION	0.79	0.86	0.71	0.75	0.80	0.58	0.57
NN WITH LSTM	1.00	1.00	0.99	0.99	1.00	0.99	0.99



Σχήμα 37: Μετρικές ταξινόμησης πειράματος με όλα τα κανάλια με διασταυρούμενη επικύρωση

Τα αποτελέσματα της διασταυρούμενης επικύρωσης είναι παρόμοια με εκείνα του πειράματος του μέσου καναλιού. Το συμπέρασμα που βγαίνει εδώ είναι ότι δε μας βοηθά ιδιαίτερα η χρήση της.

Σε σχέση με τις μετρικές ταξινόμησης, κορυφαίος ταξινομητής ήταν το νευρωνικό δίκτυο με LSTM, ενώ κοντινές τιμές παρουσιάζουν ο kNN και ο ταξινομητής τυχαίου δάσους. Οι τρεις αυτοί ταξινομητές θα τοποθετούνταν στο πρώτο επίπεδο με τις κορυφαίες επιδόσεις. Στο δεύτερο επίπεδο ξεχωρίζει το δέντρο απόφασης μαζί με τον SVM. Στο τρίτο επίπεδο, οι LDA και Λογιστική Παλινδρόμηση δείχνουν να λειτουργούν αρκετά καλύτερα με τα πολλά δεδομένα. Στο τελευταίο επίπεδο είναι και εδώ ο απλοϊκός Bayes με τα ίδια μειονεκτήματα που είδαμε και στις προηγούμενες ενότητες του παρόντος κεφαλαίου. Ο απλοϊκός Bayes δείχνει μια βελτιωμένη συμπεριφορά, αλλά απέχει σημαντικά σε επιδόσεις από τους ταξινομητές των δύο προηγούμενων επιπέδων.

Τέλος, στο πείραμα αυτό που χρησιμοποιήθηκαν όλα τα κανάλια η διαφορά στις μετρήσεις είναι πολύ μεγάλη και το πείραμα αυτό έχει τις καλύτερες μετρήσεις και από τα δύο προηγούμενα πειράματα.

8. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην τελευταία ενότητα της παρούσας διπλωματικής εργασίας θα γίνει μια ανασκόπηση των στόχων που τέθηκαν, των μεθόδων που χρησιμοποιήθηκαν και τον αποτελεσμάτων που επετεύχθησαν. Εκτός αυτών, θα δοθεί μια εικόνα των προβλημάτων που αντιμετωπίσαμε κατά την υλοποίηση και κάποιες ιδέες που μπορούν να εφαρμοστούν για μελλοντικές επεκτάσεις.

8.1 Σύνοψη και συμπεράσματα

Στις προηγούμενες ενότητες είδαμε όλη τη διαδικασία ανίχνευσης επιληπτικών δειγμάτων στο σύνολο EEG δεδομένων CHB-MIT. Τα δεδομένα χωρίζονται σε επιληπτικά και μη. Για την καλύτερη αξιοποίηση των εργαλείων μηχανικής μάθησης προχωρήσαμε σε δημιουργία χαρακτηριστικών ανά παράθυρο των 2 δευτερολέπτων στο πεδίο του χρόνου και της συχνότητας, ενώ υπολογίστηκε και ένα διμερές χαρακτηριστικό συσχέτισης καναλιών. Πριν το νέο σύνολο δεδομένων δοθεί ως είσοδος στους αλγορίθμους ταξινόμησης, κανονικοποίηθηκε, μειώθηκαν τα χαρακτηριστικά του (οι στήλες του) με τη μέθοδο PCA και εξισορροπήθηκε, ώστε οι δύο κλάσεις να αντιπροσωπεύονται ισότιμα, αφού έχει αποδειχθεί ότι οι αλγόριθμοι ML καλύτερα έτσι. Εν συνεχεία, δημιουργήσαμε τρία διαφορετικά πειράματα, τα πειράματα μέσου καναλιού, καναλιών δεξιού – αριστερού ημισφαιρίου και όλων των καναλιών, στα οποία εφαρμόσαμε τους ταξινομητές SVM, kNN, Απλοϊκός Bayes, Δέντρα Απόφασης, Τυχαίο Δάσος, LDA, Λογιστική Παλινδρόμηση και Νευρωνικό Δίκτυο με LSTM. Καθένα από τα τρία πειράματα δοκιμάστηκε τόσο με διαχωρισμό δεδομένων εκπαίδευσης και επικύρωσης σε αναλογία 70%-30% όσο και με διασταυρούμενη επικύρωση πέντε επαναλήψεων. Σε όλους τους ταξινομητές εφαρμόστηκε αναζήτηση πλέγματος για την επιλογή ενός κατάλληλου συνόλου υπερπαραμέτρων που θα δώσει την καλύτερη δυνατή ακρίβεια και αξιολογήθηκαν με τις ακόλουθες μετρικές: ακρίβεια, ευαισθησία, ειδικότητα, αξιοπιστία, βαθμολογία F1, συντελεστής συσχέτισης Matthews, συντελεστής κ του Cohen. Αυτές οι μετρικές μάς βοήθησαν, εκτός από το να δούμε πόσες σωστές και πόσες εσφαλμένες προβλέψεις υπήρχαν, να εξετάσουμε ποια κλάση εντοπίζει πιο εύκολα και που αντιμετωπίζει προβλήματα πρόβλεψης, καθώς και να τον έλεγχο συμφωνίας μεταξύ των κλάσεων. Σκοπός μας ήταν να μην ακολουθήσουμε την πεπατημένη της πρόβλεψης ανά ασθενή, αλλά να εξετάσουμε τα αποτελέσματα σε ένα σύνολο ασθενών όλα μαζί.

Από τα πειράματα που διεξήχθησαν βγήκε ως πόρισμα ότι υπάρχουν ταξινομητές που στο συγκεκριμένο σύνολο δεδομένων λειτουργούν εξαιρετικά ακόμα και με τη χρήση λιγότερων καναλιών. Συγκεκριμένα, σταθερά οι καλύτεροι ταξινομητές ήταν αυτοί του τυχαίου δάσους, του νευρωνικού δικτύου με LSTM και ο kNN, ενώ ο χειρότερος είναι ο απλοϊκός Bayes. Μια μέση απόδοση είχαν το δέντρο απόφασης, ο SVM, ο LDA και η Λογιστική Παλινδρόμηση, με τους δύο πρώτους να έχουν λίγο καλύτερα αποτελέσματα και τους άλλους δύο να έχουν πολύ κοντινά μεταξύ τους και λίγο κατώτερα αποτελέσματα. Άλλο ένα πειραματικό δεδομένο που προέκυψε είναι ότι η διασταυρούμενη επικύρωση μας βοήθησε μόνο στο πείραμα με τα κανάλια αριστερού και δεξιού ημισφαιρίου, ενώ στα άλλα δύο δεν παρατηρήθηκε αξιοσημείωτη βελτίωση. Για την ακρίβεια, σε ορισμένες περιπτώσεις παρατηρήθηκαν και χειρότερα αποτελέσματα. Τέλος, σε γενικές γραμμές παρουσιάστηκε η αναμενόμενη βελτίωση στα πειράματα που λαμβάνουν ως είσοδο δεδομένα από περισσότερα κανάλια κεφαλαίου σε όλους τους ταξινομητές πλην του απλοϊκού Bayes που ούτως ή άλλως παρουσιάζει ασταθή και αναξιόπιστη συμπεριφορά σε όλα τα πειράματα.

8.2 Προβλήματα και μελλοντικές επεκτάσεις

Το σημαντικότερο πρόβλημα που αντιμετωπίσαμε ήταν με το σύνολο δεδομένων, που, όμως, ήταν η μοναδική επιλογή που υπήρχε για το θέμα μας, όπως αναλύθηκε και στο 3.1. Όπως προαναφέραμε, το CHB-MIT είναι εντελώς μη ισορροπημένο αφού η συντριπτική πλειοψηφία ανήκει στη μη επιληπτική κατηγορία δειγμάτων (98.64%), ενώ η επιληπτική δεν εκπροσωπείται επαρκώς (1.36%). Η αναλογία αυτή αφορά μόνο τις καταγραφές που περιέχουν επιληπτικό επεισόδιο. Αν εντάξουμε και τις υπόλοιπες καταγραφές, το πρόβλημα της ανισορροπίας θα οξυνθεί. Η δημιουργία συνθετικών δεδομένων για να εξισορροπήσει ένα τέτοιο σύνολο δεδομένων θα δημιουργούσε πολλά παρόμοια σημεία, αφού όλες οι γνωστές μέθοδοι λειτουργούν με παρεμβολή (interpolation) στα υπάρχοντα σημεία. Άρα, μονόδρομος ήταν να κρατήσουμε μόνο το 20% των αρχικών στοιχείων της μη επιληπτικής κλάσης μικραίνοντας σημαντικά το σύνολο δεδομένων. Συνεπώς, θα ήταν πολύ χρήσιμο να εφαρμόσουμε τα ίδια πειράματα σε ένα σύνολο δεδομένων σημαντικά μεγαλύτερο με περισσότερα επιληπτικά δεδομένα, κάτι που θα μας έδινε πιο αξιόπιστα αποτελέσματα.

Μια ακόμα επέκταση θα ήταν να εντάξουμε στη μελέτη μας τους χρόνους εκτέλεσης των ταξινομητών, ώστε να έχουμε ένα ακόμα μέτρο σύγκρισης για επιλογή ταξινομητή. Ακόμα, θα μπορούσε κάποιος να χρησιμοποιήσει GPU για να επιταχύνει τους υπολογισμούς των μεθόδων ταξινόμησης.

Σε σχέση με τον υπολογισμό χαρακτηριστικών, μια εναλλακτική δοκιμή θα ήταν η ένταξη των χαρακτηριστικών συσχέτισης και συγκεκριμένα της μέγιστης αλληλοσυσχέτισης στο πείραμα με όλα τα κανάλια.

Τέλος, όσον αφορά τους ταξινομητές, μια τροποποίηση που μπορεί να γίνει είναι το υπάρχον νευρωνικό δίκτυο να δοκιμαστεί με ανεπεξέργαστα δεδομένα. Παράλληλα, υπάρχουν και άλλοι ταξινομητές που ενδεχομένως να παρουσιάσουν καλές επιδόσεις σε αυτό το πρόβλημα. Η αναφορά [4] προτείνει πολλούς τέτοιους, όπως και πληθώρα εναλλακτικών χαρακτηριστικών που μπορούν να δοκιμαστούν.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Accuracy	Ακρίβεια
Activation function	Συνάρτηση ενεργοποίησης
Adder	Αθροιστής
Approximate Entropy	Προσεγγιστική Εντροπία
Artificial Intelligence	Τεχνητή νοημοσύνη
Artificial Neural Networks	Τεχνητά νευρωνικά δίκτυα
Bias	Προκατάληψη
Brain Activity	Εγκεφαλική Δραστηριότητα
Centroid	Κεντροειδές
Classification Algorithms	Αλγόριθμοι Ταξινόμησης
Cluster	Συστάδα
Computer Vision	Μηχανική Όραση
Confusion Matrix	Πίνακας Σύγχυσης
Correlation Features	Χαρακτηριστικά Συσχέτισης
Cross-validation	Διασταυρούμενη επικύρωση
Decision Trees	Δέντρα Απόφασης
Deep Learning	Βαθιά Μάθηση
Dense Layer	Πυκνό επίπεδο
Dimensionality Reduction	Μείωση Διαστάσεων
Electroencephalography	Ηλεκτροεγκεφαλογράφημα
Feature Extraction	Εξαγωγή Χαρακτηριστικών
Feature Normalization	Κανονικοποίηση χαρακτηριστικών
Grid Search	Αναζήτηση Πλέγματος
Hyperparameter	Υπερπαράμετρος
Ictal state	Κατάσταση επιληπτικής κρίσης
Interictal state	Μεσοκρισική κατάσταση
Interpolation	Παρεμβολή
k Nearest Neighbors	k Εγγύτεροι Γείτονες
Kurtosis	Κύρτωση
Lazy Learning	Οκνηρή Μάθηση
Linear Combiner	Γραμμικός Συνδυαστής
Linear Discriminant Analysis	Ανάλυση Γραμμικής Διάκρισης
Logistic Regression	Λογιστική Παλινδρόμηση
Machine Learning	Μηχανική Μάθηση
Majority class	Επικρατούσα κλάση
Maximum cross-correlation	Μέγιστη αλληλοσυσχέτιση
Mean	Αριθμητική Μέση Τιμή
Mean squared error	Μέσο τετραγωνικό σφάλμα
Median	Διάμεσος
Minority class	Μειονεκτούσα κλάση
Naïve Bayes	Απλοϊκός Bayes
Normalization	Κανονικοποίηση
Optimizer	Βελτιστοποιητής
Overfitting	Υπερπροσαρμογή
Oversampling	Υπερδειγματοληψία
Peak to Peak	Απόσταση Μέγιστου-Ελάχιστου Σημείου
Postictal state	Κατάσταση μετά την επιληπτική κρίση

Power Spectral Density	Φασματική Πυκνότητα Ισχύος
Precision	Ακρίβεια
Preictal state	Κατάσταση πριν την επιληπτική κρίση
Principal Component Axes	Ανάλυση Κύριων Συνιστωσών
Random Forest	Τυχαίο Δάσος
Raw data	Ανεπεξέργαστα δεδομένα
Recall	Ανάκληση
Root Mean Square	Ενεργός τιμή σήματος
Sample Entropy	Εντροπία δείγματος
Scalp electroencephalogram	Επιφανειακό ηλεκτροεγκεφαλογράφημα
Seizure Detection	Ανίχνευση Επιληπτικών Κρίσεων
Semi-Supervised Learning	Ημιεποπτευόμενη Μάθηση
Sensitivity	Ευαισθησία
Skewness	Ασυμμετρία
Specificity	Ειδικότητα
Spectral Features	Φασματικά χαρακτηριστικά
Standard Deviation	Τυπική Απόκλιση
Supervised Learning	Εποπτευόμενη Μάθηση
Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
Time Domain Features	Χαρακτηριστικά στο πεδίο του χρόνου
Undersampling	Υποδειγματοληψία
Unsupervised Learning	Μη Εποπτευόμενη Μάθηση
Variance	Διακύμανση
Zero Crossing Rate	Ρυθμός Διάσχισης Μηδενικού Άξονα

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

ADASYN	ADAptive SYNthetic
AI	Artificial Intelligence
ANN	Artificial Neural Networks
CHB	Children's Hospital Boston
EDF	European Data Format
EEG	Electroencephalography
ECG	Electrocardiography
FFT	Fast Fourier Transform
GPU	Graphic Processing Unit
kNN	k Nearest Neighbors
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
MCC	Matthew's Correlation Coefficient
ML	Machine Learning
MLE	Maximum Likelihood Estimate
MIT	Massachusetts Institute of Technology
NB	Naïve Bayes
PCA	Principal Component Analysis
PSD	Power Spectral Density
PSG	PolySomnoGraphy
RBF	Radial Basis Function
ReLU	Rectifier Linear Unit
RMS	Root Mean Square
RNN	Recurrent Neural Network
SampEn	Sample Entropy
SMOTE	Synthetic Minority Oversampling TEchnique
SUDEP	Sudden Unexpected Death in EPilepsy
SVD	Singular Value Decomposition
SVM	Support Vector Machines
TUH	Temple University Hospital
VNS	Vagal Nerve Stimulus
ZCR	Zero Crossing Rate

ΕΕ	Ευρωπαϊκή Ένωση
ΕΚΠΑ	Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΠΑΡΑΡΤΗΜΑ Ι

Αρχεία EDF (European Data Format)

Το EDF είναι μια τυπική μορφή αρχείου που έχει σχεδιαστεί για ανταλλαγή και αποθήκευση ιατρικών χρονοσειρών. Όντας μια ανοιχτή και μη ιδιόκτητη μορφή, το EDF χρησιμοποιείται συνήθως για αρχειοθέτηση, ανταλλαγή και ανάλυση δεδομένων από εμπορικές συσκευές σε μορφή ανεξάρτητη από το σύστημα απόκτησης. Με αυτόν τον τρόπο, τα δεδομένα μπορούν να ανακτηθούν και να αναλυθούν από ανεξάρτητο λογισμικό.

Το EDF δημοσιεύθηκε το 1992 και αποθηκεύει πολυκαναλικά δεδομένα, επιτρέποντας διαφορετικά ποσοστά δειγμάτων για κάθε σήμα. Έκτοτε, το EDF έγινε το ντε φάκτο πρότυπο για εγγραφές EEG και PSG σε εμπορικό εξοπλισμό και πολυκεντρικά ερευνητικά έργα. Εσωτερικά περιλαμβάνει μια κεφαλίδα και μία ή περισσότερες εγγραφές δεδομένων. Η κεφαλίδα περιέχει μερικές γενικές πληροφορίες (αναγνώριση ασθενούς, ώρα έναρξης) και τεχνικές προδιαγραφές κάθε σήματος (βαθμονόμηση, ρυθμός δειγματοληψίας, φιλτράρισμα), κωδικοποιημένοι ως χαρακτήρες ASCII. Τα αρχεία δεδομένων περιέχουν δείγματα ως ακέραιοι αριθμοί 16-bit.

ΑΝΑΦΟΡΕΣ

[1] Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S., & Kavehei, O. (2018). Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Networks, 105, 104-111.

[2] Shoeb, A., Edwards, H., Connolly, J., Bourgeois, B., Treves, S. T., & Guttag, J. (2004). Patient-specific seizure onset detection. *Epilepsy & Behavior*, *5*(4), 483-498.

[3] Tsiouris, K. M., Pezoulas, V. C., Zervakis, M., Konitsiotis, S., Koutsouris, D. D., & Fotiadis, D. I. (2018). A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Computers in biology and medicine*, *99*, 24-37.

[4] Boonyakitanont, P., Lek-Uthai, A., Chomtho, K., & Songsiri, J. (2020). A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. *Biomedical Signal Processing and Control*, *57*, 101702.

[5] Mansouri, A., Singh, S. P., & Sayood, K. (2019). Online eeg seizure detection and localization. *Algorithms*, *12*(9), 176.

[6] Harpale, V., & Bairagi, V. (2018). An adaptive method for feature selection and extraction for classification of epileptic EEG signal in significant states. Journal of King Saud University-Computer and Information Sciences.

[7] Shoeb, A. H. (2009). *Application of machine learning to epileptic seizure onset detection and treatment* (Doctoral dissertation, Massachusetts Institute of Technology).

[8] Shoeb, A. H., & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 975-982).

[9] Mirowski, P. W., LeCun, Y., Madhavan, D., & Kuzniecky, R. (2008, October). Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG. In 2008 IEEE workshop on machine learning for signal processing (pp. 244-249). IEEE.

[10] Wen, T., & Zhang, Z. (2017). Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification. Medicine, 96(19).

[11] Usman, S. M., Usman, M., & Fong, S. (2017). Epileptic seizures prediction using machine learning methods. *Computational and mathematical methods in medicine*, *2017*.

[12] Nandy, A., Alahe, M. A., Uddin, S. N., Alam, S., Nahid, A. A., & Awal, M. A. (2019, January). Feature extraction and classification of EEG signals for seizure detection. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 480-485). IEEE.

[13] Andrzejak, R. G., Schindler, K., & Rummel, C. (2012). Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. Physical Review E, 86(4), 046206.

[14] Minasyan, G. R., Chatten, J. B., Chatten, M. J., & Harner, R. N. (2010). Patient-specific early seizure detection from scalp EEG. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, *27*(3), 163.

[15] Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology, 278(6), H2039-H2049.

[16] Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M. P., & Tobochnik, S. (2014, December). The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1-5). IEEE.

[17] Κουρσιουμπάς Νικόλαος, Μαγουλά Βασιλεία, Music Recommendation System based on EEG Sentiment Analysis using ML Techniques, 2019, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

[18] Παπανάστου Καλλιόπη, Βελτιστοποίηση παραμέτρων αλγορίθμων κατηγοριοποίησης βιολογικών δεδομένων, 2020, Βιβλιοθήκη Σχολής Θετικών Επιστημών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

[19] Σαράφης Δημήτριος, Μιχαλόπουλος Παναγιώτης. Εκτίμηση της μουσικής προτίμησης διαφορετικών τύπων μουσικής ανάμειξης μέσω της ανάλυσης της εγκεφαλικής λειτουργικής συνδεσιμότητας με φάσματα ανώτερης τάξης, 2017, ΙΚΕΕ / Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης – Βιβλιοθήκη

[20] Αναστασίου, Ε. (2019). *Αλγόριθμοι μηχανικής μάθησης και εφαρμογές σε ιατροβιολογικά προβλήματα*. Προπτυχιακή Διπλωματική Εργασία. http://hdl.handle.net/10889/13208

[21] Παϊδούση, Ε. (2016). *Δέντρα Αποφάσεων*. Μεταπτυχιακή Διπλωματική Εργασία. http://hdl.handle.net/10889/10829

[22] Δημαράκη, Α. (2017). *Εξαγωγή Χαρακτηριστικών και Ταξινόμηση Βιολογικών Σημάτων για χρήση στα Συστήματα Υποβοήθησης της Διάγνωσης*. Διπλωματική Εργασία. http://artemis.cslab.ece.ntua.gr:8080/jspui/handle/123456789/13484

[23] Archive.physionet.org. CHB-MIT Scalp EEG Database. [online] Available at: <https://archive.physionet.org/pn6/chbmit/> [Προσπελάστηκε 08/02/2020].

[24] Archive.ics.uci.edu. UCI Machine Learning Repository: Epileptic Seizure Recognition Data Set. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition> [Προσπελάστηκε 08/02/2020].
[25] Epilepsy-database.eu.[online] Available at: <http://epilepsy-database.eu/> [Προσπελάστηκε 08/02/2020].
[26] Epilepsy.uni-freiburg.de. EEG Database — Seizure Prediction Project Freiburg. [online] Available at: <http://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database> [Προσπελάστηκε 08/02/2020].
[27] Picone, J., 2020. *Temple University EEG Corpus - Downloads*. [online] Isip.piconepress.com. Available at: <https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml> [Προσπελάστηκε 08/02/2020].
[28] Kaggle. American Epilepsy Society Seizure Prediction Challenge | Kaggle. [online] Available at: <https://www.kaggle.com/c/seizure-prediction> [Προσπελάστηκε 08/02/2020].
[29] Kaggle.com. Confused student EEG brainwave data | Kaggle. [online] Available at: <https://www.kaggle.com/wanghaohan/confused-eeg> [Προσπελάστηκε 08/02/2020].
[30] Bhattacharyya, I., 2018. SMOTE And ADASYN (Handling Imbalanced Data Set). [online] Medium. Available at: <https://medium.com/coinmonks/smote-and-adasyn-handling-imbalanced-data-set-34f5223e167> [Προσπελάστηκε 22/03/2020].
[31] Rui, N., 2018. *An Introduction To ADASYN (With Code!)*. [online] Medium. Available at: <https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa> [Προσπελάστηκε 22/03/2020].
[32] Brownlee, J., 2020. Assessing And Comparing Classifier Performance With ROC Curves. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/> [Προσπελάστηκε 22/03/2020].
[33] Brownlee, J., 2020. Random Oversampling And Undersampling For Imbalanced Classification. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/> [Προσπελάστηκε 22/03/2020].
[34] Shahul, E., 2019. Tackling Class Imbalance. [online] Kaggle.com. Available at: <https://www.kaggle.com/shahules/tackling-class-imbalance> [Προσπελάστηκε 22/03/2020].
[35] Alencar, R., 2017. Resampling Strategies For Imbalanced Datasets. [online] Kaggle.com. Available at: <https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets?fbclid=IwAR2MAB59IeSpHc5lkyZWNnmrJ3cFdLX26vNv9rSAbkTXLwug3h0JsVPes-8> [Προσπελάστηκε 22/03/2020].
[36] Banerjee, P., 2019. Data Preprocessing Project - Imbalanced Classes Problem. [online] GitHub. Available at: <https://github.com/pb111/Data-Preprocessing-Project-Imbalanced-Classes-Problem> [Προσπελάστηκε 22/03/2020].
[37] Bilogar, A., 2017. Oversampling With SMOTE And ADASYN. [online] Kaggle.com. Available at: <https://www.kaggle.com/residentmario/oversampling-with-smote-and-adasyn> [Προσπελάστηκε 22/03/2020].
[38] Brownlee, J., 2015. 8 Tactics To Combat Imbalanced Classes In Your Machine Learning Dataset. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/?fbclid=IwAR1D7Mc_Ip-pbur_OeLozrSctEKcrn30LuEiU6IDJVHKHyocBywuCc0a68o> [Προσπελάστηκε 22/03/2020].
[39] Brownlee, J., 2020. SMOTE For Imbalanced Classification With Python. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/> [Προσπελάστηκε 22/03/2020].
[40] Foong, N., 2019. Tips And Tricks For Handling Configuration Files In Python. [online] Medium. Available at: <https://medium.com/better-programming/tips-and-tricks-for-handling-configuration-files-in-python-a9d7429aa50b> [Προσπελάστηκε 22/03/2020].
[41] Pugh, D., 2019. Balancing Datasets And Generating Synthetic Data With SMOTE. [online] Data Science Campus. Available at: <https://datasciencecampus.github.io/balancing-data-with-smote/> [Προσπελάστηκε 22/03/2020].
[42] Shetye, A., 2019. Feature Selection With Sklearn And Pandas. [online] Medium. Available at: <https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b> [Προσπελάστηκε 22/03/2020].

[43] En.wikipedia.org. 2020. *European Data Format*. [online] Available at: <https://en.wikipedia.org/wiki/European_Data_Format> [Προσπελάστηκε 04/04/2020].

[44] Alvarez-Estevez, D., n.d. European Data Format (EDF). [online] Edfplus.info. Available at: <https://www.edfplus.info/> [Προσπελάστηκε 04/04/2020].

[45] Nahrstaedt, H., 2020. *Pyedflib -EDF/BDF Toolbox In Python — Pyedflib Documentation*. [online] Pyedflib.readthedocs.io. Available at: <https://pyedflib.readthedocs.io/en/latest/> [Προσπελάστηκε 04/04/2020].

[46] Schachter, S., 2014. *What Happens During A Seizure?*. [online] Epilepsy Foundation. Available at: <https://www.epilepsy.com/learn/about-epilepsy-basics/what-happens-during-seizure> [Προσπελάστηκε 04/04/2020].

[47] Merlin Praveena, D., Angelin Sarah, D., & Thomas George, S. (2020). Deep Learning Techniques for EEG Signal Applications–a Review. *IETE Journal of Research*, 1-8.

[48] En.wikipedia.org. 2020. *Turing Test*. [online] Available at: <https://en.wikipedia.org/wiki/Turing_test> [Προσπελάστηκε 04/04/2020].

[49] Mathworks. n.d. What Is Deep Learning? | How It Works, Techniques & Applications. [online] Available at: <https://www.mathworks.com/discovery/deep-learning.html> [Προσπελάστηκε 04/04/2020].

[50] Bitbrain. 2020. How Deep Learning Is Changing Machine Learning AI In EEG Signal Processing. [online] Available at: <https://www.bitbrain.com/blog/ai-eeg-data-processing> [Προσπελάστηκε 04/04/2020].

[51] Heath, N., 2020. *What Is Machine Learning? Everything You Need To Know | Zdnet*. [online] ZDNet. Available at: <https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/> [Προσπελάστηκε 04/04/2020].

[52] VanderPlas, J., 2020. *In Depth: Principal Component Analysis | Python Data Science Handbook*. [online] Jakevdp.github.io. Available at: <https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html> [Προσπελάστηκε 13/04/2020].

[53] Haidara, M., 2019. Dimensionality Reduction Toolbox In Python. [online] Towards Data Science. Available at: <https://towardsdatascience.com/dimensionality-reduction-toolbox-in-python-9a18995927cd> [Προσπελάστηκε 13/04/2020].

[54] Towards Data Science. 2019. Dimension Reduction Techniques With Python. [online] Available at: <https://towardsdatascience.com/dimension-reduction-techniques-with-python-f36ca7009e5c> [Προσπελάστηκε 13/04/2020].

[55] Nelson, D., 2020. Dimensionality Reduction In Python With Scikit-Learn. [online] Stack Abuse. Available at: <https://stackabuse.com/dimensionality-reduction-in-python-with-scikit-learn/> [Προσπελάστηκε 13/04/2020].

[56] Bhattacharyya, S., 2018. Understanding PCA (Principal Component Analysis) With Python. [online] Towards Data Science. Available at: <https://towardsdatascience.com/dive-into-pca-principal-component-analysis-with-python-43ded13ead21> [Προσπελάστηκε 13/04/2020].

[57] Mikulski, B., 2019. *PCA—How To Choose The Number Of Components?*. [online] mikulskibartosz. Available at: <https://www.mikulskibartosz.name/pca-how-to-choose-the-number-of-components/> [Προσπελάστηκε 13/04/2020].

[58] Καλατζής, Ι., 2018. Ανάλυση Πρωτευόντων Συστατικών. [online] Πανεπιστήμιο Δυτικής Αττικής. Available at: <https://medisp.bme.uniwa.gr/eclass/modules/document/file.php/MTMBIT101/%ce%a5%ce%9b%ce%99%ce%9a%ce%9f%20%ce%99.%20%ce%9a%ce%91%ce%9b%ce%91%ce%a4%ce%96%ce%97%20(DESCRIPTIVE%20STATISTICS,%20HYPOTHESIS%20TESTING,%20CLUSTERING,%20PCA,%20LDA)/4.%20%ce%91%ce%bd%ce%ac%ce%bb%cf%85%cf%83%ce%b7%20%ce%a0%cf%81%cf%89%cf%84%ce%b5%cf%85%cf%8c%ce%bd%cf%84%cf%89%ce%bd%20%ce%a3%cf%85%cf%83%cf%84%ce%b1%cf%84%ce%b9%ce%ba%cf%8e%ce%bd%20(PCA)/PCA%20(I.%20Kalatzis%202018)%20(10).pdf> [Προσπελάστηκε 10/05/2020].

[59] Scikit-learn. 2020. 1.10. Decision Trees — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/tree.html> [Προσπελάστηκε 10/05/2020].

[60] Scikit-learn. 2020. Sklearn.Tree.Decisiontreeclassifier — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier> [Προσπελάστηκε 10/05/2020].

[61] Jordan, J., 2017. Decision trees. [Blog] DATA SCIENCE, Available at: <https://www.jeremyjordan.me/decision-trees/> [Προσπελάστηκε 28/05/2020].

[62] Scikit-learn. 2020. Sklearn.Neighbors.Kneighborsclassifier — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> [Προσπελάστηκε 28/05/2020].

[63] Bhivarkar, A., 2017. *KNN For Classification Using Scikit-Learn*. [online] Kaggle. Available at: <https://www.kaggle.com/amolbhivarkar/knn-for-classification-using-scikit-learn> [Προσπελάστηκε 28/05/2020].

[64] Sanjay, M., 2018. KNN Using Scikit-Learn. [online] Towards Data Science. Available at: <https://towardsdatascience.com/knn-using-scikit-learn-c6bed765be75> [Προσπελάστηκε 28/05/2020].

[65] Scikit-learn. 2020. *Sklearn.Discriminant_Analysis.Lineardiscriminantanalysis — Scikit-Learn 0.23.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html> [Προσπελάστηκε 09/06/2020].

[66] Καλατζής, Ι., 2017. Ανάλυση Γραμμικής Διάκρισης. [online] Πανεπιστήμιο Δυτικής Αττικής. Available at: <https://medisp.bme.uniwa.gr/eclass/modules/document/file.php/MTMBIT101/%CE%A5%CE%9B%CE%99%CE%9A%CE%9F%20%CE%99.%20%CE%9A%CE%91%CE%9B%CE%91%CE%A4%CE%96%CE%97%20%28DESCRIPTIVE%20STATISTICS%2C%20HYPOTHESIS%20TESTING%2C%20CLUSTERING%2C%20PCA%2C%20LDA%29/5.%20%CE%91%CE%BD%CE%AC%CE%BB%CF%85%CF%83%CE%B7%20%CE%93%CF%81%CE%B1%CE%BC%CE%BC%CE%B9%CE%BA%CE%AE%CF%82%20%CE%94%CE%B9%CE%AC%CE%BA%CF%81%CE%B9%CF%83%CE%B7%CF%82%20%28LDA%29/LDA%20%28I.%20Kalatzis%202017%29.pdf> [Προσπελάστηκε 09/06/2020].

[67] Scikit-learn. 2020. 1.1. Linear Models — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression> [Προσπελάστηκε 09/06/2020].

[68] Scikit-learn. 2020. Sklearn.Linear_Model.Logisticregression — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> [Προσπελάστηκε 09/06/2020].

[69] Li, S., 2017. Building A Logistic Regression In Python, Step By Step. [online] Towards Data Science. Available at: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> [Προσπελάστηκε 09/06/2020].

[70] Πετρίδης, Δ. 2015. ΑΝΑΛΥΣΗ ΑΝΤΙΣΤΟΙΧΙΩΝ. [Κεφάλαιο Συγγράμματος]. Στο Πετρίδης, Δ. 2015. Ανάλυση πολυμεταβλητών τεχνικών. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 3. Διαθέσιμο στο: http://hdl.handle.net/11419/2134

[71] Scikit-learn. 2020. 1.9. Naive Bayes — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/naive_bayes.html> [Προσπελάστηκε 21/06/2020].

[72] Scikit-learn. 2020. *3.2.4.3.1. Sklearn.Ensemble.Randomforestclassifier — Scikit-Learn 0.23.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [Προσπελάστηκε 21/06/2020].

[73] Koehrsen, W., 2017. Random Forest In Python. [online] Towards Data Science. Available at: <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0> [Προσπελάστηκε 21/06/2020].

[74] Scikit-learn. 2020. 1.4. Support Vector Machines — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/svm.html> [Προσπελάστηκε 21/06/2020].

[75] Malik, U., n.d. Implementing SVM And Kernel SVM With Python's Scikit-Learn. [online] Stack Abuse. Available at: <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/> [Προσπελάστηκε 04/07/2020].

[76] Sreenivasa, S., 2020. Radial Basis Function (RBF) Kernel: The Go-To Kernel. [online] Towards Data Science. Available at: <https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a> [Προσπελάστηκε 04/07/2020].

[77] Ben Fraj, M., 2018. In Depth: Parameter Tuning For SVC. [online] Medium. Available at: <https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769> [Προσπελάστηκε 04/07/2020].

[78] Team, K., 2020. Keras: The Python Deep Learning API. [online] Keras.io. Available at: <https://keras.io/> [Προσπελάστηκε 04/07/2020].

[79] Brownlee, J., 2016. *Time Series Prediction With LSTM Recurrent Neural Networks In Python With Keras*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/?fbclid=IwAR0Qs9suZCUsLW4Ef_2J9OeSQi6ffyDH19T31WRqPFgRu9rR4W1EXXKwHNc> [Προσπελάστηκε 24/07/2020].

[80] Heaton, J., 2020. T81-558: Applications Of Deep Neural Networks. [online] GitHub. Available at: <https://github.com/jeffheaton/t81_558_deep_learning/blob/master/t81_558_class_10_2_lstm.ipynb > [Προσπελάστηκε 24/07/2020].

[81] Sinha, N., 2018. Understanding LSTM And Its Quick Implementation In Keras For Sentiment Analysis. [online] Towards Data Science. Available at: <https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47> [Προσπελάστηκε 24/07/2020].

[82] Olah, C., 2015. Understanding LSTM Networks. [Blog] Colah's blog, Available at: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> [Προσπελάστηκε 24/07/2020].

[83] Brownlee, J., 2017. How To Diagnose Overfitting And Underfitting Of LSTM Models. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/> [Προσπελάστηκε 25/07/2020].

[84] Brownlee, J., 2018. Use Early Stopping To Halt The Training Of Neural Networks At The Right Time. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/> [Προσπελάστηκε 25/07/2020].

[85] Brownlee, J., 2019. *How To Use Learning Curves To Diagnose Machine Learning Model Performance*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/> [Προσπελάστηκε 26/07/2020].

[86] Schlüter, N., 2019. Don't Overfit! - How To Prevent Overfitting In Your Deep Learning Models. [online] Towards Data Science. Available at: <https://towardsdatascience.com/dont-overfit-how-to-prevent-overfitting-in-your-deep-learning-models-63274e552323> [Προσπελάστηκε 26/07/2020].

[87] Rosebrock, A., 2019. Why Is My Validation Loss Lower Than My Training Loss?. [online] PyImageSearch. Available at: <https://www.pyimagesearch.com/2019/10/14/why-is-my-validation-loss-lower-than-my-training-loss/> [Προσπελάστηκε 26/07/2020].

[88] Scikit-learn. 2020. Sklearn.Model_Selection.Gridsearchcv — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?fbclid=IwAR2FCvXCski8i2kKILuXuG6OYI671AI0DyHesYWdqFAGSWdAq2z3QsgCZ3w> [Προσπελάστηκε 02/08/2020].

[89] Brownlee, J., 2016. How To Grid Search Hyperparameters For Deep Learning Models In Python With Keras. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/> [Προσπελάστηκε 02/08/2020].

[90] Allibhai, E., 2018. Building A K-Nearest-Neighbors (K-NN) Model With Scikit-Learn. [online] Medium. Available at: <https://towardsdatascience.com/building-a-k-nearest-neighbors-k-nn-model-with-scikit-learn-51209555453a> [Προσπελάστηκε 02/08/2020].

[91] Medium. 2019. An Introduction To Grid Search. [online] Available at: <https://medium.com/datadriveninvestor/an-introduction-to-grid-search-ff57adcc0998> [Προσπελάστηκε 02/08/2020].

[92] Bakharia, A., 2016. SVM Parameter Tuning In Scikit Learn Using Gridsearchcv. [online] Medium. Available at: <https://medium.com/@aneesha/svm-parameter-tuning-in-scikit-learn-using-gridsearchcv-2413c02125a0> [Προσπελάστηκε 02/08/2020].

[93] Erik, G., 2018. K-Neighbors Classifier With Gridsearchcv Basics. [online] Medium. Available at: <https://medium.com/@erikgreenj/k-neighbors-classifier-with-gridsearchcv-basics-3c445ddeb657> [Προσπελάστηκε 02/08/2020].

[94] Polat, E., 2018. Grid Search With Logistic Regression. [online] Kaggle. Available at: <https://www.kaggle.com/enespolat/grid-search-with-logistic-regression> [Προσπελάστηκε 02/08/2020].

[95] En.wikipedia.org. 2020. *Sample Entropy*. [online] Available at: <https://en.wikipedia.org/wiki/Sample_entropy> [Προσπελάστηκε 09/09/2020].

[96] Κοντέος, Γ., n.d. [online] Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Μακεδονίας. Available at: <https://openclass.teiwm.gr/modules/document/file.php/BA-G110/%CE%A3%CE%A4%CE%91%CE%A4%CE%99%CE%A3%CE%A4%CE%99%CE%9A%CE%97_2.pdf> [Προσπελάστηκε 09/09/2020].

[97] Παπαδόπουλος, Γ., n.d. *Περιγραφική Στατιστική*. [online] Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Μακεδονίας - Εργαστήριο Μαθηματικών&Στατιστικής. Available at: <https://www.aua.gr/gpapadopoulos/files/perigrafiki11.pdf> [Προσπελάστηκε 09/09/2020].

[98] Brownlee, J., 2016. Feature Selection For Machine Learning In Python. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/feature-selection-machine-learning-python/> [Προσπελάστηκε 09/09/2020].

[99] Vallat, R., 2018. Compute The Average Bandpower Of An EEG Signal. [online] Raphaelvallat.com. Available at: <https://raphaelvallat.com/bandpower.html> [Προσπελάστηκε 12/09/2020].

[100] Pyeeg.sourceforge.net. 2010. *Pyeeg Reference Guide — Pyeeg Reference Guide V0.02 R1 Documentation*. [online] Available at: <http://pyeeg.sourceforge.net/> [Προσπελάστηκε 12/09/2020].

[101] Lakshmanan, S., 2019. How, When, And Why Should You Normalize / Standardize / Rescale Your Data?. [online] Towards AI. Available at: <https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff> [Προσπελάστηκε 12/09/2020].

[102] Scikit-learn.org. 2020. *3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.23.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/cross_validation.html> [Προσπελάστηκε 01/10/2020].

[103] Bronshtein, A., 2017. Train/Test Split And Cross Validation In Python. [online] Towards Data Science. Available at: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6> [Προσπελάστηκε 01/10/2020].

[104] Bitdegree. 2019. A Guide On Splitting Datasets With Train_Test_Split Function. [online] Available at: <https://www.bitdegree.org/learn/train-test-split> [Προσπελάστηκε 01/10/2020].

[105] Nahrstaedt, H., 2020. Pyedflib -EDF/BDF Toolbox In Python — Pyedflib Documentation. [online] Pyedflib.readthedocs.io. Available at: <https://pyedflib.readthedocs.io/en/latest/> [Προσπελάστηκε 19/02/2020].

[106] Gokul, K., 2018. A Scalable Automated Diagnostic Feature Extraction System For Eegs. [online] GitHub. Available at: <https://github.com/gokul-krishna/EEG-Feature-Extraction> [Προσπελάστηκε 19/02/2020].

[107] Rabha, J., 2016. Feature Extraction EEG. [online] GitHub. Available at: <https://github.com/JoyRabha/Feature-Extraction-EEG> [Προσπελάστηκε 19/02/2020].

[108] Simão, V., 2017. EEG Features. [online] GitHub. Available at: <https://github.com/vancleys/EEGFeatures> [Προσπελάστηκε 19/02/2020].

[109] Agarwal, M., 2020. EEG Datasets. [online] GitHub. Available at: <https://github.com/meagmohit/EEG-Datasets> [Προσπελάστηκε 08/02/2020].

[110] Yu, J., 2019. Epileptic Seizure Classification ML Algorithms. [online] Towards Data Science. Available at: <https://towardsdatascience.com/seizure-classification-d0bb92d19962> [Προσπελάστηκε 17/11/2020].