ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εφαρμογή Αλγορίθμων Επιλογής Χαρακτηριστικών για την Ανεύρεση Βιοδεικτών με στόχο την Ταξινόμηση Αυτιστικών Ατόμων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων : Γεώργιος Κ. Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Ιωσηφίνα Μ. Μαραγκού

Αθήνα, Οκτώβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εφαρμογή Αλγορίθμων Επιλογής Χαρακτηριστικών για την Ανεύρεση Βιοδεικτών με στόχο την Ταξινόμηση Αυτιστικών Ατόμων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ιωσηφίνα Μ. Μαραγκού

Επιβλέπων : Γεώργιος Κ. Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Οκτωβρίου 2020.

.....	.....	.....
Γεώργιος Κ. Ματσόπουλος	Δημήτριος-Διονύσιος Κουτσούρης	Κωνσταντίνος Πολιτόπουλος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2020

.....
Ιωσηφίνα Μ. Μαραγκού
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Σύνοψη

Στην παρούσα διπλωματική εργασία έγινε προσπάθεια για την αύξηση της ακρίβειας της αυτοματοποιημένης διαδικασίας διαχωρισμού ατόμων που ανήκουν στο φάσμα του αυτισμού από τυπικά αναπτυσσόμενους. Η προσπάθεια αυτή βασίστηκε στην εύρεση βιοδεικτών που διαφέρουν ανάμεσα στις δύο κατηγορίες. Για την επίτευξη αυτού του στόχου χρησιμοποιούνται ως εργαλεία μέθοδοι μηχανικής μάθησης. Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία, προήλθαν από τη βάση δεδομένων ABIDE (Autism Brain Imaging Data Exchange). Πρόκειται για μια βάση δεδομένων από απεικονίσεις προερχόμενες από λειτουργική απεικόνιση μαγνητικού συντονισμού (functional Magnetic Resonance Imaging - fMRI), ατόμων που ανήκουν στο φάσμα του αυτισμού αλλά και από τυπικά αναπτυσσόμενους ανθρώπους, σε κατάσταση ηρεμίας. Οι άτλαντες που χρησιμοποιήθηκαν για την απεικόνιση των περιοχών ήταν οι εξής: Harvard Oxford (HO), Automated Anatomical Labeling (AAL) και Craddock (CC-200), ενώ αξιοποιήθηκαν τα δεδομένα των περιοχών που ανήκουν στο Δίκτυο Κατάστασης Ηρεμίας (ΔΚΗ).

Ως προς τα χαρακτηριστικά που χρησιμοποιήθηκαν, πραγματοποιήθηκε υπολογισμός της στατικής και λειτουργικής συνδεσιμότητας μεταξύ των περιοχών του εγκεφάλου που ενεργοποιούνται σε κατάσταση ηρεμίας και ανήκουν στο ΔΚΗ, καθώς και υπολογισμός των στατιστικών στιγμών τους, όπως είναι η μέση τιμή, η διακύμανση, η κυρτότητα και η στρέβλωση. Κάνοντας χρήση των παραπάνω χαρακτηριστικών, των δημογραφικών στοιχείων αλλά και πληροφοριών που σχετίζονται με τις παραμέτρους που ορίστηκαν στο μαγνητικό τομογράφο για τη λήψη, δημιουργήθηκε ένα πίνακας που περιλάμβανε αυτά τα χαρακτηριστικά, τα οποία στη συνέχεια δόθηκαν ως είσοδος σε μοντέλα μηχανικής μάθησης. Τα εργαλεία μηχανικής μάθησης που χρησιμοποιήθηκαν περιλαμβάνουν νευρωνικά δίκτυα, διάφορους αλγόριθμους ταξινόμησης και αλγόριθμους μείωσης διαστατικότητας. Επιπλέον, μελετήθηκε η επίδραση επιμέρους χαρακτηριστικών στα αποτελέσματα της ταξινόμησης. Ακόμα, πραγματοποιήθηκε μελέτη μέσω δοκιμών τις επίδρασης των διαφορετικών ατλάντων στα αποτελέσματα της ταξινόμησης.

Στην εργασία αυτή παρουσιάζονται αναλυτικά οι ταξινομητές που μελετήθηκαν, οι δοκιμές που πραγματοποιήθηκαν καθώς και τα αποτελέσματα που προέκυψαν. Στην συνέχεια παρουσιάζονται τα αποτελέσματα αλγορίθμων μείωσης διαστατικότητας όταν αυτοί εφαρμόστηκαν στα παραπάνω δεδομένα και έγινε συγκερασμός των αποτελεσμάτων τους για την εύρεση του συνόλου βιοδεικτών με την καλύτερη απόδοση ταξινόμησης. Τέλος, έγινε προσπάθεια συσχετισμού περιοχών του εγκέφαλου με τα χαρακτηριστικά με την μεγαλύτερη συχνότητα εμφάνισης μεταξύ των επιτυχημένων μοντέλων, προκειμένου να αναδειχθούν οι περιοχές που τελικά σχετίζονται με το φάσμα της αυτιστικής διαταραχής.

Λέξεις κλειδιά:
Αυτισμός, fMRI, Δίκτυο Κατάστασης Ηρεμίας, μηχανική μάθηση, μείωση διαστατικότητας, νευρωνικά δίκτυα, ταξινομητές, βιοδείκτες

Abstract

In the present thesis an attempt was performed in order to increase the accuracy of an automated process of classifying people as belonging on the Autism Spectrum Disorder (ASD) against Typically Developed (TD) ones. This effort was based on finding biomarkers that are differentiated among the two groups. To achieve this goal, machine learning methods are employed. The data that were utilized for this work originated from the ABIDE (Autism Brain Imaging Data Exchange) database. ABIDE is an initiative of functional Magnetic Resonance Imaging (fMRI) data, acquired from people belonging on the autism spectrum but also of typically developed people, at rest. The atlases that were used to map the brain regions were: Harvard Oxford (HO), Automated Anatomical Labeling (AAL) and Craddock (CC-200), while data from the areas belonging to the Default Mode Network (DMN) were utilized.

Regarding the characteristics that were employed as features, a calculation of the static and dynamic functional connectivity among the areas of the brain that are activated in resting state and belong on the DMN, as well as the calculation of their statistical moments, such as mean value, variance, skewness and kurtosis, was performed. Using the above features, demographics and information related to the parameters of the MRI acquisition protocol, a table was created containing these features, which were afterwards given as input to machine learning algorithms. Artificial intelligence tools that were utilized, include neural networks, various classification algorithms and dimensionality reduction algorithms. In addition, the effect of each characteristic on the classification results was further examined. Furthermore, a variety brain atlases were tested as for their suitability, based on the derived results of the classification, for performing the distinction process among ASD and TD subjects.

This diploma thesis presents in detail the classifiers that were utilized, the trials that were performed, as well as the results that were obtained. Following, the results of the dimensionality reduction algorithms are demonstrated when these were applied to the abovementioned data. Moreover, their results were interpreted towards finding the subset of biomarkers with the best classification performance. Finally, an attempt was made to correlate areas of the brain with the features that were most commonly seen among the subsets of the biomarkers that were derived from the feature selection process and had the finest performance, in order to highlight the areas that are ultimately related with the spectrum of autistic disorder.

Keywords:
Autism, fMRI, Default Mode Network, machine learning, dimensionality reduction, feature selection, neural networks, classifiers, biomarkers

Ευχαριστίες

Θα ήθελα να ευχαριστήσω αρχικά, τον επιβλέποντα μου, κ. Ματσόπουλο Γεώργιο, για την εμπιστοσύνη που μου έδειξε καθ’ όλη τη διάρκεια της εκπόνησης αυτής της εργασίας.

Ευχαριστώ την υποψήφια διδάκτορα κα. Κατερίνα Καραμπάση για την πολύτιμη βοήθεια της και την υποστήριξη της καθ’ όλη τη διάρκεια εκπόνησης της εργασίας αυτής.

Ακόμα, θα ήθελα να πω ένα μεγάλο ευχαριστώ στους γονείς μου Φλωρεντία και Μάριο και την αδελφή μου Κατερίνα που έπαιξαν έναν πολύ σημαντικό ρόλο στην επιτυχία μου και με στήριξαν σε κάθε βήμα μου καθ’ όλη τη διάρκεια των σπουδών μου.

Τέλος, ευχαριστώ τον Κωνσταντίνο Α., για την υπομονή και την στήριξη που μου έδειξε καθ’ όλη την διάρκεια εκπόνησης αυτής της διπλωματικής.

Η παρούσα Διπλωματική Εργασία είναι αφιερωμένη στους Ιωσήφ Μ. και Κατερίνα Α..

Περιεχόμενα

Ευρετήριο εικόνων

Ευρετήριο πινάκων

1 Εισαγωγή

Ο αυτισμός είναι μια αναπτυξιακή διαταραχή, κύρια χαρακτηριστικά της οποίας είναι η αδυναμία στην κοινωνική αλληλεπίδραση, καθώς και επαναλαμβανόμενα συμπεριφορικά μοτίβα [1], [2]. Η διάγνωσή του μέχρι και σήμερα γίνεται από το γιατρό χωρίς αυτή να υποστηρίζεται από εργαστηριακό έλεγχο ή κάποια άλλη εξέταση κι έτσι βασίζεται στις υποκειμενικές παρατηρήσεις του εκάστοτε γιατρού. Προσπάθειες για να συνδεθεί ο αυτισμός με συγκεκριμένες γενετικές ή άλλου είδους διαφορές, όπως για παράδειγμα απεικονιστικές και επομένως να επιτραπεί η δημιουργία ειδικών εξετάσεων για τον εντοπισμό αυτιστικών ατόμων δεν έχει επιτευχθεί μέχρι και σήμερα [3].

Για τη λήψη των δεδομένων που θα συντελέσουν στην διάγνωση του αυτισμού απαιτείται εκτεταμένη μελέτη του εγκέφαλου, κάτι το οποίο υποβοηθάται σημαντικά από τις απεικονιστικές μεθόδους. Στην παρούσα εργασία χρησιμοποιείται η απεικονιστική μέθοδος της λειτουργικής απεικόνισης μαγνητικού συντονισμού (functional Magnetic Resonance Imaging - fMRI). Χρησιμοποιήθηκαν fMRI δεδομένα ατόμων που βρίσκονταν σε κατάσταση ηρεμίας, διότι ήταν επιθυμητή η αξιοποίηση των περιοχών του εγκέφαλου που ενεργοποιούνται όταν ένα άτομο βρίσκεται σε ηρεμία. Επίσης, χρησιμοποιήθηκαν διάφοροι άτλαντες του εγκεφάλου αφού κάθε ένας από αυτούς απεικονίζει με διαφορετικό τρόπο τις περιοχές του εγκεφάλου. Για την εύρεση των περιοχών που αντιστοιχούν στο Δίκτυο Κατάστασης Ηρεμίας από τον κάθε άτλαντα έγινε σύγκριση με τις περιοχές του Δικτύου Κατάστασης Ηρεμίας από τους Smith et al [4]. Τα δεδομένα που προέκυψαν από την επεξεργασία αυτή, χρησιμοποιήθηκαν για την δημιουργία μοντέλων αλγορίθμων μηχανικής μάθησης.

Μέχρι σήμερα, αρκετές έρευνες έχουν στραφεί στη μηχανική μάθηση για την υποβοήθηση των ιατρικών αποφάσεων που λαμβάνονται και σχετίζονται από μια απλή διάγνωση μέχρι την πορεία ενός χειρουργείου [5]–[8]. Έτσι, και στην περίπτωση του αυτισμού, η μηχανική μάθηση έρχεται να συνδράμει με στόχο την εξάλειψη της υποκειμενικότητας και με στόχο την αντικειμενική τελική διάγνωση του αυτισμού σε άτομα από μικρές έως και μεγαλύτερες ηλικίες [9]–[12]. Για την επίτευξη του παραπάνω στόχου χρησιμοποιήθηκαν χαρακτηριστικά που εξήχθησαν από υπολογισμούς χρησιμοποιώντας τα προαναφερθέντα δεδομένα, τα οποία πλέον αντιμετωπίζονται ως πιθανοί βιοδείκτες και εν συνεχεία εισήχθησαν ως δεδομένα εισόδου σε μοντέλα μηχανικής μάθησης. Ως βιοδείκτες αναφέρονται ιατρικές ενδείξεις, μετρήσιμες εργαστηριακά, οι οποίες παράγουν τα ίδια αποτελέσματα σε κάθε καταγραφή τους. Ο αριθμός των βιοδεικτών που θα δοθούν σε ένα μοντέλο ταξινόμησης μπορεί να έχει καθοριστικό ρόλο στην ακρίβεια ταξινόμησης, αλλά και στο πόσο αξιόπιστο είναι τελικά το σύστημά μας.

Αξιοποιώντας τις παραπάνω πληροφορίες και εργαλεία, η παρούσα διπλωματική εργασία αποσκοπεί στη δημιουργία ενός μοντέλου στηριγμένου στην μηχανική μάθηση για την κατάταξη των ατόμων ως τυπικά αναπτυσσόμενων ή ως άτομα που ανήκουν στο φάσμα του αυτισμού. Πιο συγκεκριμένα, μελετήθηκε η επίδοση διαφορετικών αλγορίθμων ταξινόμησης καθώς επίσης και μεταβολές στις παραμέτρους τους με σκοπό να βρεθεί εκείνος ο συνδυασμός αλγόριθμου – παραμέτρων που θα παρουσίαζε τα καλυτέρα αποτελέσματα για τα δεδομένα που χρησιμοποιήθηκαν. Ακόμα λαμβάνοντας υπόψιν την επίδραση που παρουσιάζει στην ακρίβεια ταξινόμησης ο αριθμός των βιοδεικτών που θα χρησιμοποιηθούν,

επιλέχθηκε να γίνουν δοκιμές χρησιμοποιώντας αλγορίθμους μείωσης διαστατικότητας με σκοπό να επιλεχθούν τα χαρακτηριστηκά με την μεγαλύτερη πληροφορία για την ταξινόμηση και να αφαιρεθούν εκείνα που οδηγούσαν το σύστημα σε βεβιασμένα συμπεράσματα.

Για την δόμησή τις παρούσας εργασίας χρησιμοποιήθηκαν τα κεφάλαια 2 έως 4 για την παράθεση του θεωρητικού υποβάθρου που απαιτήθηκε για αυτήν. Πιο συγκεκριμένα, στο κεφάλαιο 2 γίνεται αναφορά στις διαθέσιμες απεικονιστικές μεθόδους και επεξηγείται ο τρόπος λειτουργίας των Magnetic Resonance Imaging (MRI) και fMRI. Στην συνέχεια, αναφέρεται η προ-επεξεργασία που υφίστανται τα δεδομένα fMRI και ορισμένες λεπτομέρειες για τα δεδομένα που θα χρησιμοποιηθούν στις μετέπειτα ενότητες. Στο κεφάλαιο 3 αυτής της διπλωματικής, γίνεται μια σύντομη αναφορά στην διαταραχή του αυτισμού αλλά και τους τρόπους διάγνωσης στην σημερινή εποχή. Ακόμα, επεξηγείται τι ονομάζεται βιοδείκτης και αναλύονται εκείνοι οι βιοδείκτες που θα χρησιμοποιηθούν ως δεδομένα στους αλγορίθμους. Στην συνέχεια, αναφέρονται οι αλγόριθμοι ταξινόμησης που θα δοκιμαστούν και δίνεται ένα βασικό θεωρητικό υπόβαθρο που απαιτείται για την καλύτερη κατανόηση αυτών. Στο κεφάλαιο 4, αναλύονται οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιήθηκαν για τις δοκιμές της παρούσας μελέτης.

Λαμβάνοντας υπόψιν και αξιοποιώντας την θεωρία των κεφαλαίων που αναφέρθηκαν, πραγματοποιήθηκαν πειράματα και εξήχθησαν συμπεράσματα τα οποία παρατίθενται στα κεφάλαια 5 και 6, αντίστοιχα. Πιο αναλυτικά, στο κεφάλαιο 5 περιγράφονται όλα τα πειράματα που υλοποιήθηκαν και παρατίθενται τα αποτελέσματα που προέκυψαν από αυτά. Τα πειράματα αυτά αφορούν δοκιμές ταξινομητών και ενός νευρωνικού δικτύου τα οποία πραγματοποιούνται υπό διαφορετικές συνθήκες με την έννοια των διαφορετικών παραμέτρων αλλά και δεδομένων εισόδου, με σκοπό την διερεύνηση του πώς και αν επηρεάζονται τα τελικά αποτελέσματα από τις διαφορετικές ομάδες χαρακτηριστικών. Ακόμα μελετήθηκαν τα αποτελέσματα ταξινομητών που λαμβάνουν ως είσοδο τα χαρακτηριστικά που έχουν επιλεγεί από αλγορίθμους μείωσης διαστατικότητας. Τέλος, στο κεφάλαιο 6 παρατίθενται τα συμπεράσματα που προέκυψαν ύστερα από μελέτη των αποτελεσμάτων. Έγινε, επίσης, διερεύνηση δυνητικών βιοδεικτών – με την έννοια των περιοχών του εγκεφάλου – οι οποίοι πιθανά συνδέονται με τον αυτισμό, ύστερα από παρατήρηση των χαρακτηριστικών που επιλέχθηκαν από τους αλγορίθμους μείωσης διαστατικότητας και οι οποίοι έδωσαν τα καλύτερα τελικά αποτελέσματα στην ταξινόμηση μεταξύ αυτιστικών και τυπικά αναπτυσσόμενων ατόμων.

2 Απεικονιστικές Μέθοδοι

Η μη επεμβατική χαρτογράφηση της λειτουργίας και της δομής του εγκεφάλου καθίσταται δυνατή μέσω τεχνικών απεικόνισης αυτού. Η χαρτογράφηση επιτυγχάνεται είτε μέσω της μέτρησης συγκεκριμένων αποκρίσεων ιστών σε μια εξωτερική πηγή ενέργειας όπως ένα μαγνητικό πεδίο, είτε με απευθείας μέτρηση των ρευμάτων και των μαγνητικών πεδίων που παράγονται από την εγκεφαλική δραστηριότητα, με έγχυση ραδιοϊσοτόπων για να περιγράψουν περιοχές μέσω εκπεμπόμενης ακτινοβολίας [13]. Οι τεχνικές απεικόνισης εγκεφάλου μπορούν να διαχωριστούν σε δομικές και λειτουργικές. Η πρώτη κατηγορία προσδιορίζει τις ανατομικές περιοχές, ενώ η δεύτερη το μεταβολισμό και τη φυσιολογία των περιοχών ενδιαφέροντος [14], [15]. Μερικά παραδείγματα τεχνικών δομικής απεικόνισης αποτελούν η αξονική τομογραφία ή αλλιώς υπολογιστική τομογραφία (Computed Tomography - CT) [16], οι υπέρηχοι Doppler [17], η αγγειογραφία [18], η μυελογραφία [18] και η απεικόνιση μαγνητικού συντονισμού (Magnetic Resonance Imaging - MRI) [18]. Στην κατηγορία των τεχνικών λειτουργικής απεικόνισης ανήκουν οι τεχνικές της ποσοτικής ηλεκτροεγκεφαλογραφίας (Quantitative Electroencephalography - QEEG), της λειτουργικής απεικόνισης μαγνητικού συντονισμού (functional Magnetic Resonance Imaging ή functional MRI - fMRI), της τομογραφίας με εκπομπή ενός φωτονίου (Single-photon Emission-Computed Tomography - SPECT), της τομογραφίας εκπομπών ποζιτρονίων (Positron-Emission Tomography - PET), της μαγνητοεγκεφαλογραφίας (Magnetoencephalography - MEG), της φασματοσκοπίας μαγνητικού συντονισμού (Magnetic Resonance Spectroscopy - MRS) [14], [15], [19], [20].

Οι παραπάνω τεχνικές βασίζονται σε διάφορες αρχές για την λειτουργία τους. Συγκεκριμένα οι MEG, MRS, MRSI και η MRI χρησιμοποιούν μαγνητικά πεδία, το QEEG χρησιμοποιεί πληροφορία από την ηλεκτρική δραστηριότητα, η αγγειογραφία, η μυελογραφία και η υπολογιστική τομογραφία χρησιμοποιούν ακτίνες X, το SPECT και το PET χρησιμοποιούν ραδιενέργεια και τέλος το καρωτιδικό Doppler και το διακρανιακό Doppler χρησιμοποιούν υπέρηχους [15], [19].

2.1 Μαγνητικός Συντονισμός

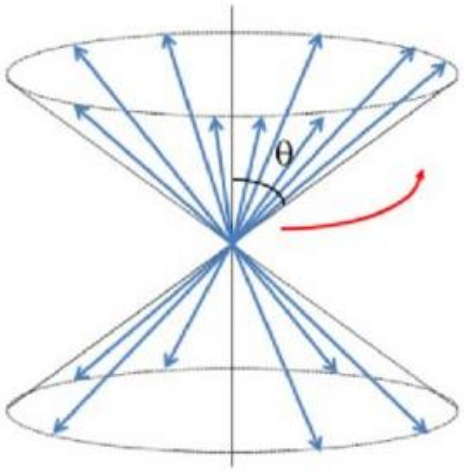
Στην συνέχεια θα αναλυθούν οι απεικονιστικές μέθοδοι MRI και fMRI έχοντας ως στόχο την κατανόηση των φυσικών αρχών στις οποίες στηρίζεται η λειτουργία τους.

2.1.1 Μαγνητική Τομογραφία

Η απεικόνιση μαγνητικού συντονισμού (MRI) χρησιμοποιείται συχνά για την ανίχνευση και διάγνωση ασθενειών καθώς και για την παρακολούθηση θεραπειών. Αυτή η μη επεμβατική τεχνική απεικόνισης παράγει τρισδιάστατες λεπτομερείς ανατομικές εικόνες σε μια συγκεκριμένη χρονική στιγμή. Βασίζεται στη διέγερση και ανίχνευση της αλλαγής της κατεύθυνσης του άξονα ιδιοστροφορμής των πρωτονίων υδρογόνου (H^+) που βρίσκονται σε μόρια νερού μέσα στο σώμα [21]. Χρησιμοποιούνται ειδικά τα πρωτόνια υδρογόνου λόγω της μεγάλης ποσότητας νερού στο ανθρώπινο σώμα και της δυνατότητας μαγνήτισής τους άνευ αρνητικών

επιπτώσεων, αλλά θεωρητικά μπορεί να χρησιμοποιηθεί οποιοδήποτε άτομο με ιδιοστροφορμή μεγαλύτερη του ½ [22].

Πιο συγκεκριμένα, κατά την διάρκεια μιας εξέτασης MRI ασκείται ισχυρό ομοιόμορφο μαγνητικό πεδίο παράλληλα με το σώμα το οποίο οδηγεί στην αλλαγή προσανατολισμού του άξονα ιδιοστροφορμής των πρωτονίων υδρογόνου μορίων νερού μέσα στο σώμα, ώστε αυτός να περιστρέφεται υπό σταθερή γωνία ϑ γύρω από άξονα του μαγνητικού πεδίου (είτε στην ίδια, είτε στην αντίθετη διεύθυνση σύμφωνα με το επιβαλλόμενο μαγνητικό πεδίο) όπως φαίνεται και στην Εικόνα 1.



Εικόνα 1: Περιστροφή αξόνων ιδιοστροφορμής κατά την εφαρμογή μαγνητικού πεδίου [22]

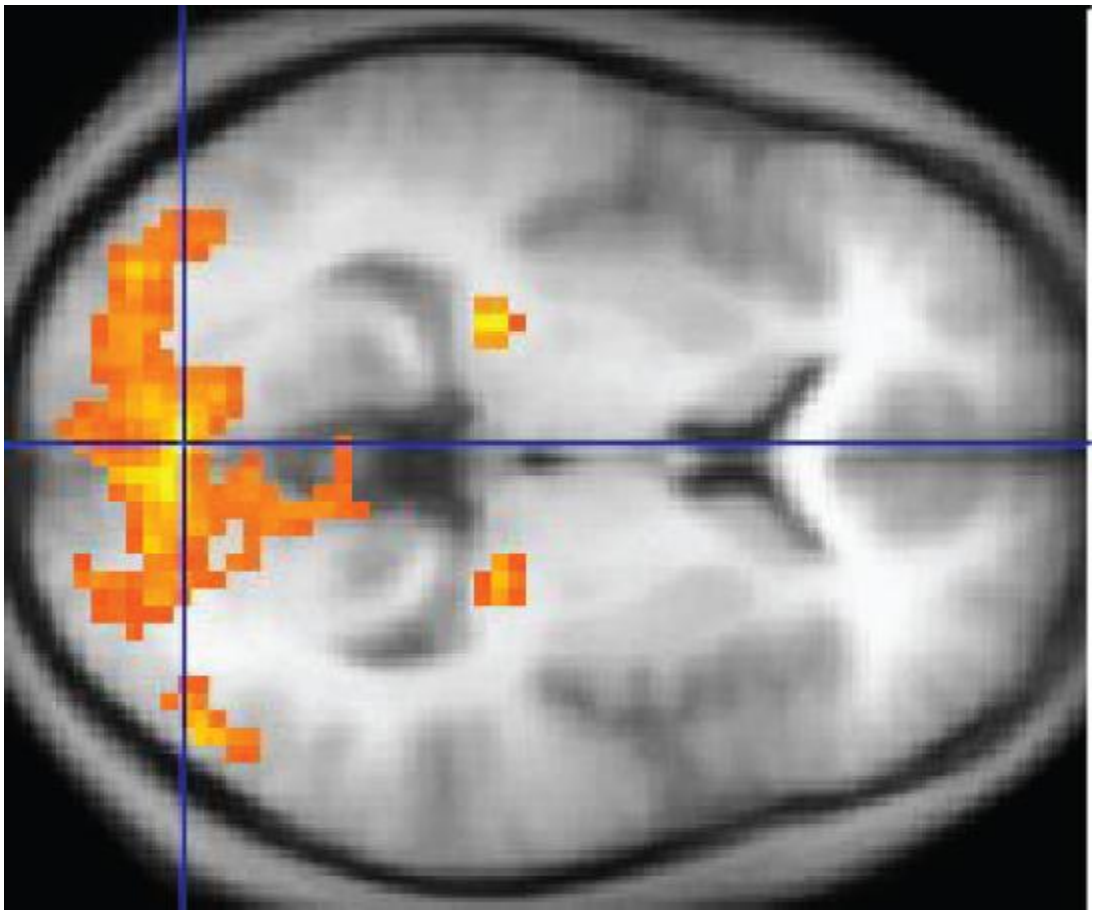
Στη συνέχεια, παράγεται μαγνητικός παλμός μικρής διάρκειας ο οποίος θα δώσει την ενέργειά του σε ένα περιστρεφόμενο πρωτόνιο, ώστε αυτό να μεταβεί από την κατάσταση χαμηλής ενέργειας (στην οποία βρίσκεται κατά την κατάσταση ηρεμίας του), στην κατάσταση υψηλής ενέργειας. Μόλις ο παλμός αυτός σταματήσει να εφαρμόζεται στα άτομα, το πρωτόνιο τείνει να αποδώσει την ενέργεια που έλαβε με τη μορφή ηλεκτρομαγνητικού κύματος. Η συχνότητα του παλμού αυτού είναι τέτοια ώστε να γίνεται επιλεκτικά διέγερση μόνο ατόμων συγκεκριμένων ουσιών/ιστών. Το κύμα αυτό που εκπέμπεται – φωτόνιο - ανιχνεύεται και έπειτα μετατρέπεται σε εικόνα από κατάλληλους αισθητήρες της συσκευής. Λόγω των διαφορετικών χαρακτηριστικών κάθε τύπου ιστού, η ενέργεια που θα απελευθερωθεί και άρα η ένταση του φωτονίου που θα εκπέμψει ο εκάστοτε ιστός, θα διαφέρει με τρόπο μοναδικό ανάλογα με τον είδος του. Η προσεκτική καταγραφή της εκπομπής αυτής επιτρέπει την δημιουργία εικόνας της υπό εξέτασης περιοχής. Με αυστηρά χωρικά ελεγχόμενη επιβολή μαγνητικών παλμών και μετέπειτα καταγραφή του εκπεμπόμενου σήματος είναι δυνατή η λήψη δεδομένων σχετικά με κάθε σημείο της υπό εξέτασης περιοχής [23], [24].

Εικόνα 2 - Συσκευή MRI στο ιατρικό κέντρο της στρατιωτικής βάσης 'Joint Operating Base', Bastion, Afghanistan (Κωδικός δημοσίευσης US Navy: 111006-O-KK908-026)

2.1.2 Λειτουργική Μαγνητική Τομογραφία - Σήμα BOLD

Η Λειτουργική Απεικόνιση Μαγνητικού Συντονισμού (functional Magnetic Resonance Imaging – fMRI) χρησιμοποιεί τα εργαλεία της MRI και αναπτύχθηκε με σκοπό την παροχή μη στατικών εικόνων του εγκέφαλου καθορισμένης διάρκειας. Διεξοδικότερα, τα δεδομένα για την κατασκευή της εικόνας MRI λαμβάνονται χρησιμοποιώντας την ενέργεια που απελευθερώνεται από τα πρωτόνια σε μορφή ηλεκτρομαγνητικού κύματος κατά την επιβολή μαγνητικού πεδίου και του μετέπειτα παλμού που δίδεται στα άτομα του υδρογόνου. Σε αντίθεση με την MRI, η fMRI βασίζεται στο οξυγονωμένο αίμα για την λήψη της εικόνας. Η ροή του αίματος ανάμεσα στις περιοχές που είναι ενεργοποιημένες και σε αυτές που δεν είναι διαφέρει. Συγκεκριμένα, οι περιοχές οι οποίες είναι ενεργοποιημένες παρουσιάζουν μεγαλύτερη ροή αίματος από τις περιοχές που δεν είναι, καθώς η ενεργοποίηση μιας περιοχής οδηγεί σε αυξημένες ενεργειακές ανάγκες και επομένως μεγαλύτερη ανάγκη για οξυγόνο [25]. Κατά την αποδέσμευση οξυγόνου από το κύριο συστατικό του αίματος, της αιμογλοβίνης, αυτή μετατρέπεται από διαμαγνητική σε παραμαγνητική. Η μεταβολή αυτή των μαγνητικών της ιδιοτήτων είναι εύκολα ανιχνεύσιμη από το μαγνητικό τομογράφο. Αυτός ο τύπος απόκρισης αποκαλείται απόκριση εξαρτώμενη από το επίπεδο οξυγόνωσης αίματος (Blood Oxygenation Level Dependent response - BOLD) [26]. Στην Εικόνα 3 εμφανίζεται το αποτέλεσμα μιας εξέτασης fMRI. Συγκεκριμένα στην εικόνα εμφανίζεται το αποτέλεσμα της στατιστικής ανάλυσης του σήματος BOLD σε όλο τον εγκέφαλο και όχι το σήμα αυτό καθ' αυτό. Με πορτοκαλί χρώμα εμφανίζονται οι περιοχές που ενεργοποιούνται, δηλαδή οι περιοχές με αυξημένη αιμάτωση, κάτι το οποίο επηρεάζει τις τοπικές μαγνητικές ιδιότητες.

Η fMRI στον εγκέφαλο μετράει την εγκεφαλική δραστηριότητα ανιχνεύοντας τις μεταβολές στην οξυγόνωση και στον εγκεφαλικό μεταβολισμό οι οποίες οφείλονται σε νευρική δραστηριότητα [27], [28]. Μέσω της fMRI παρέχεται η δυνατότητα να ανιχνευθούν περιοχές του εγκεφάλου που παρουσιάζουν λειτουργική συνδεσιμότητα [29]. Με τον όρο αυτό αναφέρεται η κοινή, δηλαδή ταυτόχρονη ενεργοποίηση περιοχών του εγκεφάλου οι οποίες μπορεί να είναι και απομακρυσμένες η μία από την άλλη, να μη συνδέονται δηλαδή άμεσα δομικά. Η δραστηριότητα αυτή μπορεί να είναι κάποια συγκεκριμένη εργασία (πχ ανάγνωση κειμένου, ακρόαση μουσικής, κτλ) ή ακόμα και απλά να παραμένει το άτομο σε κατάσταση ηρεμίας [30].



Εικόνα 3: Περιοχές εγκεφάλου που παρουσιάζουν αυξημένη ενεργοποίηση. Αποτέλεσμα εξέτασης fMRI.

2.2 Προεπεξεργασία δεδομένων fMRI

2.2.1 Βήματα προεπεξεργασίας fMRI δεδομένων

Έχοντας ως μέλημα τη βελτίωση της ποιότητας των δεδομένων που έχουν ληφθεί γίνεται λειτουργική και δομική προεπεξεργασία αυτών. Η προεπεξεργασία αυτή περιλαμβάνει τα παρακάτω στάδια:

I. Έλεγχος ποιότητας (Quality check)
II. Αφαίρεση ψευδοενδείξεων (Artifact correction)

III. Διόρθωση παραμόρφωσης λόγω μαγνητικών πεδίων (Distortion correction)
IV. Συγχρονισμός των τομών στο πεδίο του χρόνου (Slice timing correction)
V. Διόρθωση των παραμορφώσεων λόγω κινήσεων (Motion correction)
VI. Χωρική ομαλοποίηση (Spatial normalization)
VII. Χωρική εξομάλυνση (Spatial smoothing)
VIII. Φιλτράρισμα στο πεδίο του χρόνου (Temporal filtering)

Αναλυτικότερα, αρχικά γίνεται έλεγχος ποιότητας στα καταγεγραμμένα δεδομένα. Σε 2ο στάδιο γίνεται αφαίρεση του θορύβου για να αποφευχθεί η παρουσία ψευδών ενδείξεων στο λαμβανόμενο σήμα με σκοπό να απομονωθεί το σήμα που παράχθηκε από τον εγκέφαλο από αυτό που οφειλόταν σε πιθανούς θορύβους (για παράδειγμα κίνηση, θόρυβος από τα μαγνητικά πεδία κ.ά.). Στο 3ο στάδιο διορθώνονται οι παραμορφώσεις που ίσως να υπήρξαν στις fMRI εικόνες, λόγω των μαγνητικών πεδίων που επιβάλλονται. Στη συνέχεια, γίνεται συγχρονισμός των τομών στο πεδίο του χρόνου προκειμένου να εξαλειφθούν προβλήματα ετεροχρονισμού. Αυτό γίνεται, καθώς η fMRI είναι μία λήψη που λαμβάνει απεικόνιση από κάθε τομή του εγκεφάλου σε διακριτές χρονικές στιγμές. Έτσι, προκειμένου να υπάρχει το σωστό σήμα για όλες τις τομές σε μία χρονική στιγμή, απαιτείται χρονική διόρθωση. Στο επόμενο στάδιο γίνεται η διόρθωση των παραμορφώσεων λόγω κίνησης του εξεταζόμενου ατόμου σε κάθε τομή. Η ανάγκη αυτή προέκυψε διότι η απεικονιστική τεχνική fMRI είναι μια εξέταση που απαιτεί πλήρη ακινησία για αρκετό χρονικό διάστημα, πράγμα που δεν είναι δυνατόν να επιτευχθεί εύκολα. Στο 6ο στάδιο πραγματοποιείται κανονικοποίηση των δεδομένων σε έναν κοινό χώρο δηλαδή μια κοινή εικόνα εγκεφάλου. Ο χώρος που χρησιμοποιείται συνήθως είναι ο MNI152. Ο MNI152 βασίζεται στην απεικόνιση ενός προτύπου εγκεφάλου, όπως προέκυψε από τον μέσο όρο δομικών εικόνων MRI από 152 νεαρά υγιή ενήλικα άτομα. Ο σκοπός αυτής της ενέργειας είναι παροχή ενός κοινού σημείου αναφοράς για να υπάρχει η δυνατότητα συγκρίσεων τόσο μεταξύ ατόμων αλλά και μεταξύ ερευνών. Στο στάδιο 7 οι εικόνες ομαλοποιούνται χωρικά με στόχο την περαιτέρω μείωση του θορύβου. Τέλος, γίνεται χρήση ενός ζωνοπερατού φίλτρου προκειμένου να διατηρηθούν μόνο οι χαμηλές συχνότητες καθώς αυτές είναι που ανταποκρίνονται στο σήμα που θέλουμε να κρατήσουμε [31]–[34].

2.2.2 Βάση Δεδομένων Αυτιστικών Ατόμων

Η ABIDE (Autism Brain Imaging Data Exchange) είναι μια βάση δεδομένων η οποία περιέχει απεικονίσεις εγκεφάλου ατόμων που ανήκουν στο φάσμα του αυτισμού αλλά και τυπικά αναπτυσσόμενων. Η βάση αυτή δημιουργήθηκε από 17 κλινικές και αποτελείται από δεδομένα fMRI σε κατάσταση ηρεμίας 1034 ατόμων εκ των οποίων στην παρούσα διπλωματική γίνεται χρήση των 868. Στην παρούσα διπλωματική αξιοποιήθηκαν τα άτομα που δεν είχαν κουνηθεί περισσότερο από 0.2mm κατά τη διάρκειας της λήψης, όπως υπολογίστηκε κατά μέσο όρο, από κάθε τομή της λήψης. Από τα άτομα που εντέλει χρησιμοποιήθηκαν, τα 403 ανήκαν στο φάσμα του αυτισμού και τα 465 ήταν τυπικά αναπτυσσόμενα. Ο μέσος όρος ηλικίας αυτών των ατόμων είναι τα 14.7 έτη και το εύρος της ηλικίας τους κυμαίνεται ανάμεσα στα 7 με 64 χρόνια [35]. Η ABIDE περιλαμβάνει αποτελέσματα τα οποία προέρχονται από τη διαδικασία προ-επεξεργασίας των δεδομένων, ώστε αυτά να είναι άμεσα διαθέσιμα σε ερευνητικές ομάδες με χαμηλή υπολογιστική ισχύ στις

δομές τους (ABIDE I Preprocessed). Τέσσερα λογισμικά χρησιμοποιήθηκαν για την πραγματοποίησή της προεπεξεργασίας των δεδομένων της: το Connectome Computation System (CCS) [36], το Configurable Pipeline for the Analysis of Connectomes (CPAC) [37], το Data Processing Assistant for Resting-State fMRI (DPARSF) [38] και το NeuroImaging Analysis Kit (NIAK) [39], [40]. Ειδικότερα, το λογισμικό Configurable Pipeline for the Analysis of Connectomes (CPAC) χρησιμοποιεί τα εργαλεία AFNI [41], ANTs [42], FSL [43] και προσαρμοσμένο κώδικα σε γλώσσα python [37], ενώ περιλαμβάνει όλα τα στάδια που αναφέρθηκαν για την λειτουργική προεπεξεργασία των δεδομένων fMRI στην προηγούμενη υποενότητα. Τα δεδομένα υποβλήθηκαν πέρα από την λειτουργική προεπεξεργασία και σε δομική. Συγκεκριμένα, κατά την δομική προεπεξεργασία έγινε η αφαίρεση του κρανίου και χωρίστηκε ο ιστός του εγκέφαλου σε 3 τύπους (μέσω του εργαλείου FAST της βιβλιοθήκης FSL), χωρίς να αφαιρεθούν τα πρώτα volumes. Διορθώθηκε ο χρονισμός των διαφόρων τομών και αφαιρέθηκε πιθανή κίνηση των εξεταζόμενων, ενώ ομαλοποιήθηκε η ένταση της φωτεινότητας της εικόνας. Οι ανά υποκείμενο διαχωρισμοί στον ιστό του εγκεφάλου περιορίστηκαν με χρήση του εργαλείου FSL και έγινε η προσαρμογή των δεδομένων σε έναν κοινό χώρο, ο οποίος ήταν ο MNI152 με χρήση γραμμικών και μη προσαρμογέων της βιβλιοθήκης ANTs [32].

2.3 Άτλαντες

Ένας τρόπος για να προσδιοριστεί μια περιοχή ενδιαφέροντος με τελικό σκοπό τη διερεύνησή της, είναι να χρησιμοποιηθεί ένας άτλαντας δηλαδή ένας «χάρτης» που χωρίζει τον εγκέφαλο σε ανατομικά διακριτές περιοχές [44]. Κάθε άτλαντας αποτελείται από συνεχόμενα, περιεκτικά αποτελέσματα της οπτικής χαρτογράφησης του εγκεφάλου και μπορεί να περιλαμβάνει ανατομικά, γενετικά ή λειτουργικά χαρακτηριστικά [45], [46]. Μερικά παραδείγματα ατλάντων, των οποίων οι διαχωρισμοί των περιοχών του εγκεφάλου παρέχονται και από τη βάση δεδομένων ABIDE, είναι ο Harvard Oxford (HO), ο Automated Anatomical Labeling (AAL) και ο Craddock (CC-200). Με γνώμονα το να χρησιμοποιηθούν μόνο οι περιοχές του ΔΚΗ, αφαιρέθηκαν εκείνα τα άτομα που είχαν έστω και μία από τις χρονοσειρές από τις περιοχές ενδιαφέροντος μηδενική. Αυτό συνέβη καθώς μηδενική στήλη θα οδηγούσε σε μηδενική συσχέτιση με οποιαδήποτε άλλη περιοχή κι επομένως σε εμφάνιση καμίας συσχέτισης.

Σχετικά με τα δεδομένα fMRI ατόμων σε κατάσταση ηρεμίας που παρέχονται από τη βάση δεδομένων ABIDE, μετά από την αφαίρεση των ατόμων που κουνήθηκαν αρκετά κατά την διαδικασία αλλά και των ατόμων που είχαν κάποια μηδενική χρονοσειρά υπάρχουν αναλυτικά:

- για τον άτλαντα HO 868 άτομα εκ των οποίων τα 403 παρουσίαζαν διαταραχές αυτιστικού φάσματος, ενώ τα 465 ήταν τυπικά αναπτυσσόμενα,
- για τον άτλαντα AAL 736 άτομα εκ των οποίων τα 324 παρουσίαζαν διαταραχές αυτιστικού φάσματος, ενώ τα 412 ήταν τυπικά αναπτυσσόμενα,
- για τον άτλαντα CC-200 856 άτομα εκ των οποίων τα 391 παρουσίαζαν διαταραχές αυτιστικού φάσματος, ενώ τα 465 ήταν τυπικά αναπτυσσόμενα.

2.4 fMRI σε Κατάσταση ηρεμίας / εργασίας

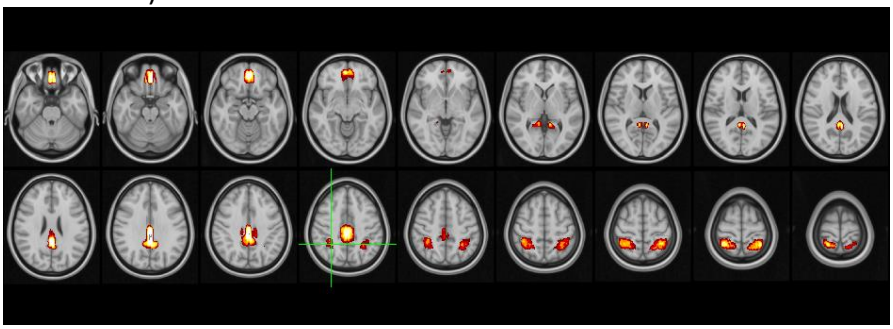
Στην fMRI με εργασία (task-based), στόχος είναι να συνδεθεί η ενεργοποίηση περιοχών του εγκεφάλου με την εκτέλεση συγκεκριμένων εργασιών [47]. Για την επίτευξη αυτού, μελετάται η αιμάτωση των αγγείων κατά την εκτέλεση μιας σαφώς καθορισμένης εργασίας ενώ ταυτόχρονα πραγματοποιείται η εξέταση [48]. Εν αντιθέσει, η εξέταση fMRI σε κατάσταση ηρεμίας (resting state fMRI – rs-fMRI), μελετά την εγκεφαλική δραστηριότητα απουσία οποιασδήποτε εργασίας. Κατά την διάρκεια της εξέτασης εμφανίζεται στο εξεταζόμενο άτομο η εικόνα ενός σταυρού στο κέντρο μιας οθόνης και του ζητείται να εστιάσει την προσοχή του σε αυτόν. Η συγκέντρωση του ατόμου εκεί, επιτρέπει να ελαχιστοποιηθεί η απόσπαση της προσοχής του καθ' όλη την διάρκεια της εξέτασης. Εναλλακτικά, του ζητείται να κλείσει τα μάτια του και να προσπαθήσει να μη σκέφτεται τίποτα [49], [50]. Παρότι το άτομο βρίσκεται σε ηρεμία, παρατηρούνται συνεχείς αυθόρμητες και αργές διακυμάνσεις στα μοτίβα ενεργοποίησής του εγκεφάλου. Η fMRI σε κατάσταση ηρεμίας επικεντρώνεται στην ανάλυση των διακυμάνσεων αυτών [51]. Επιπλέον, τα δεδομένα που λαμβάνονται σε κατάσταση ηρεμίας μπορεί να αποτελούν δυνητικούς βιοδείκτες [52].

2.5 ΔΚΗ σε κατάσταση ηρεμίας αυτιστικών

Το Δίκτυο Προεπιλεγμένης Λειτουργιάς ή Δίκτυο Κατάστασης Ηρεμίας (Default Mode Network – DMN) αντιπροσωπεύει μια ομάδα περιοχών του εγκεφάλου που παρουσιάζουν υψηλή ενεργητικότητα τις χρονικές στιγμές που ο εγκέφαλος του ατόμου βρίσκεται σε κατάσταση ηρεμίας, συγκριτικά με τις στιγμές που βρίσκεται σε νοητική δραστηριότητα [53], [54]. Το DMN αποτελείται από τις αμφίπλευρες περιοχές του κάτω – πλευρικού - βρεγματικού και εμπρόσθιου - μεσαίου φλοιού (inferior-lateral-parietal and ventro-medial frontal cortex), την οπίσθια περιοχή του προσαγωγίου (posterior cingulate) και το προσφηνοειδές λοβίο (precuneus) [55].

Στον άτλαντα Harvard Oxford (HO) το DMN, όπως φαίνεται και στην Εικόνα 4, περιλαμβάνει τις παρακάτω 6 περιοχές:

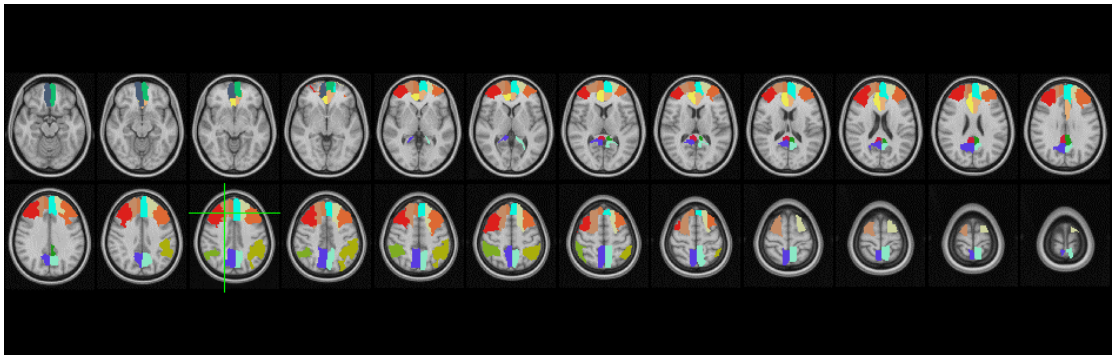
- Τη δεξιά και αριστερή περιοχή του μέσου μετωπιαίου φλοιού (Frontal Medial Cortex).
- Τη δεξιά και αριστερή οπίσθια περιοχή του προσαγωγίου (cingulate gyrus – posterior division).
- Τη δεξιά και αριστερή περιοχή του άνω βρεγματικού λοβίου (Superior Parietal Lobule).



Εικόνα 4: Οι περιοχές του άτλα HO που ανήκουν στο DMN.

Στον άτλαντα Automated Anatomical Labeling (AAL) το DMN, όπως φαίνεται και στην Εικόνα 5, αποτελείται από τις παρακάτω 16 περιοχές:

- Τη δεξιά και αριστερή περιοχή στο προσφηνοειδές λοβίο (precuneus).
- Τις δεξιές και αριστερές περιοχές του προσαγωγίου άνω και κάτω (anterior & posterior cingulate gyrus).
- Την άνω μετωπιαία έλικα μέση (δεξιά και αριστερή) και κογχική (δεξιά και αριστερή) περιοχή (superior frontal gyrus medial & medial orbital area).
- Τη δεξιά και αριστερή περιοχή της μέσης μετωπιαίας έλικας (middle frontal gyrus).
- Τη δεξιά και αριστερή περιοχή της άνω μετωπιαίας έλικας (superior frontal gyrus).
- Τη δεξιά και αριστερή περιοχή του κάτω βρεγματικού λοβίου (inferior parietal lobule).

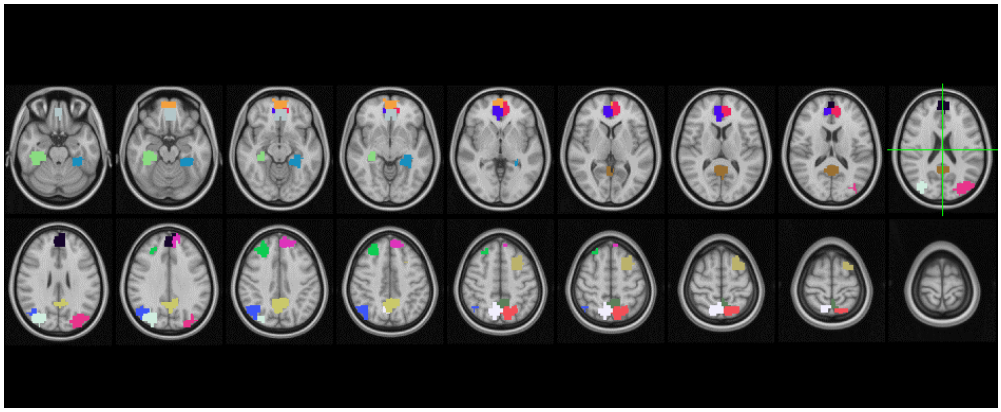


Εικόνα 5 : Οι περιοχές του άτλα AAL που ανήκουν στο DMN.

Στον άτλαντα Craddock (CC-200) το DMN, όπως φαίνεται και στην Εικόνα 6,περιλαμβάνει τις παρακάτω 18 περιοχές:

- Αριστερή μετωπική περιοχή (Frontal_L) με κωδικό περιοχής 133,
- Δεξιά μετωπική περιοχή (Frontal_R) με κωδικό περιοχής 106,
- Αριστερή μετωπική περιοχή (Frontal_L) με κωδικό περιοχής 95,
- Μεσαία γραμμή μετωπικής περιοχής(Frontal_L/R) με κωδικό περιοχής 91,
- Μεσαία γραμμή μετωπικής περιοχής (Frontal_L/R) με κωδικό περιοχής 109,
- Μεσαία γραμμή μετωπικής περιοχής (Frontal_L/R) με κωδικό περιοχής 51,
  - Δεξιά μετωπική περιοχή (Frontal_R) με κωδικό περιοχής 22,
  - Αριστερή μετωπική περιοχή (Frontal_L) με κωδικο περιοχής 5,
  - Δεξιά πάρα-ιπποκάμπεια περιοχή (Parahippo_R) με κωδικό περιοχής 62,
  - Αριστερή πάρα-ιπποκάμπεια περιοχή (Parahippo_L) με κωδικό περιοχής 122,
  - Δεξιός ινιακός λοβός (Occipital_R) με κωδικό περιοχής 170,
  - Δεξιός ινιακός λοβός (Occipital_R) με κωδικό περιοχής 166,
  - Αριστερός ινιακός λοβός (Occipital_L) με κωδικό περιοχής 97,
  - Αριστερή περιοχή προσφηνοειδούς λοβίου (Precuneus_L) με κωδικό περιοχής 136,
  - Δεξιά περιοχή προσφηνοειδούς λοβίου (Precuneus_R) με κωδικό περιοχής 163,
  - Αριστερή περιοχή προσφηνοειδούς λοβίου (Precuneus_L) με κωδικό περιοχής 197,
  - Δεξιά περιοχή προσφηνοειδούς λοβίου (Precuneus_R) με κωδικό περιοχής 174,

- Περιοχή του παρά-προσαγωγίου(Paracingulate με κωδικό περιοχής 58.



Εικόνα 6: Οι περιοχές του άτλα CC-200 που ανήκουν στο DMN.

3 Αυτισμός και Μηχανική Μάθηση

Η διαταραχή του φάσματος του αυτισμού (Autism Spectrum Disorder – ASD) είναι μια ευρεία ομάδα νευροαναπτυξιακών διαταραχών, η οποία παρουσιάζει αλλαγές στη λειτουργία του αναπτυσσόμενου εγκεφάλου του ατόμου. Τα άτομα που παρουσιάζουν αυτή τη διαταραχή έχει παρατηρηθεί ότι έχουν ελλιπή λειτουργικότητα σε ικανότητες όπως η κοινωνική αλληλεπίδραση και έχουν την τάση να παρουσιάζουν περιορισμένου αριθμού και επαναλαμβανόμενα μοτίβα συμπεριφοράς [1]. Αν και σε κάποια άτομα η διάγνωση μπορεί να γίνει πολύ νωρίς (ακόμα και από τους πρώτους 12 μήνες ζωής) δεν υπάρχει συγκεκριμένη ηλικία, ενώ διαφοροποιείται από άτομο σε άτομο καθώς οι ενδείξεις και η ηλικία εμφάνισης αυτών διαφέρει [56]–[58].

Σήμερα, η διάγνωση για το αν ανήκει ένα άτομο στο φάσμα του αυτισμού πραγματοποιείται από τον γιατρό σύμφωνα με τις παρατηρήσεις του και με τη βοήθεια ερωτηματολογίων. Πιο συγκεκριμένα, το Διαγνωστικό και Στατιστικό Εγχειρίδιο Ψυχικών Διαταραχών (Diagnostic and Statistical Manual of Mental Disorders – DSM-V) είναι το εγχειρίδιο εκείνο που αποτελεί γνώμονα για τη διάγνωση του ατόμου να ανήκει στο φάσμα του αυτισμού, ταυτόχρονα με συγκεκριμένα ερωτηματολόγια. Ο τρόπος αυτός έχει σαν αποτέλεσμα η διάγνωση να είναι υποκειμενική και να επηρεάζεται από τον εκάστοτε γιατρό που πραγματοποιεί την εξέταση. Εμφανίζεται, λοιπόν, η ανάγκη λήψης έγκυρων και αξιόπιστων διαγνώσεων απαλλαγμένων από υποκειμενικά κριτήρια για την αποφυγή σφαλμάτων και την εξασφάλιση της ίσης διαχείρισης των ατόμων στο κοινωνικό και εργασιακό τους περιβάλλον. Η έλλειψη έστω και ενός δείκτη για την κατάταξη ενός ατόμου στο φάσμα του αυτισμού (όπως για παράδειγμα μια βιοχημική εξέταση αίματος), έστρεψε το ενδιαφέρον των ερευνητών στην δημιουργία μοντέλων μηχανικής μάθησης τα οποία βασιζόμενα σε σειρά βιοδεικτών εξαγόμενων από απλές σχετικά εξετάσεις, θα μπορούσαν να απλοποιήσουν ή ακόμα και να αυτοματοποιήσουν την διάγνωση αυτής της διαταραχής [11], [12], [59].

Η Μηχανική Μάθηση (Μ.Μ. / Machine Learning – ML) αποτελεί κλάδο της τεχνητής νοημοσύνης (Τ.Ν. / Artificial Intelligence – AI) και έχει σκοπό τον προγραμματισμό ενός υπολογιστή με στόχο τη βελτιστοποίηση της λειτουργίας ενός αλγορίθμου βάσει κριτήριων μέτρησης επιδόσεων, χρησιμοποιώντας δεδομένα από παραδείγματα ή προηγούμενα δεδομένα που έχει επεξεργαστεί ο αλγόριθμος. Για την επίτευξη του παραπάνω σκοπού η Μ.Μ. χρησιμοποιεί θεωρία στατιστικής για την δημιουργία μαθηματικών μοντέλων με τελικό στόχο την εξαγωγή συμπερασμάτων από πολλαπλά δείγματα. Στην προκειμένη περίπτωση, δείγματα αποτελούν τα άτομα, τα δεδομένα των οποίων χρησιμοποιούνται στην παρούσα έρευνα. Η Μ.Μ. αποτελεί σημαντική βοήθεια στην εύρεση λύσεων σε πολλά προβλήματα αναγνώρισης εικόνας ή/και φωνής, ρομποτικής κ.ά. [60]. Πλέον η συμβολή της Μ.Μ. έγκειται μέχρι και στη χρήση και λειτουργία ρομποτικών βραχιόνων σε χειρουργεία [61] αλλά και σε άλλες εξίσου σημαντικές εκφάνσεις της καθημερινότητας. Ένα παράδειγμα μοντέλων που βασίζονται στη Μ.Μ. αποτελούν οι αλγόριθμοι ταξινόμησης (ταξινομητές – classifiers) όπου ένας ταξινομητής χρησιμοποιεί τα δεδομένα εισόδου του για προσαρμογή βαρών και γενικότερα του τρόπου λειτουργίας του (εκμάθηση)

ούτως ώστε σε μετέπειτα εκτελέσεις του και βασιζόμενος σε αυτές τις παραμετροποιήσεις να μπορέσει επιτυχώς να διενεργήσει το διαχωρισμό για τον οποίο έχει εκπαιδευτεί, σε δεδομένα που δεν έχει ξαναδεί.

3.1 Αυτισμός

3.1.1 Διαγνωστικό και Στατιστικό Εγχειρίδιο Ψυχικών Διαταραχών

Το Διαγνωστικό και Στατιστικό Εγχειρίδιο Ψυχικών Διαταραχών (Diagnostic and Statistical Manual of Mental Disorders – DSM) Πέμπτη Έκδοση (DSM-5; APA, 2013) παρέχει πληροφορίες του βαθμού σοβαρότητας και παρουσιάζει κλινικά χαρακτηριστικά τα οποία αποτελούν καθοδήγηση για τον εκάστοτε γιατρό ώστε αυτός να προβεί στη διάγνωση. Αυτό σημαίνει πως ανάλογα με τις παρατηρήσεις που μπορεί να κάνει ο γιατρός, θα αποφανθεί για την κατηγορία του φάσματος στην οποία ανήκει το άτομο. Ένα άτομο μπορεί να ανήκει στο φάσμα του αυτισμού εάν παρατηρηθούν προσπάθειες για κοινωνική αλληλεπίδραση μέσω μη λεκτικών επικοινωνιακών συμπεριφορών, περιορισμένα και επαναλαμβανόμενα μοτίβα συμπεριφοράς ή δραστηριότητες, ελλείμματα στην κοινωνική επικοινωνία, ή ακόμα και αδυναμία διατήρησης και κατανόησης των διαπροσωπικών σχέσεων. Εάν έχουν παρατηρηθεί στο παρελθόν κάποια διαγνωστικά κριτήρια σε ένα άτομο είναι ικανή συνθήκη ώστε βασιζόμενοι στο ιστορικό του να ληφθούν υπόψιν, καθώς οι συμπεριφορές αλλάζουν με την ανάπτυξη [1], [2].

3.1.2 Ερωτηματολόγια

Κατά το παρελθόν, η διάγνωση για το αν ένα άτομο ανήκε στο φάσμα της αναπτυξιακής διαταραχής του αυτισμού γινόταν αποκλειστικά από τις παρατηρήσεις του γιατρού. Το γεγονός αυτό εισήγαγε τον παράγοντα της υποκειμενικότητας του γιατρού στη διάγνωση. Στην προσπάθεια μείωσης της υποκειμενικότητας η διάγνωση ξεκίνησε να πραγματοποιείται από τον γιατρό με γνώμονα τις παρατηρήσεις του και βάσει των κατευθυντήριων γραμμών που υπάρχουν στο Διαγνωστικό και Στατιστικό Εγχειρίδιο καθώς επίσης και ερωτηματολογίων, τα οποία έχουν σκοπό την υποβοήθηση της διάγνωσης του γιατρού. Αυτός ο τρόπος διάγνωσης είναι εκείνος που ακολουθείται ακόμα και σήμερα. Τα ερωτηματολόγια που λαμβάνονται υπόψιν στην διαδικασία διάγνωσης του αυτισμού συμπληρώνονται από το ίδιο το άτομο αν είναι σε θέση να το πραγματοποιήσει, αλλιώς από κάποιο συγγενικό του πρόσωπο. Υπάρχουν πολλά είδη ερωτηματολογίων, ανάλογα για παράδειγμα με την ηλικία κατά την οποία το άτομο επισκέπτεται τον θεράποντα ιατρό, να γίνεται πιο έγκυρη διάγνωση, αφού τα κριτήρια μπορούν να μεταβληθούν [31]. Άλλο παράδειγμα διακριτής κατηγορίας ερωτηματολογίου είναι το σε ποιον απευθύνεται, ποιος, δηλαδή, πρόκειται να το συμπληρώσει. Μερικά παραδείγματα ερωτηματολόγιων αποτελούν τα M-CHAT (Modified Checklist for Autism in Toddlers) [62], AQ (Autism Quotient) [63], RBQ-2A (Repetitive Behaviours Questionnaire – 2) [64], DISCO (Diagnostic Interview for Social and Communication Disorders) [65], ADOS (Autism Diagnostic Observation) [66].

3.1.3 Βιοδείκτες

Ο όρος βιοδείκτης (βιολογικός δείκτης), αναφέρεται σε μια ευρεία κατηγορία ιατρικών ενδείξεων οι οποίες είναι δυνατόν να μετρηθούν εργαστηριακά. Αναλυτικότερα, λαμβάνονται υπόψιν αντικειμενικές ενδείξεις της ιατρικής κατάστασης του εξεταζόμενου ατόμου οι οποίες μπορούν να μετρηθούν με ακρίβεια και επαναληψιμότητα [67].

Κατά τον τωρινό τρόπο αξιολόγησης όσον αφορά στο εάν ένα άτομο βρίσκεται στο φάσμα του αυτισμού, ο παράγοντας της υποκειμενικότητας που εισάγει ο γιατρός μπορεί να οδηγήσει σε λανθασμένα και αναξιόπιστα συμπεράσματα. Για την αντιστάθμιση του παράγοντα αυτού, γίνεται μία προσπάθεια για την ανεύρεση πιο αντικειμενικών τρόπων διάγνωσης και κατ' επέκταση, κριτηρίων. Πολλά υποσχόμενα φαίνεται να είναι τα μοντέλα μηχανικής μάθησης για τον σκοπό αυτό. Έρευνες πραγματοποιούνται με στόχο να ελεγχθεί αν η μηχανική μάθηση με τη χρήση κατάλληλων βιοδεικτών, έχει τη δυνατότητα να παρέχει αποτελεσματική κατάταξη ενός ατόμου ως τυπικά αναπτυσσόμενου ή ως αυτιστικού. Δυστυχώς, μέχρι στιγμής δεν έχει βρεθεί κάποιος δείκτης που να συνδέεται άμεσα με τον αυτισμό ώστε να μπορεί να χρησιμοποιηθεί με ακρίβεια για την κατάταξη ενός ατόμου, με αποτέλεσμα να υπάρχει μεγάλο ενδιαφέρον για την εύρεση βιοδεικτών οι οποίοι θα ενισχύσουν τα αποτελέσματα. Για τον σκοπό αυτό, στην παρούσα διπλωματική εργασία, έχουν χρησιμοποιηθεί διάφοροι βιοδείκτες όπως η ηλικία, το φύλο, καθώς και μετρήσεις που προήλθαν από την επεξεργασία της fMRI εικόνας όπως είναι η λειτουργική συνδεσιμότητα περιοχών του εγκέφαλου, στατική και δυναμική [11], [12], [59], [68], [69].

Λειτουργική συνδεσιμότητα

Ως λειτουργική συνδεσιμότητα ορίζεται η χρονική συσχέτιση μεταξύ χρονοσειρών των διαφόρων περιοχών του εγκεφάλου [70]. Η μεταβολή στη λειτουργική συνδεσιμότητα (Functional Connectivity – FC) μεταξύ περιοχών του εγκεφάλου αναμένεται να αποτελεί δυνητικό βιοδείκτη για ταξινόμηση ή πρόβλεψη διαταραχών του εγκεφάλου [29], [71], [72]. Η λειτουργική συνδεσιμότητα χωρίζεται σε δυο κατηγορίες: στη στατική λειτουργική συνδεσιμότητα και στη δυναμική λειτουργική συνδεσιμότητα, ανάλογα με το χρονικό παράθυρο στο οποίο υπολογίζονται οι προαναφερόμενες συσχετίσεις.

3.1.3..1 Στατική Λειτουργική συνδεσιμότητα

Η στατική λειτουργική συνδεσιμότητα εκφράζει την συσχέτιση ενεργοποίησης περιοχών ενδιαφέροντος καθ' όλη τη διάρκεια λήψης της εικόνας. Ένας τρόπος υπολογισμού της είναι ο υπολογισμός των συντελεστών συσχέτισης Pearson μεταξύ κάθε δυνατού ζεύγους των περιοχών ενδιαφέροντος και στη συνέχεια η χρήση του μετασχηματισμού Fisher στους συντελεστές συσχέτισης Pearson ώστε να βελτιωθεί η κανονικότητα της κατανομής συσχετισμού και τελικά να διευκολυνθεί η ερμηνεία των αποτελεσμάτων [73]. Εν συντομία, ο μετασχηματισμός Fisher του συντελεστή συσχέτισης δειγμάτων r δίνεται από τον ακόλουθο τύπο [74]:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r) \quad \textbf{(1)}$$

όπου ο συντελεστής συσχέτισης δειγμάτων Pearson, r, για N διμερή ζεύγη δειγμάτων (X_i, Y_i), i=1,2,…N είναι:

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \quad \textbf{(2)}$$

3.1.3..2 Δυναμική Λειτουργική συνδεσιμότητα

Η δυναμική λειτουργική συνδεσιμότητα αφορά μικρότερα χρονικά διαστήματα και όχι όλη την διάρκεια λήψης της εικόνας. Η τεχνική κυλιόμενου παραθύρου η οποία καταγράφει τις συσχετίσεις μεταξύ των χρονοσειρών μέσα σε συγκεκριμένες χρονικές ζώνες, είναι ένας τρόπος υπολογισμού της δυναμικής λειτουργικής συνδεσιμότητας [75]. Το μέγεθος του παραθύρου και το πόσο μετακινείται είναι σημαντική παράμετρος στους υπολογισμούς δυναμικής λειτουργικής συνδεσιμότητας που βασίζονται σε παράθυρα. Το αποτέλεσμα ενός τέτοιου υπολογισμού οδηγεί στη δημιουργία ενός τρισδιάστατου πίνακα, ο οποίος στην τρίτη διάσταση περιλαμβάνει τις συσχετίσεις μεταξύ των περιοχών για το εκάστοτε ζεύγος περιοχών για το συγκεκριμένο χρονικό παράθυρο. Διάφορες στατιστικές στιγμές λαμβάνονται υπόψιν και εξάγονται, προκειμένου να εξεταστεί ο τρόπος με τον οποίο μεταβάλλεται η δυναμική λειτουργική συνδεσιμότητα στην παρούσα διπλωματική εργασία. Τέτοιες μπορεί να είναι η μέση τιμή, η διακύμανση, η κυρτότητα και η στρέβλωση οι οποίες ορίζονται για N δείγματα, όπου x_i είναι το δείγμα i:

- Μέση τιμή

$$\mu = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad \textbf{(3)}$$

- Διακύμανση

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1} \quad \textbf{(4)}$$

- Κυρτότητα

$$\alpha_4 = \frac{\mu^4}{s^4} - 3 \quad \textbf{(5)}$$

- Στρέβλωση

$$a_3 = \frac{\mu^3}{s^3} \quad \textbf{(6)}$$

Στις προηγούμενες συναρτήσεις, το σύμβολο μ_r είναι η κεντρική ροπή τάξης r, όπου r φυσικός αριθμός, για την οποία ισχύει:

$$\alpha_r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^r}{N} \quad \textbf{(7)}$$

3.2 Αλγόριθμοι Ταξινόμησης

Σε αυτήν την υποενότητα θα αναφερθούν οι ταξινομητές που χρησιμοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας. Για τους σκοπούς υλοποίησης των αλγορίθμων που θα δημιουργούν τα μοντέλα εκμάθησης χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn. Η scikit-learn είναι μια βιβλιοθήκη,

ανοιχτού κώδικα, η οποία παρέχεται για την γλώσσα προγραμματισμού python και η οποία περιλαμβάνει μεταξύ άλλων και τους ταξινομητές που θα αναφερθούν σε αυτήν την ενότητα. Για τους ταξινομητές αυτούς, θα γίνει και μια σύντομη αναφορά στα ορίσματα που δέχονται. Εκτενέστερη αναφορά γίνεται και στην ιστοσελίδα της scikit-learn [76], [77].

3.2.1 Νευρωνικά δίκτυα

Ένα τεχνητό νευρωνικό δίκτυο (Artificial Neural Network – ANN) είναι μια ευέλικτη μαθηματική δομή που, μεταξύ συνόλων δεδομένων εισόδου και εξόδου, έχει την δυνατότητα να αναγνωρίζει πολύπλοκες μη γραμμικές σχέσεις [78]. Η αρχιτεκτονική τους έχει επιρροές από την αρχιτεκτονική των βιολογικών νευρωνικών δικτύων του Κεντρικού Νευρικού Συστήματος (ΚΝΣ). Η βασική διαφορά των βιολογικών και τεχνητών νευρωνικών δικτύων είναι ότι τα δεύτερα εκπαιδεύονται με τη βοήθεια παραδειγμάτων, έτσι ώστε να μαθαίνουν από το περιβάλλον τους μέσα από μια διαδικασία μάθησης [79].

Ένα νευρωνικό δίκτυο απαρτίζεται από ένα αρχικό στρώμα (στάδιο επεξεργασίας) εισόδου, ένα ή περισσότερα κρυμμένα στρώματα και ένα στρώμα εξόδου [78]. Τα στοιχεία επεξεργασίας σε κάθε στρώμα ονομάζονται κόμβοι ή μονάδες. Κάθε κόμβος συνδέεται με έναν αριθμό κόμβων των γειτονικών στρωμάτων. Οι παράμετροι που σχετίζονται με κάθε μια από αυτές τις συνδέσεις ονομάζονται βάρη. Το πρώτο στρώμα συνδέεται με τις μεταβλητές εισόδου και ονομάζεται στρώμα εισόδου. Το τελευταίο στρώμα συνδέεται με τις μεταβλητές εξόδου και ονομάζεται στρώμα εξόδου. Τα επίπεδα μεταξύ των στρωμάτων εισόδου και εξόδου ονομάζονται κρυμμένα στρώματα. Αν όλες οι συνδέσεις είναι προσωτροφοδοτούμενες (feed forward), δηλαδή επιτρέπουν τη μεταφορά πληροφοριών μόνο από ένα προηγούμενο στρώμα στα επόμενα διαδοχικά επίπεδα χωρίς να υπάρχουν κύκλοι ή βρόχοι στο δίκτυο, τότε το νευρωνικό δίκτυο χαρακτηρίζεται ως νευρωνικό δίκτυο πρόσθιας τροφοδότησης (feed forward neural network) ενός επιπέδου, εάν δεν υπάρχουν κρυφά στρώματα, ή πολλαπλών επιπέδων εάν υπάρχουν κρυφά στρώματα. Εάν σε ένα νευρωνικό δίκτυο υπάρχουν ένας ή περισσότεροι βρόγχοι ανάδρασης τότε το νευρωνικό δίκτυο χαρακτηρίζεται ως αναδρομικό νευρωνικο δίκτυο (Recurrent Neural Network –RNN) [80].

Καθοριστική για την λειτουργία ενός νευρωνικού δικτύου είναι η συνάρτηση ενεργοποίησης που θα επιλεγεί. Οι συναρτήσεις ενεργοποίησης καθορίζουν την έξοδο ενός νευρωνικού δικτύου, την ακρίβειά του και την υπολογιστική αποτελεσματικότητα της εκπαίδευσης του. Δέχονται σαν είσοδο το άθροισμα των εισόδων κάθε νευρώνα πολλαπλασιασμένων με τα αντίστοιχα βάρη τους, ενώ η έξοδός τους καθορίζει εάν μπορεί να ενεργοποιηθεί ένας νευρώνας ή όχι με βάση το αν η είσοδος είναι σχετική με την πρόβλεψη του μοντέλου [81]. Συμβάλλουν έτσι, στην ομαλοποίηση της εξόδου κάθε νευρώνα σε ένα εύρος μεταξύ 0 και 1 ή μεταξύ -1 και 1 [82].

MLP

Το νευρωνικό δίκτυο MLP (Multi-Layer Perceptron classifier – πολυστρωματικός ταξινομητής) είναι ένας ταξινομητής πολλαπλών επιπέδων

που βασίζεται στον αλγόριθμο Perceptron. Το Perceptron, ήταν το πρώτο εκπαιδευτικό νευρωνικό δίκτυο και δημιουργήθηκε από τον ψυχολόγο του Πανεπιστημίου Cornell, Frank Rosenblatt το 1957 [83]. Το νευρωνικό δίκτυο Perceptron είναι παρόμοιο με τα μοντέρνα νευρωνικά δίκτυα, με την διαφορά ότι έχει μόνο ένα κρυφό στρώμα με ρυθμιζόμενα βάρη και κατώφλια μεταξύ των στρωμάτων εισόδου και εξόδου [84]. Τα MLP ανήκουν στον κλάδο των feedforward τεχνητών νευρωνικών δικτύων [77].

Η βιβλιοθήκη scikit-learn παρέχει τον ταξινομητή MLPClassifier για δίκτυα MLP. Ο MLPClassifier εκπαιδεύεται επαναληπτικά αφού σε κάθε βήμα του χρόνου υπολογίζονται οι μερικοί παράγωγοι της συνάρτησης απώλειας σε σχέση με τις παραμέτρους του μοντέλου για την ενημέρωση αυτών. Το μοντέλο αυτό βελτιστοποιεί την συνάρτηση της λογαριθμικής απώλειας (log-loss function) ή της στοχαστικής κλίσης καθόδου (stochastic gradient descent) ανάλογα με τον επιλυτή που έχει οριστεί από το χρήστη (αν είναι ο 'lbfgs' συμβαίνει το πρώτο ενδεχόμενο, ενώ αν είναι ο 'sgd' ή ο 'adam' το δεύτερο). Πολύ σημαντικό ρόλο στην λειτουργία και στην επίδοση του MLP έχουν οι παράμετροι που χρησιμοποιούνται για την εκπαίδευσή του. Στην συνέχεια αναφέρονται κάποιες από αυτές:

- 'hidden layer sizes': είναι η παράμετρος πλήθους των κρυφών επίπεδων και προσδιορίζει τον αριθμό νευρώνων και κρυφών επιπέδων.
- 'activation': είναι η συνάρτηση ενεργοποίησης για τα εσωτερικά επίπεδα. Μπορεί να είναι:
  - η «ταυτότητα» ('identity') η οποία επιστρέφει την

$$f(x) = x \quad \textbf{(8)}$$

  - η λογαριθμική σιγμοειδής συνάρτηση:

$$f(x) = 1/(1 + e^{-x}) \quad \textbf{(9)}$$

  - η υπερβολική εφαπτομένη:

$$f(x) = \tanh(x) \quad \textbf{(10)}$$

  - ή η διορθωμένη γραμμική συνάρτηση μονάδας ('Rectified Linear Unit function – ReLU'):

$$f(x) = \max(0, x) \quad \textbf{(11)}$$

- 'Solver': είναι ο επιλυτής για την βελτιστοποίηση των βαρών ο οποίος μπορεί να είναι:
  - 'lbfgs': ο οποίος είναι ένας βελτιστοποιητής στην οικογένεια μεθόδων quasi-Newton [85],
  - 'Sgd': ο οποίος αναφέρεται σε στοχαστική κλίση καθόδου [86],
  - 'Adam': ο οποίος αναφέρεται σε ένα στοχαστικό βελτιστοποιητή βασισμένο στην κλίση που προτείνουν οι Kingma, Diederik και Jimmy Ba. [87].

Αξίζει να σημειωθεί ότι ο 'adam' είναι αρκετά αποτελεσματικός για μεγάλα σύνολα δεδομένων τόσο ως προς το χρόνο εκπαίδευσης όσο και για τα αποτελέσματα αξιολόγησης ενώ ο 'lbfgs' είναι αποδοτικότερος στα μικρά σύνολα δεδομένων.

- 'learning_rate': είναι ο ρυθμός εκμάθησης, ο οποίος αφορά στον τρόπο με το οποίο μεταβάλλονται και ενημερώνονται τα βάρη. Αυτός μπορεί να είναι:

- «Σταθερός» (‘Constant’): όπου τότε μια σταθερή ταχύτητα εκμάθησης δίνεται από το ‘learning_rate_init’
- ‘Invscaling’: όπου τότε μειώνεται σταδιακά ο ρυθμός εκμάθησης σε κάθε βήμα ‘t’ χρησιμοποιώντας έναν αντίστροφο συντελεστή κλιμάκωσης του ‘power_t’
- «Προσαρμοστικός» (‘Adaptive’): ο οποίος διατηρεί το ρυθμό εκμάθησης σταθερό στο ‘learning_rate_init’ όσο η απώλεια της κατάρτισης συνεχίζει να μειώνεται. Κάθε φορά που δύο διαδοχικοί πλήρεις κύκλοι εκπαίδευσης (epoch – εποχές) αποτυγχάνουν να μειώσουν την απώλεια της εκπαίδευσης κατά τουλάχιστον ‘tol’, ή δεν αυξάνουν την βαθμολογία αξιολόγησης τουλάχιστον κατά το ‘tol’ εάν το ‘early_stopping’ είναι ενεργοποιημένο, ο τρέχων ρυθμός εκμάθησης διαιρείται με 5.

- ‘learning_rate_init’: είναι ο αρχικός ρυθμός εκμάθησης που χρησιμοποιείται. Ελέγχει το βήμα στην ενημέρωση των βαρών. Χρησιμοποιείται μόνο όταν για ‘solver’ έχει επιλεγεί ο ‘sgd’ ή ο ‘adam’.
- ‘power_t’: είναι ο εκθέτης για την αντίστροφη κλιμάκωση του ρυθμού εκμάθησης. Χρησιμοποιείται στην αξιολόγηση του πραγματικού ρυθμού εκμάθησης όταν ο ρυθμός εκμάθησης (‘learning_rate’) έχει οριστεί σε ‘invscaling’. Επίσης, χρησιμοποιείται μόνο όταν ο επιλυτής (‘solver’) είναι ο ‘sgd’.
- ‘batch_size’: είναι το πλήθος των στοιχείων που θα ληφθούν ανά εποχή υπόψιν για την εκπαίδευση, ενώ μπορεί να προσδιορίσει το μέγεθος των μικροπαρτίδων για τους στοχαστικούς βελτιστοποιητές.
- ‘max_iter’: είναι ο μέγιστος αριθμός επαναλήψεων που μπορεί να εκτελέσει το δίκτυο προκειμένου να βελτιώσει το ρυθμό εκμάθησης. Ο επιλυτής επαναλαμβάνει μέχρι τη σύγκλιση (που καθορίζεται από το ‘tol’) ή το μέγιστο αριθμό των επαναλήψεων. Είναι σημαντικό να αναφερθεί ότι για τους στοχαστικούς επιλυτές (‘sgd’, ‘adam’) αυτό καθορίζει τον αριθμό των εποχών δηλαδή το πόσες φορές θα χρησιμοποιηθεί κάθε σημείο δεδομένων και όχι τον αριθμό των βαθμίδων κλίσης.
- ‘verbose’: ορίζει το αν θα εκτυπώνονται μηνύματα προόδου στην κονσόλα.
- ‘shuffle’: προσδιορίζει το αν θα ανακατευθούν ή όχι τα δείγματα σε κάθε επανάληψη/εποχή. Χρησιμοποιείται μόνο όταν ο επιλυτής είναι ‘sgd’ ή ’adam’.
- ‘random_state’ (τυχαία κατάσταση ή γεννήτρια ψευδοτυχαίων αριθμών): μπορεί να οριστεί από έναν ακέραιο αριθμό (integer) ή με παράδειγμα τυχαίας κατάστασης (RandomState instance) ή ως τίποτα (None). Ουσιαστικά εδώ ο ορίζεται η λειτουργία της γεννήτριας ψευδοτυχαίων αριθμών. Αν οριστεί ακέραιος αριθμός τότε αυτός θα είναι η αρχική τιμή (‘seed’) που θα χρησιμοποιείται από τη γεννήτρια τυχαίων αριθμών και άρα θα υπάρχει επαναληψιμότητα μεταξύ εκτελέσεων του προγράμματος. Στην δεύτερη επιλογή δίνεται ως random_state μία ήδη υπάρχουσα γεννήτρια ψευδοτυχαίων αριθμών (δημιουργημένη σε προηγούμενο σημείο του προγράμματος). Στην τρίτη περίπτωση, χρησιμοποιείται ως γεννήτρια το numpy.random.
- ‘tol’ (tolerance): αντιπροσωπεύει την ανοχή για τη βελτιστοποίηση. Όταν η συνάρτηση απώλειας ή η βαθμολογία δεν βελτιώνεται τουλάχιστον κατά tol για n_iter_no_change διαδοχικές επαναλήψεις (εκτός και αν η παράμετρος ‘learning_rate’ έχει οριστεί ως «προσαρμοστική») θεωρείται ότι η μέθοδος συνέκλινε και η εκπαίδευση σταματά.

- ‘epsilon’: καθορίζει την τιμή για αριθμητική σταθερότητα στον adam. Χρησιμοποιείται μόνο όταν ο επιλυτής είναι ‘adam’.
- ‘n_iter_no_change’: είναι ο μέγιστος αριθμός εποχών κατά τις οποίες δεν επιτεύχθηκε τουλάχιστον ‘tol’ βελτίωση.
- ‘momentum’: σχετίζεται με την στιγμή για την ενημέρωση της κλίσης καθόδου. Πρέπει να είναι μεταξύ 0 και 1. Χρησιμοποιείται μόνο όταν ο solver είναι ‘sgd’.

3.2.2 Ταξινομητές

Support Vector Machines (SVMs) - C-Support Vector Classification (SVC)

Οι ταξινομητές SVM (Support Vector Machine) είναι δυαδικοί ταξινομητές που επιτρέπουν την δημιουργία ενός μοντέλου το οποίο αντιστοιχώντας παρατηρήσεις σε σημεία ενός επιπέδου, δημιουργεί μία γραμμική συνάρτηση ώστε να ταξινομήσει τα δείγματα βάσει της θέσης τους. Το όνομα της μεθόδου προέρχεται από τα σημεία που βρίσκονται στην μικρότερη απόσταση από την συνάρτηση ταξινόμησης και άρα αυτά τα οποία παρουσιάζουν την μεγαλύτερη «δυσκολία» ταξινόμησης [88]. Οι SVM προσπαθούν να μεγιστοποιήσουν το περιθώριο γύρω από την συνάρτηση ταξινόμησης. Αν και γραμμικοί ταξινομητές αρχικά, μπορούν να μετατραπούν σε μη γραμμικούς με τη χρήση συναρτήσεων πυρήνων (kernels) οι οποίες μεταφέρουν τα σημεία - παρατηρήσεις σε άλλο χώρο (ν+1-διάστατο χώρο, όπου ν το πλήθος των χαρακτηριστικών με βάση τα οποία ταξινομούν οι αλγόριθμοι αυτοί). Η συνάρτηση που χωρίζει σε αυτήν την περίπτωση τα σημεία, αποκτάει διαφορετική και πιο πολύπλοκη μορφή, επιτρέποντας την ταξινόμηση των παρατηρήσεων, τα οποία δεν είναι δυνατό να διαχωριστούν γραμμικά, με λιγότερα σφάλματα. Χαρακτηριστικά παραδείγματα πυρήνων που χρησιμοποιούνται είναι ο γραμμικός, ο γκαουσιανός (radial basis function – rbf) και ο πολυωνυμικός (polynomial – poly) [77], [88], [89]. Επιπλέον, κάθε διαφορετική συνάρτηση πυρήνα χρησιμοποιεί διαφορετική συνάρτηση κόστους, την οποία προσπαθεί να ελαχιστοποιήσει με στόχο την ορθότερη κατάταξη των δειγμάτων, σύμφωνα με τα χαρακτηριστικά που έχει στη διάθεσή της.

Σε αυτόν τον αλγόριθμο οι τιμές που μπορούν να οριστούν ως kernel είναι οι ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’ ή μία συνάρτηση κατασκευασμένη από το χρήστη με διαφορετικά κριτήρια βελτιστοποίησης της συνάρτησης διαχωρισμού. Μια ακόμα σημαντική παράμετρος αυτού του αλγορίθμου αποτελεί το ‘gamma’. Η μεταβλητή αυτή αποτελεί τον συντελεστή του kernel αν είναι ορισμένος ως ‘rbf’, ‘poly’ ή ‘sigmoid’. Εάν για gamma επιλεγεί η τιμή ‘scale’, τότε η συνάρτηση θα χρησιμοποιήσει σαν τιμή για το gamma την σχέση:

$$gamma = \frac{1}{n\_features * var(X)} \quad \textbf{(12)}$$

όπου το n_features είναι ο αριθμός των χαρακτηριστικών και το var(X) είναι η διακύμανση των δειγμάτων X που δίνονται ως είσοδος. Εναλλακτικά, εάν είναι ‘auto’, τότε παίρνει την τιμή:

$$gamma = \frac{1}{n\_features} \quad \textbf{(13)}$$

Ως παράμετρος κανονικοποίησης χρησιμοποιείται η παράμετρος C. Η ισχύς της κανονικοποίησης είναι αντιστρόφως ανάλογη με το C. Πρέπει να είναι αυστηρά θετικό, ενώ η ποινή είναι μια τετράγωνη ποινή τύπου L2. Εδώ ως ποινή, νοείται μια αρνητική «βαθμολόγηση» της απόδοσης του αλγορίθμου η οποία ισχύει όταν γίνεται λανθασμένη ταξινόμηση παρατηρήσεων.
Στην εργασία αυτή, χρησιμοποιήθηκε η κλάση SVC του πακέτου scikit-learn της python. Η κλήση αυτής έγινε με όρισμα πυρήνα τον rbf. Πρόκειται για την ακόλουθη εκθετική συνάρτηση:

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right) \quad \textbf{(14)}$$

Στη συγκεκριμένη συνάρτηση η παράμετρος γ είναι το gamma που ορίστηκε προηγουμένως, ενώ τα x και x' αποτελούν 2 δείγματα.

Nearest Neighbors Classification

Η ταξινόμηση που βασίζεται στους κοντινότερους γείτονες (Nearest Neighbors Classification) είναι ένας τύπος ταξινόμησης που δεν επιχειρεί να κατασκευάσει ένα γενικό εσωτερικό μοντέλο, αλλά αποθηκεύει στιγμιότυπα των δεδομένων εκπαίδευσης. Η ταξινόμηση αυτού του τύπου υπολογίζεται με απλή πλειοψηφία των πλησιέστερων γειτόνων κάθε σημείου. Συγκεκριμένα, ένα δείγμα μη επισημασμένο θα ταξινομηθεί στην ίδια κατηγορία που ανήκει η πλειοψηφία των κοντινότερων γειτόνων του [77].

Ο ταξινομητής KNeighborsClassifier της βιβλιοθήκης scikit-learn εφαρμόζει τη μάθηση βασισμένη στους k πλησιέστερους γείτονες κάθε δείγματος, όπου k είναι ένας ακέραιος αριθμός που προσδιορίζεται από τον χρήστη, ενώ η ιδανική τιμή του εξαρτάται άμεσα από τα δεδομένα του προβλήματος. Αν ο στόχος του χρήστη είναι η κατάταξη που έχει εκχωρηθεί σε ένα δείγμα να προσδιορίζεται μόνο από την πλειοψηφία των πλησιέστερων γειτόνων, πρέπει να χρησιμοποιηθούν όμοια βάρη (weights) στους γείτονες. Για την επιλογή αυτή, ορίζουμε την τιμή της παραμέτρου 'weight' του ταξινομητή ως 'uniform'. Αν ο στόχος του χρήστη είναι να επηρεάζεται η επισήμανση των δειγμάτων από την απόσταση του μη επισημασμένου δείγματος που εξετάζεται, ορίζεται η τιμή της παραμέτρου 'weight' του ταξινομητή ως 'distance'. Αυτό θα έχει σαν αποτέλεσμα τον προσδιορισμό των βαρών ανάλογα με το αντίστροφο της απόστασης από το δείγμα αναζήτησης. Ακόμα, παρέχεται στον χρήστη αν το επιθυμεί, η δυνατότητα να χρησιμοποιήσει μια δική του συνάρτηση για τον προσδιορισμό των βαρών.

Linear Discriminant Analysis (LDA)

Ο ταξινομητής Γραμμικής Ανάλυσης Διακρίσεων (Linear Discriminant Analysis-LDA) έχει ένα γραμμικό όριο απόφασης, που δημιουργείται από την τοποθέτηση κατηγοριοποιημένων συνθηκών πυκνότητας στα δεδομένα και τη χρήση του κανόνα Bayes. Το μοντέλο αποδίδει μια πυκνότητα (κατανομή) Gauss σε κάθε κατηγορία, υποθέτοντας ότι όλες οι κλάσεις μοιράζονται τον ίδιο πίνακα συνδιασποράς (covariance). Το προσαρμοσμένο μοντέλο μπορεί επίσης να χρησιμοποιηθεί για να μειώσει τις διαστάσεις της εισόδου, μέσω προβολής της σε κατευθύνσεις με την μέγιστη δυνατότητα διάκρισης.

Η χρήση αυτού του μοντέλου που έχει επιλεγεί από την βιβλιοθήκη scikit-learn, παρέχει τρεις επιλογές ανάθεσης στον ‘επιλυτή’ (solver) ώστε να μπορεί ο χρήστης να επιλέξει τον ιδανικό για την εργασία που επιθυμεί να εκτελέσει σύμφωνα με τα δεδομένα που έχει στη διάθεσή του. Ο ‘επιλυτής’ μπορεί να είναι ο ‘svd’, ο ‘lsq’ ή ο ‘eigen’. Συγκριτικά ο πρώτος ενδείκνυται για μεγάλο αριθμό χαρακτηριστικών καθώς δεν βασίζεται στον υπολογισμό της μήτρας (πίνακα) συνδιακύμανσης, σε αντίθεση με τον τρίτο που δεν ενδείκνυται για μεγάλο αριθμό χαρακτηριστικών διότι βασίζεται στον υπολογισμό της μήτρας συνδιακύμανσης. Η ‘συρρίκνωση’ (‘shrinkage’) μπορεί να οριστεί ως καθόλου (‘None’), ως αυτόματη χρησιμοποιώντας την λήμμα του Ledoit-Wolf (‘auto’), ή ως μια τιμή μεταξύ 0 και 1. Με τον όρο συρρίκνωση εννοείται εδώ η στατιστική τεχνική μέσω της οποίας, ένα πολύπλοκο μοντέλο με πολλαπλές καταστάσεις μπορεί να απλοποιηθεί και να δημιουργηθεί ένα νέο, απλούστερο με ικανοποιητική εκτίμηση παραμέτρων [90].

Logistic Regression

Ο ταξινομητής Logistic Regression είναι μια στατιστική μέθοδος που μπορεί να χρησιμοποιηθεί για ταξινόμηση δυο κλάσεων. Ο ταξινομητής αυτός είναι ένα γραμμικό μοντέλο ταξινόμησης που βασίζεται στην logistic function. Για λόγους πληρότητας θα αναφερθούν κάποιες βασικές μεταβλητές αυτού του αλγορίθμου. Η παράμετρος ‘C’ σχετίζεται με την αντίστροφη τιμή της ισχύος ομαλοποίησης των δεδομένων. Μικρότερες τιμές της οδηγούν σε ισχυρότερη ομαλοποίηση. Μέσω της μεταβλητής ‘solver’ ορίζεται ο αλγόριθμος που θα χρησιμοποιηθεί για την επίλυση του προβλήματος βελτιστοποίησης. Οι τιμές που μπορεί να έχει η μεταβλητή solver είναι ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’ και ‘saga’. Ο ‘liblinear’ είναι καλή επιλογή για μικρά σύνολα δεδομένων. Ο ‘sag’ και ‘saga’ είναι γρηγορότεροι για μεγάλα σύνολα δεδομένων. Οι ‘newton-cg’, ‘lbfgs’, ‘sag’ και ‘saga’ μπορούν να χειριστούν ποινή L2 ή καθόλου ποινή (όπου η ποινή, όπως και προηγουμένως, αναφέρεται σε λανθασμένες ταξινομήσεις). Ακόμα μια σημαντική μεταβλητή αποτελεί η ‘max_iter’, όπου μέσω της τιμής αυτής της παραμέτρου καθορίζεται ο μέγιστος αριθμός επαναλήψεων για σύγκλιση των επιλυτών (καθώς η επίλυση γίνεται με επαναληπτικούς αλγορίθμους αφού το πρόβλημα δεν επιλύεται αλγεβρικά).

3.2.3 Δέντρα

Οι ταξινομητές δέντρων απόφασης χρησιμοποιούνται με επιτυχία σε πολλές διαφορετικές περιοχές όπως στις χρηματιστηριακές συναλλαγές [91], αναγνώριση γραπτού κειμένου [92] αλλά και στην γεωλογική μελέτη εδαφών [93]. Στόχος τους είναι να δημιουργηθεί ένα μοντέλο που να προβλέπει την αξία μιας μεταβλητής στόχου με την εκμάθηση απλών κανόνων απόφασης που προκύπτουν από τα χαρακτηριστικά δεδομένων.

Decision trees

Τα decision trees (δέντρα αποφάσεων) αποτελούν μια πολύ γνωστή τεχνική ταξινόμησης σε διάφορα θέματα αναγνώρισης προτύπων [94]. Ένα δέντρο αποφάσεων αποτελείται από κόμβους για έλεγχο και απόφαση σε χαρακτηριστικά, άκρα για διακλάδωση βάσει τιμών του επιλεγμένου χαρακτηριστικού και «φύλλα»

όπου επισυνάπτεται μια μοναδική κλάση ανά φύλλο [95]. Η λειτουργία ενός δέντρου αποφάσεων μπορεί να χωριστεί σε κάποια στάδια. Στο πρώτο στάδιο γίνεται η κατασκευή του και στη συνέχεια γίνεται η αποκοπή (αναρρόφηση) των λιγότερο χρήσιμων κλάδων (όπου κλάδοι είναι οι κανόνες για την κατάταξη των δεδομένων σε κλάσεις) του δέντρου για την καλύτερη εξισορρόπηση της ακρίβειας των δεδομένων εκπαίδευσης με την πολυπλοκότητα του μοντέλου [96]. Το τελικό δέντρο χωρίζει το χώρο των χαρακτηριστικών σε έναν αριθμό επισημασμένων περιοχών.

Το μέγιστο βάθος του δέντρου (‘max_depth’) μπορεί να οριστεί ως κανένα (‘None’) ή ως ένας ακέραιος αριθμός. Ειδικότερα, για την πρώτη περίπτωση, οι κόμβοι επεκτείνονται έως ότου όλα τα φύλλα είναι «καθαρά» (δηλαδή όταν κάθε ένα αντιστοιχεί σε μία μόνο κλάση) ή έως ότου όλα τα φύλλα περιέχουν λιγότερα από τον ελάχιστο αριθμό δειγμάτων που απαιτείται για τον διαχωρισμό ενός εσωτερικού κόμβου (όπου κόμβος νοείται ένα κανόνας που κατατάσσει βάσει χαρακτηριστικών, μια παρατήρηση σε κλάση).

AdaBoost

Ο ταξινομητής AdaBoost βασίζεται για την παραγωγή της τελικής πρόβλεψης στην ιδέα της δημιουργίας ενός κανόνα υψηλής ακρίβειας πρόβλεψης συνδυάζοντας πολλούς σχετικά αδύναμους και ανακριβείς κανόνες, όπως αυτούς που δημιουργούνται από μικρά δέντρα αποφάσεων, που προέκυψαν από επανειλημμένα τροποποιημένες εκδόσεις των δεδομένων [97]. Πρακτικά γίνονται πολλαπλές εκτελέσεις ενός αλγορίθμου ταξινόμησης ρυθμίζοντας κάθε φορά παραμέτρους του. Για να παραχθεί η τελική πρόβλεψη, οι προβλέψεις από όλους τους συνδυασμούς ομαδοποιούνται με βάρη (με ειδικό βάρος για την κάθε επανάληψη) [77]. Για κάθε διαδοχική επανάληψη μέσα στο μοντέλο, τα βάρη του δείγματος επαναπροσδιορίζονται ανάλογα με την επιτυχία ή την αποτυχία της ταξινόμησης. Τα βάρη των εσφαλμένα ταξινομημένων περιπτώσεων ρυθμίζονται έτσι ώστε οι επόμενοι ταξινομητές να εστιάζουν περισσότερο σε δύσκολες περιπτώσεις.

Στον ταξινομητή AdaBoost μπορεί να προσδιοριστεί, αν είναι επιθυμητό, ο μέγιστος αριθμός εκτιμητών στους οποίους τερματίζεται η ενίσχυση. Σε περίπτωση τέλειας τοποθέτησης, η διαδικασία της μάθησης διακόπτεται νωρίτερα. Ως «εκτιμητής βάσης» ορίζεται ο αλγόριθμος ταξινόμησης που χρησιμοποιείται κατά την εκτέλεση του ταξινομητή Adaboost. Αν η ανάθεση της παραμέτρου του εκτιμητή βάσης (‘base estimator’) είναι η τιμή κανένα (’None’), τότε ο εκτιμητής βάσης είναι το δέντρο αποφάσεων με ‘max_depth’ ίσο με 1 (όπου το ‘max_depth’ είναι η ίδια μεταβλητή που αναφέρεται και στο δέντρο αποφάσεων).

Τυχαία Δάση - Random Forest

Τα Τυχαία Δάση (Random Forest) είναι ένας συνδυασμός δέντρων πρόβλεψης (tree predictors), όπως είναι τα δέντρα απόφασης, όπου κάθε δέντρο εξαρτάται από τις τιμές ενός τυχαίου διανύσματος, ανεξάρτητου από τα υπόλοιπα διανύσματα αλλά με την ίδια κατανομή τιμών με αυτά [98]. Η τυχαιότητα στα δάση παράγει δέντρα απόφασης με ελαφρώς αποσυνδεδεμένα σφάλματα πρόβλεψης αντιμετωπίζοντας το πρόβλημα ότι τα δέντρα απόφασης

εμφανίζουν συνήθως μεγάλη διακύμανση και τείνουν να έχουν προβλήματα υπερπροσαρμογής (over-fitting). Ένα τυχαίο δάσος ταιριάζει έναν αριθμό ταξινομητών δέντρων απόφασης στα δεδομένα και χρησιμοποιεί τον μέσο όρο για τη βελτίωση της ακρίβειας πρόβλεψης και να ελέγξει το over-fitting.

Μια ρυθμιζόμενη παράμετρος από τον χρήστη είναι ο αριθμός των δέντρων του δάσους και το μέγιστο βάθος κάθε δέντρου. Το μέγιστο βάθος του δέντρου μπορεί να οριστεί ως κανένα (’None’) ή ως ένας ακέραιος αριθμός. Ειδικά για την πρώτη περίπτωση, οι κόμβοι επεκτείνονται έως ότου όλα τα φύλλα είναι «καθαρά»(δηλαδή όταν κάθε ένα αντιστοιχεί σε μία μόνο κλάση) ή έως ότου όλα τα φύλλα περιέχουν λιγότερα από τον ελάχιστο αριθμό δειγμάτων που απαιτείται για τον διαχωρισμό ενός εσωτερικού κόμβου, όπου τα δείγματα είναι οι παρατηρήσεις (άτομα) και κόμβοι είναι σημεία διαχωρισμού των παρατηρήσεων σε υποσύνολα βάσει της τιμής ενός χαρακτηριστικού.

4 Μείωση Διαστατικότητας

Το πλήθος των δεδομένων εισόδου ενός ταξινομητή έχει σημαντικό ρόλο στην απόδοση μιας ταξινόμησης, καθώς η μείωση των διαστάσεων μπορεί να μην επιφέρει πάντα βελτίωση σε ένα σύστημα ταξινόμησης [99]. Το παραπάνω συμπέρασμα είναι απόρροια του γεγονότος ότι ένα σύστημα αναγνώρισης σχεδιάζεται χρησιμοποιώντας ένα πεπερασμένο σύνολο εισόδων. Ενώ η απόδοση αυτού του συστήματος αυξάνεται αν προσθέσουμε επιπλέον χαρακτηριστικά, υπάρχει ένα ασαφές όριο πέραν του οποίου μια περαιτέρω προσθήκη μπορεί να οδηγεί σε υποβάθμιση της απόδοσης.

Η επιλογή χαρακτηριστικών (ΕΧ) είναι η διαδικασία επιλογής ενός υποσυνόλου των αρχικών χαρακτηριστικών για χρήση στην κατασκευή ενός μοντέλου το οποίο θα χρησιμοποιηθεί για την ταξινόμηση των χαρακτηριστικών βάσει της προσφοράς τους σε συνεχώς επιτυχείς κατατάξεις των υγιών ή των αυτιστικών ατόμων. Το κριτήριο επιλογής των χαρακτηριστικών εξαρτάται από τον επιλογέα χαρακτηριστικών που θα χρησιμοποιηθεί. Η χρήση της επιλογής χαρακτηριστικών για την μείωση της διαστατικότητας έχει σαν στόχο κυρίως την βελτίωση των επιδόσεων του μοντέλου, ώστε να παραχθούν πιο γρήγορα, αποδοτικά μοντέλα και να αποκτηθεί μια βαθύτερη γνώση των υποκείμενων διαδικασιών που παρήγαγαν τα δεδομένα, εάν δηλαδή υπάρχει κάποιο κοινό μοτίβο πίσω από τη δημιουργία τους [100]. Ένας ακόμα λόγος που χρησιμοποιείται η επιλογή χαρακτηριστικών είναι η αποφυγή της δημιουργίας ενός μοντέλου το οποίο έχει προσαρμοστεί στα χαρακτηριστικά και στα δεδομένα που έχει κι έτσι βγάζει βεβιασμένα/λανθασμένα συμπεράσματα σε επόμενες δοκιμές, άλλων δειγμάτων. Στη συνέχεια του κεφαλαίου θα γίνει αναφορά σε αλγορίθμους μείωσης της διαστατικότητας οι οποίοι έχουν προέλθει από τις βιβλιοθήκες scikit-feature και scikit-learn της python.

Στην μηχανική εκμάθηση γίνεται συχνή χρήση εκτιμητών (estimator). Ένας εκτιμητής είναι μια συνάρτηση που θα επιλέξει το πιο ακριβές μοντέλο για τα δεδομένα βασιζόμενο σε πραγματικές παρατηρήσεις. Ο εκτιμητής είναι ο κώδικας (συνάρτηση) που αξιολογεί μια δεδομένη ποσότητα και δημιουργεί μια εκτίμηση. Αυτή η εκτίμηση εισάγεται στη συνέχεια στο σύστημα ταξινόμησης για να προσδιοριστεί τι δράση πρέπει να αναληφθεί ως προς το διαχωρισμό των ομάδων των δεδομένων. Πρακτικά, οι εκτιμητές είναι οι συναρτήσεις που επιτρέπουν τον προσδιορισμό παραμέτρων ενός συνόλου δεδομένων και δίνουν την δυνατότητα σε ένα σύστημα βαθιάς μάθησης να «μαθαίνει». Υπάρχουν δύο κύριες κατηγορίες εκτιμητών βάσει του αποτελέσματος τους. Η πρώτη κατηγορία αποτελείται από εκτιμητές που επιστρέφουν απλές τιμές όπως η διακύμανση που μπορούν στη συνέχεια να χρησιμοποιηθούν από αλγόριθμους ταξινόμησης. Η δεύτερη κατηγορία αποτελείται από ταξινομητές οι οποίοι επιστρέφουν διαστήματα πιθανών τιμών για περεταίρω ανάλυση [101]. Παραδείγματα αλγορίθμων που μπορούν να χρησιμοποιηθούν ως εκτιμητές είναι ο SVM και ο Logistic Regression.

4.1.1 Scikit-Feature και Μείωση Διαστατικότητας

Οι παρακάτω αλγόριθμοι επιλογής χαρακτηριστικών επιλέχθηκαν από τη scikit-feature η οποία είναι μια βιβλιοθήκη που παρέχεται για την γλώσσα

python [102]. Το είδος του κάθε αλγόριθμου επιλογής χαρακτηριστικών καθορίζεται από τις συναρτήσεις που χρησιμοποιεί για να αποφασίσει ποια χαρακτηριστικά είναι σημαντικά και ποια όχι. Έτσι, οι γενικότερες κατηγορίες είναι: βασισμένα στη στατιστική (statistical based), βασισμένα στην ομοιότητα (similarity based), βασισμένα στις πληροφορίες (theoretical information based), βασισμένα σε ροές δεδομένων (streaming based) και βασισμένα στην μάθηση με αραιά δεδομένα (sparse learning based). Γίνεται αναφορά σε αλγορίθμους μηχανικής μάθησης με επίβλεψη, δηλαδή αλγορίθμους στους οποίους δίνονται στην φάση της εκπαίδευσης δείγματα που έχουν ήδη κατηγοριοποιηθεί, αλλά και χωρίς επίβλεψη, αλγορίθμους δηλαδή οι οποίοι δημιουργούν κατηγορίες ταξινόμησης χωρίς πρότερη γνώση σχετικά με αυτές ή τις παρατηρήσεις.

Statistical based

- *Correlation based Feature selection - CFS*

Ο επιλογέας χαρακτηριστικών Correlation-based Feature Selection (CFS) είναι ένας αλγόριθμος με επίβλεψη. Ο CFS αξιολογεί και υπολογίζει τη αξία ενός χαρακτηριστικού εξετάζοντας τον βαθμό συσχέτισης και την δυνατότητα πρόβλεψης που εισάγει στο επιλεγμένο σύνολο χαρακτηριστικών. Οι συσχετίσεις υπολογίζονται με την χρήση της συμμετρικής αβεβαιότητας και στην συνέχεια επιλέγεται ένα υποσύνολο από τα χαρακτηριστικά, με κριτήριο διακοπής την ύπαρξη πέντε διαδοχικών πλήρως αναπτυγμένων υποσυνόλων στα οποία δεν παρουσιάζεται βελτίωση, χρησιμοποιώντας την Best First Search μεθοδολογία.

Τα πλεονεκτήματα του CFS είναι ότι για την πραγματοποίηση της αξιολόγησής του δεν είναι απαραίτητο να κρατήσει κανένα τμήμα των δεδομένων εκπαίδευσης [102].

- *T_score*

Ο επιλογέας χαρακτηριστικών T_score είναι ένας αλγόριθμος με επίβλεψη ο οποίος βρίσκει εφαρμογή στην επίλυση δυαδικών προβλημάτων. Η παρακάτω εξίσωση δίνει τον τύπο υπολογισμού του t_score στην περίπτωση που χρησιμοποιούνται άνισα μεγέθη δειγμάτων και η διακύμανση είναι άνιση. Εδώ ως δείγματα εννοούνται οι παρατηρήσεις (άτομα) για κάθε ομάδα στην οποία μπορούν να ταξινομηθούν, ενώ άνισα μεγέθη σημαίνει ότι ο πληθυσμός των ομάδων είναι άνισος, δηλαδή το πλήθος της μίας ομάδας είναι τουλάχιστον διπλάσιο από το άλλο. Το t_score βασίζεται στον λόγο μεταξύ της μέσης διαφοράς και της διακύμανσης δύο τάξεων, ώστε να εκτιμήσει αν ένα χαρακτηριστικό μεταξύ δυο κλάσεων έχει την δυνατότητα να παρουσιάσει στατιστική διαφορά στην μέση τιμή τους. Η τιμή του t_score είναι ανάλογη της σημαντικότητας ενός χαρακτηριστικού.

$$R_t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \textbf{(15)}$$

όπου τα μ_1, μ_2 δηλώνουν την μέση τιμή του χαρακτηριστικού f_i δυο διαφορετικών κλάσεων (για παράδειγμα υγιής ή ασθενής), τα σ_1^2, σ_2^2 δηλώνουν την τιμή της διακύμανσης του χαρακτηριστικού f_i δυο διαφορετικών κλάσεων και τα n_1, n_2 δηλώνουν τον αριθμό των δειγμάτων δυο διαφορετικών κλάσεων [102].

- *F-score*

Ο επιλογέας χαρακτηριστικών F-score είναι ένας αλγόριθμος με επίβλεψη. Ο αλγόριθμος αυτός κρίνει την δυνατότητα ενός χαρακτηριστικού να διαχωρίσει αποδοτικά δείγματα που ανήκουν σε διαφορετικές κατηγορίες ανεξάρτητα από τα υπόλοιπα χαρακτηριστικά. Δίνει δηλαδή, τη δυνατότητα αξιολόγησης και σύγκρισης της διακριτικής ικανότητας κάθε χαρακτηριστικού. Η τιμή F-score ενός χαρακτηριστικού είναι ανάλογη της διακριτικής ικανότητας προκειμένου να αποφανθεί ο αλγόριθμος ταξινόμησης ως προς την ομάδα στην οποία θα τοποθετήσει το δείγμα [103].

- *Gini_index*

Ο επιλογέας χαρακτηριστικών Gini βασίζεται στην στατιστική και είναι ένας αλγόριθμος με επίβλεψη. Ο Gini βασίζεται στην επιλογή χαρακτηριστικών με κριτήριο την ικανότητα ενός χαρακτηριστικού να ξεχωρίζει μεταξύ δυο κλάσεων. Η τιμή του δείκτη Gini κάθε χαρακτηριστικού είναι ανεξάρτητη από τα άλλα χαρακτηριστικά και υποδηλώνει την σημαντικότητα ενός χαρακτηριστικού. Η τιμή του είναι αντιστρόφως ανάλογη της αξίας του γι' αυτό και πιο σημαντικά είναι τα χαρακτηριστικά με το μικρότερο δείκτη Gini. Στην παρακάτω εξίσωση παρατίθεται ο τύπος υπολογισμού του δείκτη Gini για ένα χαρακτηριστικό f όταν ο αριθμός των κλάσεων είναι c [102] και p(i|f) η πιθανότητα να ανήκει το χαρακτηριστικό f στην κλάση i:

$$gini\ index\ (f) = 1 - \sum_{i=1}^{c} [p(i|f)]^2 \quad \textbf{(16)}$$

Similarity based

- *Fisher Score*

Ο επιλογέας χαρακτηριστικών Fisher Score είναι ένας αλγόριθμος με επίβλεψη που χωρίζει τα χαρακτηριστικά σε κατηγορίες με στόχο οι τιμές χαρακτηριστικών των δειγμάτων μιας κατηγορίας να είναι παρόμοιες μεταξύ τους και να μην παρουσιάζουν ομοιότητα με αυτές από τις άλλες κλάσεις. Στην επόμενη εξίσωση παρατίθεται ο τύπος υπολογισμού του Fisher Score:

$$fisher\_score(f_i) = \frac{\sum_{j=1}^{c} n_j \left(\mu_{ij} - \mu_i\right)^2}{\sum_{j=1}^{c} n_j \sigma_{ij}^2} \quad \textbf{(17)}$$

όπου n_j δηλώνει τον αριθμό των δειγμάτων στην κλάση j, το μ_{ij} δηλώνει την μέση τιμή του χαρακτηριστικού f_i των δειγμάτων που ανήκουν στην κλάση j, το μ_i δηλώνει την μέση τιμή του χαρακτηριστικού f_i, το σ_{ij}^2 δηλώνει την τιμή διακύμανσης του χαρακτηριστικού f_i των δειγμάτων που ανήκουν στην κλάση j [102].

- *Laplacian score*

Ο επιλογέας χαρακτηριστικών Laplacian score είναι ένας αλγόριθμος χωρίς επιτήρηση που έχει ως αρχή ότι η ύπαρξη δυο σημείων που βρίσκονται κοντά μεταξύ τους στο χώρο των παρατηρήσεων, σημαίνει ότι αφορούν στην ίδια κατηγορία. Σημαντική βοήθεια προσφέρουν εδώ τα γραφήματα κοντινότερων γειτόνων, τα οποία διευκολύνουν την κατηγοριοποίηση σε ομάδες των παρατηρήσεων [104].

- *Spec - Spectral*

Ο επιλογέας χαρακτηριστικών Spectral είναι ένας αλγόριθμος χωρίς επιτήρηση ο οποίος έχει παρόμοια βασική ιδέα με το Laplacian Score [105]. Λόγω της πολυπλοκότητας της λειτουργίας του, περαιτέρω ανάλυση ξεπερνάει το επίπεδο αυτής της διπλωματικής.

- *Trace Ratio*

Ο επιλογέας χαρακτηριστικών Trace Ratio είναι ένας αλγόριθμος με επιτήρηση στον οποίο η επιλογή ενός υποσυνόλου χαρακτηριστικών στηρίζεται στη βαθμολογία του συνόλου των χαρακτηριστικών. Ο τρόπος λειτουργίας του περιλαμβάνει την δημιουργία δυο πινάκων συγγένειας για την περιγραφή των ομοιοτήτων των χαρακτηριστικών μεταξύ των κλάσεων και εντός μιας κλάσης [102].

- *Relief*

Ο επιλογέας χαρακτηριστικών Relief είναι ένας επιλογέας χαρακτηριστικών πολλαπλών κλάσεων με επιτήρηση που έχει ως βασική ιδέα τους πλησιέστερους γείτονες. Αναλυτικότερα, επιλέγει μια τυχαία περίπτωση και εντοπίζει τους k πλησιέστερους γείτονές του που ανήκουν στην ίδια κλάση με αυτή και τους k - πλησιέστερους γείτονες οι οποίοι δεν ανήκουν στην ίδια κλάση με αυτή, βάσει τυχαίας (αρχικά) βαθμολογίας του καθενός. Στη συνέχεια επαναϋπολογίζει την βαθμολογία - βάρος ώστε να αυξηθούν οι γείτονες που ανήκουν στην ίδια κατηγορία [106].

Information Theoretical Based

- *Conditional Informative Feature Extraction - CIFE*

Ο επιλογέας χαρακτηριστικών Conditional Informative Feature Extraction - CIFE είναι ένας αλγόριθμος με επιτήρηση που έχει ως βασική ιδέα την επιλογή χαρακτηριστικών τέτοιων ώστε, δοσμένων των πιθανών κατηγοριών, να ελαχιστοποιηθεί η αλληλεξάρτηση χαρακτηριστικών για το τελικό αποτέλεσμα ταξινόμησης δείγματος σε μια κλάση [102],[107].

- *Conditional Mutual Information Maximisation - CMIM*

Ο επιλογέας χαρακτηριστικών CMIM είναι ένας αλγόριθμος με επιτήρηση που βασίζεται στην επιλογή χαρακτηριστικών τα οποία ενώ είναι ανεξάρτητα μεταξύ τους βελτιώνουν την ταξινόμηση, αλλά και εξαρτώνται ελαφρά σε ζευγάρια το ένα από το άλλο. Το CMIM είναι μια προς τα εμπρός επιλογή των χαρακτηριστικών βασισμένη στο κριτήριο των Markov blankets των Koller και Sahami [108]. Εδώ, δημιουργούνται ομάδες χαρακτηριστικών βασισμένες σε ένα ήδη επιλεγμένο χαρακτηριστικό. Στη συνέχεια, γίνεται απόρριψη χαρακτηριστικών μέσω μιας διαδικασίας κατά την οποία κάθε χαρακτηριστικό X μπορεί να απορριφθεί εφόσον υπάρχει ήδη επιλεγμένο χαρακτηριστικό X' τέτοιο ώστε τα χαρακτηριστικά της ομάδας X, Y να είναι ανεξάρτητα μεταξύ τους εφόσον υπάρχει το X' [109].

- *Double Input Symmetrical Relevance - DISR*

Ο επιλογέας χαρακτηριστικών Double Input Symmetrical Relevance – DISR είναι ένας αλγόριθμος με επιτήρηση που βασίζεται σε δύο αρχές. Η πρώτη του αρχή αφορά στην έννοια της μεταβλητής συμπληρωματικότητας, δηλαδή βασίζεται στο γεγονός ότι η αντιμετώπιση των χαρακτηριστικών μεμονωμένα δεν αποφέρει πάντα

τα αποδοτικότερα αποτελέσματα καθώς ο συνδυασμός μεταβλητών ίσως να αποφέρει καλύτερη επιλογή χαρακτηριστικών. Η δεύτερη αρχή στην οποία βασίζεται είναι ότι ο μέσος όρος των πληροφοριών όλων των υποσυνόλων προσδιορίζει το κατώτερο όριο στις πληροφορίες ενός συνόλου μεταβλητών [110].

- *Interactive Cluster Analysis Procedure - ICAP*

Ο επιλογέας χαρακτηριστικών Interactive Cluster Analysis Procedure (ICAP) είναι ένας αλγόριθμος με επιτήρηση που βασίζεται στη μελέτη της αλληλεξάρτησης χαρακτηριστικών και κυρίως στα χαρακτηριστικά που επηρεάζουν όμοια το τελικό αποτέλεσμα ταξινόμησης [111]. Στο αλγόριθμο επιλογής χαρακτηριστικών (ICAP) τα χαρακτηριστικά ταξινομούνται χρησιμοποιώντας την αλληλεπίδραση της ετικέτας της κλάσης με διαφορετικά σύνολα χαρακτηριστικών [112].

- *Joint Mutual Information - JMI*

Ο επιλογέας χαρακτηριστικών Joint Mutual Information (JMI) είναι ένας αλγόριθμος με επιτήρηση που βασίζεται για την επιλογή των χαρακτηριστικών στην αμοιβαία πληροφόρηση μαζί με την εντροπία μεταξύ οποιωνδήποτε τυχαίων μεταβλητών [102].

- *Mutual Information Feature Selection – MIFS*

Ο επιλογέας χαρακτηριστικών Mutual Information Feature Selection (MIFS) είναι ένας αλγόριθμος με επιτήρηση που λαμβάνει υπόψη για την επιλογή των χαρακτηριστικών τη σχετικότητα των χαρακτηριστικών και τον πλεονασμό των χαρακτηριστικών [102]. Ο MIFS εστιάζει αποκλειστικά στις αμοιβαίες πληροφορίες που υπολογίζονται για κάθε στοιχείο – παρατήρηση του αρχικού συνόλου δεδομένων [112].

- *Mutual Information Maximisation - MIM*

Ο επιλογέας χαρακτηριστικών Mutual Information Maximisation (MIM) είναι ένας αλγόριθμος με επιτήρηση που λαμβάνει υπόψη αποκλειστικά την συσχέτιση του χαρακτηριστικού με της ταμπέλες της κλάσης (αν δηλαδή, στην προκειμένη περίπτωση, κάποιος είναι υγιής ή ασθενής). Για αυτό τον επιλογέα η αξία του χαρακτηριστικού είναι ανάλογη με την τιμή της συσχέτισης του [102].

- *Minimal Redundancy Maximal Relevance Criterion - mRMR*

Ο επιλογέας χαρακτηριστικών Minimal Redundancy Maximal Relevance Criterion (mRMR) είναι ένας αλγόριθμος με επιτήρηση. Ο τρόπος επιλογής των χαρακτηριστικών βασίστηκε σε μια σειρά διαισθητικών μέτρων πλεονασμού και σχετικότητας με στόχο την μείωση του πλεονασμού δηλαδή αφαίρεση χαρακτηριστικών των οποίων η επίδραση στο αποτέλεσμα ταξινόμησης υπερκαλύπτεται από άλλα χαρακτηριστικά [113].

Streaming

- *Alpha investing*

Ο επιλογέας χαρακτηριστικών Alpha investing είναι ένας αλγόριθμος με επιτήρηση. Ο τρόπος λειτουργίας του εστιάζει στην δυναμική προσαρμογή του

ορίου μείωσης σφάλματος ταξινόμησης για κάθε νέο χαρακτηριστικό που εισέρχεται στο μοντέλο. Το παραπάνω αποσκοπεί στον έλεγχο του ποσοστού «ψευδούς ανακάλυψης» δηλαδή το ποσοστό των χαρακτηριστικών που εισάγονται στο μοντέλο χωρίς να βελτιώνουν την ακρίβεια ταξινόμησης σε ένα θεωρητικά άπειρο σετ δεδομένων εκπαίδευσης[114].

Sparse learning based

- *Discriminative Feature Selection - UDFS*

Ο επιλογέας χαρακτηριστικών Discriminative Feature Selection (UDFS) είναι ένας αλγόριθμος χωρίς επιτήρηση ο οποίος εστιάζει στην επιλογή των χαρακτηριστικών με βάση τα διακριτά χαρακτηριστικά [115]. Αναλυτικότερα επιδιώκει την μεγαλύτερη δυνατή απόσταση μεταξύ των χαρακτηριστικών δυο διαφορετικών κλάσεων και την ελάχιστη απόσταση μεταξύ των χαρακτηριστικών που ανήκουν στην ίδια κλάση [116].

- *Multi-Cluster Feature Selection - MCFS*

Ο επιλογέας χαρακτηριστικών Multi-Cluster Feature Selection (MCFS) είναι ένας αλγόριθμος χωρίς επιτήρηση. Ο MCFS έχει την δυνατότητα να παρουσιάζει καλό χειρισμό πολλαπλών ομαδοποιημένων δομών. Αυτό οφείλεται στο γεγονός ότι ο MCFS βασίζεται στη μέτρηση των συσχετίσεων μεταξύ διαφορετικών χαρακτηριστικών χωρίς πληροφορίες ετικέτας, δηλαδή χωρίς να λαμβάνει υπόψιν του την κλάση στην οποία ανήκει κάθε δείγμα (στην προκειμένη περίπτωση ο άνθρωπος να είναι υγιής ή ασθενής) [117].

- *Nonnegative Discriminative Feature Selection - NDFS*

Ο επιλογέας χαρακτηριστικών Non negative Discriminative Feature Selection (NDFS) είναι ένας αλγόριθμος χωρίς επιτήρηση. Η επιλογή ενός υποσυνόλου διακριτικών χαρακτηριστικών στηρίζεται στην από κοινού φασματική ομαδοποίηση και την επιλογή χαρακτηριστικών σε ένα κοινό πλαίσιο. Δημιουργεί σε πρώτη φάση μια σειρά ψευδο-κατηγοριών μέσω της φασματικής ανάλυσης και επιβάλλει περιορισμούς σχετικά με την μη αρνητικότητα και ορθογωνιότητα κατά τον υπολογισμό της φάσης για τη φασματική ομαδοποίηση. Η φάση της επιλογής χαρακτηριστικών καθοδηγείται από τις μη αρνητικές ετικέτες ψευδό-κατηγορίας οι οποίες λειτουργούν ως περιορισμοί παλινδρόμησης. Σύμφωνα με τους δημιουργούς του, οι περιορισμοί αυτοί οδηγούν τις αρχικά δημιουργημένες ψευδείς κατηγορίες να προσεγγίσουν τις πραγματικές [118].

4.1.2 Scikit-Learn και Μείωση Διαστατικότητας

Οι παρακάτω αλγόριθμοι επιλογής χαρακτηριστικών επιλεχθήκαν από το scikit-learn που είναι μια βιβλιοθήκη που παρέχεται για την γλώσσα python. Θα γίνει αναφορά για κάθε αλγόριθμο σε μερικές από τις παραμέτρους του scikit-learn που επηρεάζουν την λειτουργία του καθενός και οι οποίες χρησιμοποιήθηκαν στην παρούσα εργασία [76], [77].

Recursive Feature Elimination – RFE

Ο επιλογέας χαρακτηριστικών Recursive Feature Elimination(RFE) έχει ως στόχο την εξάλειψη επαναλαμβανόμενων χαρακτηριστικών χρησιμοποιώντας ως

κριτήριο κατάταξης μία βαθμολογία (βάρος) που έχει αποδοθεί σε κάθε χαρακτηριστικό από έναν εξωτερικό εκτιμητή με επιτήρηση (π.χ. SVM) [119]. Για τον παραπάνω σκοπό πραγματοποιείται μία επαναλαμβανόμενη διαδικασία κατά την οποία ο αριθμός του συνόλου των χαρακτηριστικών μειώνεται καθώς αφαιρούνται τα λιγότερο σημαντικά χαρακτηριστικά σε κάθε επανάληψη μέχρι να καταλήξει στον επιθυμητό αριθμό χαρακτηριστικών [1]. Στον RFE που παρέχεται από τη βιβλιοθήκη scikit-learn ο αριθμός των επιθυμητών χαρακτηριστικών ορίζεται από την μεταβλητή 'n_features_to_select'. Ο αριθμός των χαρακτηριστικών που θα αφαιρεθούν σε κάθε επανάληψη ορίζεται από την μεταβλητή 'step'. Αν η τιμή αυτής είναι μεταξύ 0 και 1 αντιστοιχεί στο ποσοστό των χαρακτηριστικών (στρογγυλοποίηση προς τα κάτω) που θα αφαιρεθούν σε κάθε επανάληψη, αλλιώς αντιστοιχεί στο ακριβές νούμερο. Τέλος, ο εκτιμητής ορίζεται από την μεταβλητή 'estimator' και η αύξηση, μείωση ή απενεργοποίηση πλήρως των μηνύματων που εμφανίζονται στην κονσόλα κατά την εκτέλεση του αλγορίθμου ορίζεται από την μεταβλητή 'verbose'.

Select From Model

Ο επιλογέας χαρακτηριστικών Select From Model χρησιμοποιεί ως κριτήριο κατάταξης το βάρος που έχει αποδοθεί σε κάθε χαρακτηριστικό από έναν εκτιμητή. Ο εκτιμητής αυτός είναι απαραίτητο να περιέχει τη μεταβλητή της σημαντικότητητας των χαρακτηριστικών (feature importances) ή συντελεστών (coefficients) γιατί έχουν καθοριστικό ρολό στην επιλογή των σημαντικών χαρακτηριστικών. Εάν κάποια από τις δύο παραπάνω μεταβλητές ενός χαρακτηριστικού έχει τιμή μικρότερη από το προβλεπόμενο όριο τότε το χαρακτηριστικό αυτό δεν θεωρείται σημαντικό και αφαιρείται [120]. Σημαντικές παράμετροι αυτού του επιλογέα είναι:

- 'threshold': το όριο επιλογής χαρακτηριστικών. Τα χαρακτηριστικά με βάρος μικρότερο από το όριο απορρίπτονται. Μπορεί να είναι 'None' ή 'median' / 'mean'. Στην πρώτη περίπτωση ελέγχεται αν ο εκτιμητής έχει ρυθμισμένη ποινή στις παραμέτρους ίση με L1 οπότε και ορίζεται το όριο σε 10-5. Σε αντίθετη περίπτωση ορίζεται σε 'mean' όπου χρησιμοποιείται ο μέσος όρος των βαρών των χαρακτηριστικών. Ειδικά για το 'mean' μπορεί να συνδυαστεί και με έναν συντελεστή (πχ 1.25*mean).
- 'max_features': η παράμετρος αυτή ορίζει τον μέγιστο αριθμό χαρακτηριστικών πάνω από το 'threshold' τα οποία επιτρέπεται να επιλεχθούν. Μπορεί να απενεργοποιηθεί τελείως η χρήση 'threshold' ορίζοντας το ως np.inf, δηλαδή δίνοντάς του άπειρη τιμή.

Lasso

Το Lasso είναι ένα γραμμικό μοντέλο που υπολογίζει τους «αραιούς» (sparse) συντελεστές. Μπορεί να χρησιμοποιηθεί για την μείωση διαστάσεων του προβλήματος καθώς τείνει να επιλέγει λύσεις που έχουν λιγότερους μη μηδενικούς συντελεστές. Για το λόγο αυτό το μοντέλο Lasso και οι παραλλαγές του έχουν σημαντικό ρόλο για το πεδίο της συμπιεσμένης ανίχνευσης. Στην πράξη, αποτελείται από ένα γραμμικό μοντέλο με έναν πρόσθετο όρο

κανονικοποίησης. Παρακάτω παρατίθεται η αντικειμενική συνάρτηση για ελαχιστοποίηση, η οποία είναι η:

$$\min_{w} \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + a\|w\|_1 \quad \textbf{(18)}$$

Στην εξίσωση αυτή, το $n_{samples}$ είναι ο αριθμός δειγμάτων / παρατηρήσεων, το α είναι μια σταθερά ορισμένη από τον χρήστη και το w το διάνυσμα συντελεστών / βαρών για τα χαρακτηριστικά που έχουν δοθεί.

Η εκτίμηση Lasso επιλύει έτσι την ελαχιστοποίηση της ποινής ελαχίστων τετραγώνων με προσθήκη του όρου $\alpha\|w\|_1$, όπου το α είναι μια σταθερά και $\|w\|_1$ είναι η νόρμα l_1 του συντελεστή διανύσματος.

Η υλοποίηση στην κλάση Lasso χρησιμοποιεί τον αλγόριθμο coordinate descent για την προσαρμογή συντελεστών. Μερικές βασικές παράμετροι του αλγορίθμου Lasso που θα αναφερθούν είναι οι 'tol', 'selection', 'max_iter', 'alpha'. Πιο συγκεκριμένα:

- 'tol': καθορίζει την ανοχή για τη βελτιστοποίηση. Εάν οι ενημερώσεις είναι μικρότερες από το 'tol', ο κώδικας βελτιστοποίησης ελέγχει το διπλό κενό για τη βελτιστοποίηση και συνεχίζει έως ότου αυτό είναι μικρότερο από 'tol'.
- 'selection': σχετίζεται με την ενημέρωση των συντελεστών. Εάν λάβει την τιμή 'cyclic' διαδοχικά ο κάθε συντελεστής θα ενημερώνεται σε κάθε επανάληψη, ενώ εάν οριστεί με τιμή 'random' ένας τυχαίος συντελεστής ενημερώνεται σε κάθε επανάληψη. Η δεύτερη επιλογή συχνά οδηγεί σε σημαντικά ταχύτερη σύγκλιση ειδικά όταν το 'tol' είναι υψηλότερο από 10^{-4}
- 'max_iter'. παράμετρος μέσω της οποίας είναι δυνατός ο καθορισμός του μέγιστου αριθμού επαναλήψεων,
- 'alpha': αποτελεί την σταθερά που πολλαπλασιάζεται με τον όρο L1. Αν δεν δοθεί ισούται με 1. Στην περίπτωση που 'alpha' = 0 τότε ισοδυναμεί με συνηθισμένο τετράγωνο, το οποίο επιλύεται μέσω του LinearRegression.

Select Percentile

Ο επιλογέας χαρακτηριστικών SelectPercentile επιλέγει τα χαρακτηριστικά με κριτήριο το ποσοστό των δεδομένων που έχουν την υψηλότερη βαθμολογία. Την τιμή του ποσοστού αυτού την ορίζει ο χρήστης μέσω της μεταβλητής percentile. Γίνεται χρήση συναρτήσεων για τον υπολογισμό των χαρακτηριστικών που έχουν την υψηλότερη βαθμολογία. Μερικά παραδείγματα συναρτήσεων που μπορούν να χρησιμοποιηθούν για αυτόν το σκοπό αποτελούν η συνάρτηση f_classif και η συνάρτηση mutual_info_classif [77]. Η συνάρτηση f_classif είναι η αντίστοιχη f_score της βιβλιοθήκης scikit-feature της ενότητας 4.1.1. Η συνάρτηση mutual_info_classif βασίζεται στις αμοιβαίες πληροφορίες (Mutual information - MI) μεταξύ δύο τυχαίων μεταβλητών, δηλαδή μελετά την εξάρτηση αυτών χρησιμοποιώντας την εκτίμηση της εντροπίας η οποία έχει προκύψει από τις k-πλησιέστερες αποστάσεις των γειτόνων [121], [122].

Principal component analysis – PCA

Ο επιλογέας χαρακτηριστικών Principal component analysis (PCA) βασίζεται στη γραμμική μείωση διαστάσεων των δεδομένων μέσω της χρήσης Singular Value Decomposition (SVD), ώστε να γίνει προβολή τους σε ένα μικρότερο χώρο διαστάσεων. Το PCA επεξεργάζεται τα δεδομένα εισόδου, ώστε αυτά να είναι κεντραρισμένα αλλά όχι κλιμακωτά για κάθε χαρακτηριστικό που μπορεί να χρησιμοποιηθεί για ταξινόμηση, πριν την εφαρμογή του SVD. Αναλυτικότερα, το PCA έχει ως στόχο να αποσυνθέσει ένα dataset πολλαπλών μεταβλητών σε ένα σύνολο διαδοχικών ορθογώνιων συνιστωσών οι οποίες διατηρούν ως χαρακτηριστικό το μέγιστο ποσό της διακύμανσης. Το PCA της βιβλιοθήκης scikit-learn υλοποιείται ως μια συνάρτηση - αντικείμενο μετασχηματισμού που «μαθαίνει» τις συνιστώσες κατά την εκπαίδευση και μπορεί να χρησιμοποιηθεί στη συνέχεια σε νέα δεδομένα ώστε να τα προβάλει στις συνιστώσες που δημιούργησε.

Μια σημαντική παράμετρος του PCA είναι η ‘svd_solver’ η οποία μπορεί να οριστεί ως ‘auto’, ‘full’, ‘arpack’ και ‘randomized’. Πρόκειται ουσιαστικά για τον αλγόριθμο επίλυσης του προβλήματος SVD και τις παραμέτρους λειτουργίας αυτού.

Εάν η παράμετρος svd_solver οριστεί ως ‘auto’, ο επιλυτής επιλέγεται μεταξύ των ‘full’ και ‘randomized’ βάσει του μεγέθους του πίνακα δεδομένων εισόδου και του αριθμού συνιστωσών στις οποίες θα γίνει οι αποσύνθεση. Εάν οι διαστάσεις των δεδομένων εισόδου είναι πάνω από 500x500 και ο αριθμός των συνιστωσών που θα εξαχθούν είναι χαμηλότερος από το 80% της μικρότερης διάστασης των δεδομένων, τότε η πιο αποτελεσματική ‘randomized’ μέθοδος θα χρησιμοποιηθεί. Διαφορετικά, η ‘full’ αποσύνθεση υπολογίζεται και ενδεχομένως να περικοπεί στη συνέχεια.

Εάν η παράμετρος svd_solver πάρει την τιμή ‘full’, θα χρησιμοποιηθεί ο αλγόριθμος αποσύνθεσης της βιβλιοθήκης γραμμικής άλγεβρας ‘LAPACK’ και μέσω μετέπειτα επεξεργασίας θα γίνει η τελική επιλογή των συνιστωσών.

Εάν για την svd_solver δοθεί η τιμή ‘arpack’, τότε θα εκτελεστεί ο SVD περικομμένος στον αριθμό συνιστωσών που ορίζονται από την μεταβλητή n_components, η οποία ορίζει των πλήθος των χαρακτηριστικών που θα κρατηθούν μετά την εφαρμογή της μείωσης διαστατικότητας. Για την επίλυση του προβλήματος αποσύνθεσης θα χρησιμοποιηθούν εσωτερικά οι αλγόριθμοι της βιβλιοθήκης ARPACK. Απαιτεί αυστηρά η τιμή των συνιστωσών που θα αποσυντεθούν τα δεδομένα εισόδου να είναι μεταξύ 0 και της μικρότερης διάστασης του πίνακα των δεδομένων εισόδου.

Τέλος, εάν οριστεί η svd_solver ως ‘randomized’, θα εκτελεστεί «τυχαιοποιημένο» SVD βασισμένο στην μέθοδο των Halko et al[123].

5 Πειραματική Διάταξη και Αποτελέσματα

Όπως προαναφέρθηκε, σκοπός αυτής της διπλωματικής ήταν η αύξηση της ακρίβειας ταξινόμησης ατόμων ως τυπικά αναπτυσσόμενα ή μη κάνοντας χρήση μεθόδων μηχανικής μάθησης με δεδομένα που αφορούν σε δημογραφικά στοιχεία, συνδεσιμότητα υπολογισμένη από τις χρονοσειρές της λήψης, αλλά και παραμέτρους του μαγνήτη ως είσοδο. Για την επίτευξη του παραπάνω στόχου έγιναν δοκιμές κάνοντας χρήση νευρωνικού δικτύου και διαφόρων άλλων ταξινομητών. Ταυτόχρονα, έγιναν δοκιμές με αλγορίθμους μείωσης όγκου των χαρακτηριστικών που θα χρησιμοποιούνταν ως είσοδος σε έναν ταξινομητή. Τα συστήματα που χρησιμοποιήθηκαν έπρεπε να αποφασίσουν αν ένα άτομο ανήκε στο φάσμα του αυτισμού ή όχι, δηλαδή οι ταξινομητές καλούνταν να επιλύσουν ένα πρόβλημα δύο (2) κλάσεων. Επιπλέον, πραγματοποιήθηκε έρευνα ως προς την ανταπόκριση των προαναφερθέντων όταν χρησιμοποιήθηκαν διαφορετικοί άτλαντες (AAL, HO, CC-200).

5.1 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από απεικονίσεις fMRI του εγκεφάλου ατόμων που βρισκόντουσαν σε κατάσταση ηρεμίας. Τα δεδομένα προήλθαν από την βάση δεδομένων ABIDE. Για την εκτέλεση της πειραματικής διαδικασίας χρησιμοποιήθηκαν διάφοροι άτλαντες από τους οποίους αξιοποιήθηκαν οι περιοχές που αφορούσαν στο DMN. Τέλος, δημιουργήθηκε ο τελικός πίνακας με τα δεδομένα που ήταν επιθυμητό να δοθούν ως είσοδος στον ταξινομητή. Συγκεκριμένα, ο πίνακας αποτελούνταν από την στατική λειτουργική συνδεσιμότητα ακολουθούμενη από τη δυναμική λειτουργική συνδεσιμότητα των προαναφερθέντων περιοχών και κάποια δημογραφικά χαρακτηριστικά. Επιπλέον αυτών των χαρακτηριστικών, στοιχεία όπως η ηλικία των ατόμων, το αν τα άτομα ήταν δεξιόχειρες ή αριστερόχειρες, το φύλο προστέθηκαν στον πίνακα των χαρακτηριστικών, μαζί με τις παραμέτρους της λήψης του μαγνήτη, για κάθε άτομο. Παρακάτω αναλύονται διεξοδικότερα οι λόγοι εκτέλεσης των παραπάνω ενεργειών.

5.1.1 ABIDE

Στην παρούσα διπλωματική, για την υλοποίηση του συστήματος απόφασης, ήταν απαραίτητη η ύπαρξη ικανοποιητικού αριθμού χαρακτηριστικών από άτομα που έχουν την αναπτυξιακή διαταραχή του αυτισμού. Για τον σκοπό αυτό, είναι σημαντικό να χρησιμοποιηθεί μεγάλος αριθμός αξιόπιστων δειγμάτων ως είσοδος στον ταξινομητή. Για την επίτευξη του στόχου αυτού, χρησιμοποιήθηκε η βάση δεδομένων ABIDE (Autism Brain Imaging Data Exchange – ABIDE), η οποία όπως αναφέρθηκε και στο κεφάλαιο 2.2.2 είναι μια βάση δεδομένων από απεικονίσεις του εγκεφάλου ατόμων με διαταραχές αυτιστικού φάσματος και τυπικά αναπτυσσόμενων ατόμων. Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από 1112 άτομα. Η συμμέτοχη διαφορετικών κλινικών για την συλλογή των δεδομένων της ABIDE είχε σαν αποτέλεσμα την ύπαρξη ποικίλων πρωτοκόλλων λήψης fMRI, γεγονός που αύξησε τον αριθμό παραμέτρων που αξιοποιήθηκαν ως είσοδοι στο σύστημα απόφασης που υλοποιήθηκε [31], [35]. Τα δεδομένα που παρέχονται από την ABIDE και χρησιμοποιήθηκαν για την διπλωματική αυτή, προήλθαν από προεπεξεργασία

που έγινε με την βοήθεια του λογισμικού Configurable Pipeline for the Analysis of Connectomes (CPAC), η οποία περιλάμβανε όλα τα στάδια προεπεξεργασίας που αναφέρθηκαν στο κεφάλαιο 2.2.2 και των προγραμμάτων ANTS [124], CIVET [125], FreeSurfer [126]. Αξίζει να αναφερθεί, ότι δεν ήταν το σύνολο των ατόμων διαθέσιμα προς αξιοποίηση για τους χάρτες που χρησιμοποιήθηκαν στα πλαίσια αυτής της διπλωματικής κι έτσι το δείγμα ήταν εξαρχής μειωμένο. Εκτός αυτού του γεγονότος στην παρούσα εργασία χρησιμοποιήθηκαν τα δεδομένα από όσα άτομα δεν είχαν κουνηθεί κατά μέσο όρο πάνω από 0.2mm κατά τη διάρκεια τη λήψης κι έτσι τα άμεσα αξιοποιήσιμα δεδομένα ανέρχονταν σε 883. Επιπροσθέτως, όσα άτομα είχαν μηδενική χρονοσειρά σε κάποια από τις περιοχές ενδιαφέροντος αφαιρέθηκαν καθώς η συσχέτιση αυτής με οποιαδήποτε άλλη χρονοσειρά θα οδηγούσε σε μηδενικό αποτέλεσμα κι έτσι το αποτέλεσμα της συνδεσιμότητας θα ήταν μηδενικό, γεγονός που δε θέλαμε να οδηγήσει τους ταξινομητές σε βεβιασμένα συμπεράσματα.

5.1.2 Άτλας - Περιοχές ενδιαφέροντος και Συνδεσιμότητες

Άτλας - Περιοχές ενδιαφέροντος

Ύστερα από την προεπεξεργασία, πραγματοποιήθηκε προσδιορισμός των περιοχών ενδιαφέροντος (Region of Interest – ROI) που ήταν χρήσιμες για τους σκοπούς αυτής της εργασίας. Χρησιμοποιήθηκαν οι κατάλληλες περιοχές ενδιαφέροντος από 3 διαφορετικούς άτλαντες. Οι περιοχές ενδιαφέροντος ήταν οι περιοχές του κάθε άτλαντα που ανήκαν σε περιοχές του DMN (DMN ROIs). Οι άτλαντες που επιλέχθηκαν ήταν οι ανατομικοί Harvard Oxford (HO), Automated Anatomical Labeling (AAL) και ο λειτουργικός Craddock (CC-200) εκ των οποίων το πλήθος των περιοχών ενδιαφέροντος ήταν 6, 16 και 18 αντίστοιχα.

Χρησιμοποιήθηκε η βιβλιοθήκη FSLeyes (FMRIB's Software Library, version 5.0) κατά την διαδικασία επιλογής περιοχών που δομούν κάθε δίκτυο σε κάθε άτλαντα [43], [127]. Η επιλογή αυτή διευκόλυνε σημαντικά την διαδικασία της οπτικοποίησης περιοχών. Σαν βάση για την επιλογή αυτή, χρησιμοποιήθηκε η εργασία των Smith et al [128].

Τα δεδομένα που χρησιμοποιήθηκαν περιλαμβάνουν στοιχεία που έχουν προκύψει υστέρα από τον υπολογισμό της στατικής και δυναμικής συνδεσιμότητας των περιοχών DMN-ROIs. Ο άτλας HO αποτελείται από 15 στοιχεία στατικής λειτουργικής συνδεσιμότητας και 15 στοιχεία δυναμικής λειτουργικής συνδεσιμότητας, από τα οποία υπολογίστηκαν η μέση τιμή, η διακύμανση, η κυρτότητα και η στρέβλωση. Ο AAL από 120 και τον αντίστοιχο αριθμό χαρακτηριστικών από τη δυναμική λειτουργική συνδεσιμότητα όπως αναφέρθηκε προηγουμένως και ο CC-200 από 153 και 153 αντίστοιχα σε κάθε κατηγορία της δυναμικής λειτουργικής συνδεσιμότητας.

Συνδεσιμότητες

Η ανάλυση της λειτουργικής συνδεσιμότητας μεταξύ περιοχών ενδιαφέροντος (ROIs) είναι από τις πιο ευρέως χρησιμοποιούμενες μεθόδους στην βιβλιογραφία [71], [129]. Η μεταβολή στη λειτουργική συνδεσιμότητα εγκεφάλου (Functional Connectivity – FC) αναμένεται να παρέχει δυνητικούς

βιοδείκτες για ταξινόμηση ή πρόβλεψη διαταραχών του εγκεφάλου [71]. Για τους παραπάνω λόγους επιλέχθηκε να γίνει υπολογισμός της στατικής και της λειτουργικής συνδεσιμότητας των περιοχών ενδιαφέροντος. Στην μελέτη αυτή, για των υπολογισμό των δυναμικών συσχετίσεων χρησιμοποιήθηκε η τεχνική του συρόμενου παραθύρου. Το μήκος του παραθύρου που χρησιμοποιήθηκε κυμαίνονταν μεταξύ 60 - 65 s.

5.1.3 Διασταυρούμενη Επικύρωση – Cross Validation

Για την υλοποίηση της διαδικασίας της ταξινόμησης, απαιτείται αρχικά η «εκπαίδευση» του αλγορίθμου, δηλαδή η δημιουργία του μοντέλου και στην συνέχεια ο έλεγχος αυτού μέσω της αξιολόγησης δεδομένων για τα οποία είναι a-priori γνωστή η κλάση στην οποία ανήκουν. Συνηθίζεται, για να αποφευχθεί να γίνει ο έλεγχος του μοντέλου σε δεδομένα τα οποία έχουν χρησιμοποιηθεί κατά την εκπαίδευση (και άρα να γίνει overfitting, δηλαδή το μοντέλο να υπερ-προσαρμοστεί στην πρόβλεψη ενός συνόλου δεδομένων και να αδυνατεί να κάνει ικανοποιητική πρόβλεψη σε διαφορετικά δεδομένα [130]), να γίνεται ο διαχωρισμός των αρχικών δεδομένων σε 2 ομάδες (δεδομένα εκπαίδευσης και δεδομένα testing) διαφορετικού μεγέθους ώστε να γίνει δημιουργία του μοντέλου από την πρώτη και έλεγχος μέσω της δεύτερης. Ο διαχωρισμός αυτός επιτρέπει τον έλεγχο του μοντέλου με δεδομένα τα οποία ποτέ δεν έχει ξαναδεί. Για να αυξηθεί περαιτέρω η αξία των δεδομένων στην δημιουργία του μοντέλου χρησιμοποιήθηκε η τεχνική k-fold Cross Validation (CV). Στην τεχνική αυτή, γίνεται αρχικά η διάσπαση των δεδομένων εκπαίδευσης σε k κομμάτια ίσου μεγέθους. Στην συνέχεια γίνονται k επαναλήψεις κατά τις οποίες k-1 κομμάτια δεδομένων χρησιμοποιούνται για την δημιουργία του μοντέλου, ενώ το τελευταίο κομμάτι χρησιμοποιείται για την αξιολόγηση του μοντέλου, το οποίο στο τέλος θα εφαρμοστεί στο testing κομμάτι δεδομένων. Η διαδικασία επαναλαμβάνεται και κάθε φορά επιλέγεται διαφορετικό testing κομμάτι για αξιολόγηση. Στο τέλος εξάγεται ο μέσος όρος των μετρικών απόδοσης από όλες τις επαναλήψεις. Το τελικό μοντέλο αξιολογήθηκε τελικά και με τα αρχικά δεδομένα αξιολόγησης τα οποία ποτέ δεν είχε δει κατά την δημιουργία του.

5.1.4 GridSearchCV

Οι τιμές των παραμέτρων των ταξινομητών μπορούν να επηρεάσουν σημαντικά τις επιδόσεις ολόκληρου του μοντέλου. Η αναζήτηση πλέγματος (Grid Search) είναι η διαδικασία του προσδιορισμού των παραμέτρων για την εύρεση βέλτιστων τιμών για ένα συγκεκριμένο μοντέλο. Το GridSearchCV χρησιμοποιεί την τεχνική διασταυρούμενης επικύρωσης (Cross - Validation) για να προσδιοριστεί η τιμή κάθε παραμέτρου η οποία παρέχει τα καλύτερα επίπεδα ακρίβειας, ελαχιστοποιώντας μία τιμή, η οποία θεωρείται κόστος και επιλέγεται από το χρήστη. Συνηθίζεται η τιμή που ψάχνουμε να ελαχιστοποιήσουμε να είναι η ελαχιστοποίηση κάποια μετρικής αξιολόγησης όταν αυτή αφαιρεθεί από τη μονάδα. Η λίστα με τις παραμέτρους και την τιμή εύρους για κάθε παράμετρο του καθορισμένου εκτιμητή, πρέπει να προσδιοριστεί από τον χρήστη μέσω ενός λεξιλογίου το οποίο στην python μπορεί να περιλαμβάνει αρκετές διαφορετικού είδους τιμές τις οποίες θέλουμε να συνδυάσουμε [131].

5.2 Μετρικές αξιολόγησης

Η επιτυχία διαχωρισμού των κλάσεων και κατά συνέπεια η επιτυχία του μοντέλου, αξιολογείται με την βοήθεια μετρικών αξιολόγησης. Στην παρούσα εργασία επιλέχθηκαν να χρησιμοποιηθούν για την αξιολόγηση οι μετρικές της ακρίβειας (accuracy), ευαισθησίας (sensitivity), ειδικότητας (specificity) και περιοχή κάτω από την καμπύλη (Area Under the Curve - AUC). Οι παραπάνω μετρικές επιλέχθηκαν καθώς είναι οι πιο συνηθισμένες στην βιβλιογραφία.

Ακολουθούν παρακάτω οι τύποι που ορίζουν τις μετρικές που αναφέρθηκαν.

$$\text{Ακρίβεια (Accuracy)} = \frac{\text{ΠΘ} + \text{ΠΑ}}{\Pi\Theta + \Pi A + \Lambda\Theta + \Lambda A} \quad \textbf{(19)}$$

$$\text{Ευαισθησία (Sensitivity)} = \frac{\Pi\Theta}{\Pi A + \Lambda A} \quad \textbf{(20)}$$

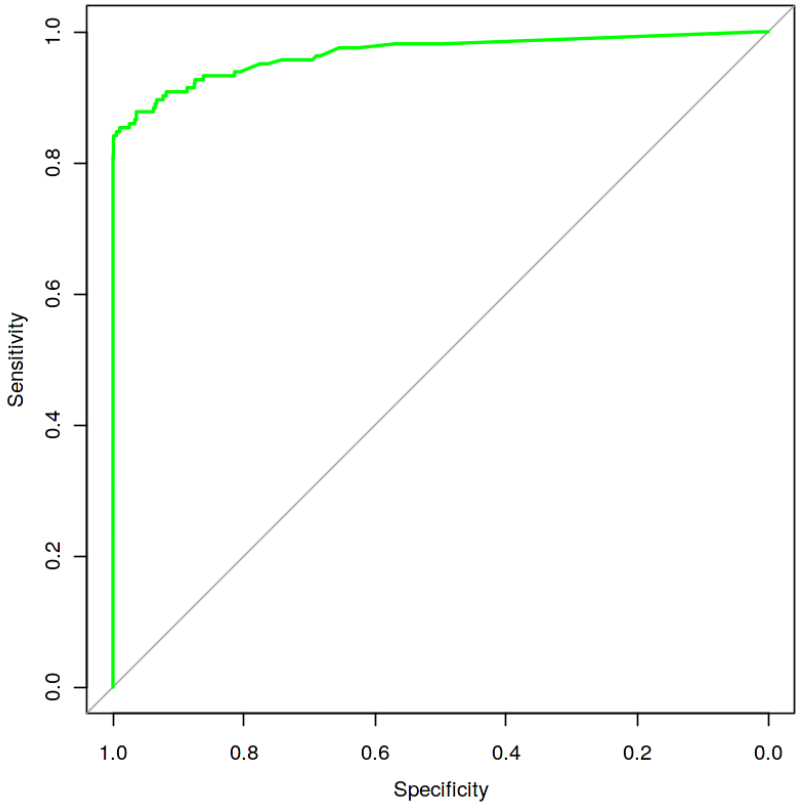
$$\text{Ειδικότητα (Specificity)} = \frac{\Pi A}{\Pi A + \Lambda\Theta} \quad \textbf{(21)}$$

$$\text{AUC} = 1 - \sum_{k=1}^{n} (X_K - X_{k-1})(Y_k + Y_{k-1}) \quad \textbf{(22)}$$

Στις παραπάνω σχέσεις, οι παράμετροι που απαιτούνται είναι οι ΠΘ, ΠΑ, ΛΘ, ΛΑ οι οποίες σημαίνουν αντίστοιχα: Πραγματικά Θετική πρόβλεψη (True Positive – ΠΘ), Πραγματικά Αρνητική πρόβλεψη (True Negative – ΠΑ), Λανθασμένα Θετική πρόβλεψη (False Positive - FP) και τέλος, Λανθασμένα Αρνητική πρόβλεψη (False Negative- ΛΑ). Με τον όρο «Θετική πρόβλεψη» είναι η πρόβλεψη ότι το άτομο είναι υγιές και με τον όρο «Αρνητική πρόβλεψη» είναι η πρόβλεψη ότι το άτομο ανήκει στο φάσμα.

Οι τιμές των προαναφερθέντων μετρικών κυμαίνονται από 1 (απόλυτα ακριβής ταξινόμηση) εως και 0 (απόλυτα ανακριβής). Συνδυάζοντας τις μετρικές της ευαισθησίας και της ειδικότητας, δημιουργείται η καμπύλή ROC (Receiver Operating Characteristic – ROC / Λειτουργικός Χαρακτηριστικός Δέκτης). Το εμβαδό της περιοχής που ορίζεται κάτω από την καμπύλη αυτή είναι η μετρική AUC. Καθώς η καμπύλη ROC είναι αποτέλεσμα των sensitivity και specificity, θα έχει τιμές και αυτή στο σύνολο [0,1] και στους 2 άξονες. Ένα σύστημα με μεγάλη ευστοχία στην ταξινόμηση θα έχει τιμή AUC κοντά στην μονάδα.

Για την εύρεση του αποδοτικότερου ταξινομητή, ικανή και αναγκαία συνθήκη αποτελεί η μετρική αξιολόγησης AUC να έχει την υψηλότερη τιμή σε συνδυασμό με την υψηλότερη τιμή της μετρικής αξιολόγησης της ακριβείας λήψης απόφασης . Η υψηλή τιμή της AUC είναι ανάλογη με τον καλό συνδυασμό της μετρικής της ειδικότητας με την μετρική της ευαισθησίας. Η ύπαρξη καλής ακρίβειας συνδυαστικά με καλή τιμή ευαισθησία και ειδικότητα δηλώνει την ευστάθεια του συστήματος μας.



Εικόνα 7: Παράδειγμα καμπύλης ROC

5.3 Περιβάλλον Υλοποίησης

Για την εκτέλεση του πειράματος χρησιμοποιήθηκε το περιβάλλον της python 3.6.8. Το πλαίσιο (workflow) εύρεσης των καλύτερων παραμέτρων για τον ταξινομητή (hyperparameters) υλοποιήθηκε μέσω του 'GridSearchCV'.

Το workflow αυτό έχει ως στόχο την εύρεση των παραμέτρων της συνάρτησης που ικανοποιούν καλύτερα τον ταξινομητή για τα δεδομένα που έχει στη διάθεσή του. Για τον σκοπό αυτό προσπαθεί να ελαχιστοποιήσει μία τιμή, η οποία μπορεί να δοθεί από μία συνάρτηση της επιλογής του εκάστοτε χρήστη. Για τους σκοπούς αυτής της διπλωματικής η τιμή που επιλέχθηκε ήταν το 1 - AUC.

5.4 Νευρωνικό Δίκτυο για ταξινόμηση

Στα πλαίσια της ταξινόμησης των ατόμων με αυτισμό χρησιμοποιήθηκε ένα νευρωνικό δίκτυο, το οποίο ήταν ο αλγόριθμος MLP που παρέχεται από την βιβλιοθήκη scikit-learn της python. Ο αλγόριθμος βελτιστοποιήθηκε χρησιμοποιώντας το workflow GridSearchCV όπως αναφέρθηκε στην 5.1.4.

5.4.1 Δόκιμες εύρεσης των αποδοτικότερων δεδομένων

Για τον προσδιορισμό της δυναμικής λειτουργικής συνδεσιμότητας χρησιμοποιήθηκαν οι στατιστικές στιγμές της μέσης τιμής, διακύμανσης, λοξότητας και κύρτωσης, οι οποίες υπολογίστηκαν από τον αρχικό τρισδιάστατο πίνακα στον οποίο υπήρχαν όλα τα ζευγάρια συνδεδιμότητας με τρίτη διάσταση το εκάστοτε

παράθυρο. Σ’ αυτή την μελέτη πραγματοποιήθηκαν δοκιμές διαφόρων συνδυασμών των παραπάνω παραμέτρων με σκοπό την εξαγωγή συμπερασμάτων σχετικά με την επιρροή που παρουσίασαν στην ακρίβεια της λήψης απόφασης από τον ταξινομητή.

Αρχικά, εκτελέστηκε ο MLP με τις διαφοροποιήσεις δεδομένων στο τμήμα της δυναμικής λειτουργικής συνδεσιμότητας που θα αναφερθούν στη συνέχεια, κρατώντας σταθερά τις άλλες παραμέτρους που αναφέρονται σε δημογραφικά στοιχεία και παραμέτρους του μαγνήτη. Αναλυτικότερα, στο τμήμα της δυναμικής λειτουργικής συνδεσιμότητας σε κάθε δοκιμή χρησιμοποιήθηκαν τα εξής χαρακτηριστικά από τον αρχικό υπολογισμό της:

- Της μέσης τιμής, διακύμανσης, λοξότητας και κύρτωσης
- Της μέσης τιμής, διακύμανσης
- Της μέσης τιμής
- Της διακύμανσης

Ο Πίνακας 1 παραθέτει τα αποτελέσματα της ακρίβειας λήψης απόφασης από τον ταξινομητή των δοκιμών. Οι παράμετροι του νευρωνικού ορίστηκαν ως

- verbose=1
- learning_rate=‘constant’
- tol=1e-4
- n_iter_no_change=10
- random_state=1
- max_iter=1500
- activation=‘relu’
- solver=‘adam’
- hidden_layer_sizes = 2

ενώ όσες παράμετροι δεν αναφέρονται ήταν οι προκαθορισμένες από τον αλγόριθμο.

Πίνακας 1: Αποτελέσματα των μετρικών αξιολόγησης του νευρωνικού MLP υστέρα από δοκιμές στα δεδομένα εισόδου του χρησιμοποιώντας τους άτλαντες CC-200, HO, AAL.

		Μετρικές αξιολόγησης	**Craddock (CC-200)**	**Harvard Oxford (HO)**	**Automated Anatomical Labeling (AAL)**
Δεδομένα	***Μέση τιμή Διακύμανση Λοξότητα Κύρτωση***	**Ακρίβεια**	0,59	0,58	0,55
		AUC	0,58	0,57	0,5
		Ευαισθησία	0,51	0,42	0
		Ειδικότητα	0,65	0,71	1
	Μέση τιμή Διακύμανση	**Ακρίβεια**	0,65	0,59	0,57
		AUC	0,64	0,57	0,56
		Ευαισθησία	0,57	0,41	0,51
		Ειδικότητα	0,7	0,73	0,61
	Μέση τιμή	**Ακρίβεια**	0,63	0,6	0,56
		AUC	0,62	0,58	0,55
		Ευαισθησία	0,54	0,44	0,49
		Ειδικότητα	0,7	0,73	0,62
	Διακύμανση	**Ακρίβεια**	0,6	0,59	0,58
		AUC	0,59	0,57	0,57
		Ευαισθησία	0,53	0,44	0,52
		Ειδικότητα	0,65	0,7	0,62

Όπως παρατηρείται στον Πίνακας 1 το καλύτερο αποτέλεσμα είναι εκείνο με τιμή ακρίβειας 65% και AUC 64% καθώς συνδυάζει το υψηλότερο AUC με την υψηλότερη ακρίβεια. Το αποτέλεσμα αυτό πρόκυψε χρησιμοποιώντας τον άτλαντα Craddock και μόνο τις στατιστικές στιγμές της μέσης τιμής και της διακύμανσης για την δυναμική λειτουργική συνδεσιμότητα. Λαμβάνοντας υπόψιν τα αποτελέσματα που προέκυψαν, οι δοκιμές της υποενότητας 5.4.2 πραγματοποιήθηκαν χωρίς την συνεισφορά των στατιστικών παραμέτρων της λοξότητας και της κύρτωσης στην δυναμική λειτουργική συνδεσιμότητα.

5.4.2 Δοκιμές ατλάντων, επιλυτών, ρυθμού εκμάθησης

Στην συνέχεια πραγματοποιήθηκε μία περαιτέρω διερεύνηση του άτλαντα που προσδίδει τα καλύτερα αποτελέσματα στο νευρωνικό MLP. Οι άτλαντες οι οποίοι εξετάστηκαν ήταν ο Craddock (CC-200), ο Harvard Oxford (HO) και ο Automated Anatomical Labeling (AAL). Ακόμα μελετήθηκε αν προσδίδει καλύτερα αποτελέσματα για τα δεδομένα ο επιλυτής ‘adam’ ή ο ‘sgd’ και ο ρυθμός εκμάθησης ‘constant’ ή ‘adaptive’ (όλες οι υπόλοιπες μεταβλητές διατηρήθηκαν όπως στην ενότητά 5.4.1). Τα αποτελέσματα αυτών των δοκιμών παρατίθενται στον Πίνακας 2 και στον Πίνακας 3.

Πίνακας 2: Αποτελέσματα των μετρικών αξιολόγησης του νευρωνικού MLP για διαφορετικούς επιλυτές και άτλαντες όταν ο ρυθμός εκμάθησης είναι 'constant' .

		Άτλαντες					
		Craddock (CC-200)		**Harvard Oxford (HO)**		**Automated Anatomical Labeling (AAL)**	
	Επιλυτής	**adam**	**sgd**	**adam**	**sgd**	**adam**	**sgd**
Μετρικές αξιολόγησης	**Ακρίβεια**	0,65	0,67	0,59	0,62	0,57	0,6
	AUC	0,64	0,64	0,57	0,59	0,56	0,59
	Ευαισθησία	0,57	0,42	0,41	0,35	0,51	0,44
	Ειδικότητα	0,7	0,86	0,73	0,82	0,61	0,73

Πίνακας 3: Αποτελέσματα των μετρικών αξιολόγησης του νευρωνικού MLP υπό συνθήκη διαφορετικών επιλυτών και ατλάντων όταν ο ρυθμός εκμάθησης είναι 'adaptive' .

		Άτλαντες					
		Craddock (CC-200)		**Harvard Oxford (HO)**		**Automated Anatomical Labeling (AAL)**	
	Επιλυτής	**adam**	**sgd**	**adam**	**sgd**	**adam**	**sgd**
Μετρικές αξιολόγησης	**Ακρίβεια**	0,62	0,66	0,59	0,6	0,57	0,61
	AUC	0,61	0,64	0,57	0,56	0,56	0,59
	Ευαισθησία	0,54	0,45	0,41	0,32	0,52	0,48
	Ειδικότητα	0,69	0,83	0,73	0,81	0,61	0,7

Όπως παρατηρείται στον Πίνακας 2 το καλύτερο αποτέλεσμα προκύπτει χρησιμοποιώντας τον άτλαντα Craddock με επιλυτή 'sgd' και ρυθμό εκμάθησης 'constant', καθώς παρουσιάζει την υψηλότερη ακρίβεια τιμής 67% και το υψηλότερο AUC τιμής 64%. Συμπερασματικά, οι παραπάνω δοκιμές οδηγήσαν στην εκπαίδευση ενός νευρωνικού MLP που είχε ακρίβεια λήψης απόφασης 67%, AUC 64%, ευαισθησία 42% και ειδικότητα 86% για το αν ένα άτομο έχει την αναπτυξιακή διαταραχή του αυτισμού. Αξιοσημείωτο είναι πως ο άτλας AAL εμφανίζει τις χαμηλότερες επιδόσεις και ο άτλας CC-200 παρουσιάζει τις υψηλότερες.

5.5 Ταξινομητές

Στα πλαίσια της ταξινόμησης των ατόμων με αυτισμό χρησιμοποιήθηκαν ταξινομητές από την βιβλιοθήκη scikit-learn της python. Στην υποενότητα 5.5.1 χρησιμοποιήθηκαν ταξινομητές δέντρων χωρίς μείωση διαστατικότητας, στην υποενότητα 5.5.2 απλοί ταξινομητές χωρίς μείωση διαστατικότητας, ενώ τέλος στην υποενότητα 5.5.3, χρησιμοποιήθηκαν ταξινομητές συνδυαστικά με αλγόριθμους μείωσης της διαστατικότητας των δεδομένων εισόδου τους. Οι

αλγόριθμοι βελτιστοποιήθηκαν χρησιμοποιώντας το workflow ‘GridSearchCV’ όπως αναφέρθηκε στην υποενότητα 5.1.4.

5.5.1 Ταξινομητές δέντρων χωρίς μείωση διαστατικότητας

Οι ταξινομητές δέντρων που εξετάστηκαν είναι οι ‘Decision tree’, ‘AdaBoost’ και ‘Random Forest’. Όπως αναφέρθηκε και στην υποενότητα 5.4.2, ο άτλας CC-200 παρουσίασε τα πιο αποδοτικά αποτελέσματα, οπότε η παρακάτω μελέτη στηρίχθηκε αρχικά στις ενδείξεις του νευρωνικού δικτύου MLP με χρήση του άτλαντα CC-200 για τις δοκιμές που πραγματοποιήθηκαν. Στην συνέχεια θα γίνει μερική ανάλυση των ταξινομητών που εξετάστηκαν καθώς επίσης και παρουσίαση των αποτελεσμάτων τους (Πίνακας 4).

- Decision tree

Πρόκειται για τον αλγόριθμο ‘DecisionTreeClassifier’ του πακέτου ‘scikit-learn’ της python. Μέσω του ‘GridSearchCV’ υλοποιήθηκε η εύρεση των καλύτερων τιμών για την παράμετρο ‘max_depth’ του ταξινομητή. Η βέλτιστη τιμή του ‘max_depth’ ήταν 4. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους.

- AdaBoost

Ο αλγόριθμος ‘AdaBoostClassifier’ του πακέτου ‘scikit-learn’ της python χρησιμοποιήθηκε για αυτή την δοκιμή. Με χρήση του ‘GridSearchCV’ χρησιμοποιήθηκε η βέλτιστη τιμή για την παράμετρο ‘n_estimators’ για τον ταξινομητή, η οποία ήταν 417. Για τις παραμέτρους ‘learning_rate’ και ‘algorithm’ ορίστηκαν οι τιμές 0.5 και ‘SAMME.R’ αντίστοιχα. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους.

- Random Forest

Στην μελέτη αυτή, χρησιμοποιήθηκε ο αλγόριθμος ‘RandomForestClassifier’ του πακέτου ‘scikit-learn’ της python. Χρησιμοποιώντας το ‘GridSearchCV’ χρησιμοποιήθηκε η βέλτιστη τιμή του ‘n_estimators’, η οποία ήταν 507. Για την παράμετρο ‘max_features’ ορίστηκε η τιμή 200. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους.

Πίνακας 4: Αποτελέσματα των μετρικών αξιολόγησης των ταξινομητών δέντρων Decision tree, AdaBoost και Random Forest.

		Μετρικές αξιολόγησης			
		Ακρίβεια	**AUC**	**Ευαισθησία**	**Ειδικότητα**
Ταξινομητές	**Decision tree**	0,61	0,59	0,43	0,75
	AdaBoost	0,59	0,57	0,5	0,65
	Random Forest	0,67	0,65	0,45	0,84

Παρατηρείται ότι ο αλγόριθμος ‘Random Forest’ παρουσιάζει τα καλύτερα αποτελέσματα καθώς επιτυγχάνει την μεγαλύτερη ακρίβεια λήψης απόφασης συνδυαστικά με την μεγαλύτερη τιμή AUC με τιμές 67% και 65% αντίστοιχα.

5.5.2 Ταξινομητές χωρίς μείωση διαστατικότητας

Οι ταξινομητές που εξετάστηκαν είναι οι 'SVM', 'Logistic Regression' και 'KNN'. Λόγω της παρατήρησης της υποενότητας 5.4.2, ο άτλας που χρησιμοποιήθηκε για τις παρακάτω δοκιμές είναι ο 'CC-200'. Τα τελικά αποτελέσματα των ταξινομητών εμφανίζονται στον Πίνακας 5. Οι παράμετροι για κάθε ταξινομητή δίνονται στην συνέχεια.

- Support Vector Machine – SVM

Στην δοκιμή αυτή, χρησιμοποιήθηκε η κλάση 'svm.SVC' του πακέτου 'scikit-learn' της python και ως kernel δόθηκε ο 'rbf'. Χρησιμοποιώντας το GridSearchCV υλοποιήθηκε η εύρεση των καλύτερων τιμών για τις παραμέτρους 'C' και 'gamma' για τον ταξινομητή. Η τιμές αυτές ήταν για το 'C' η τιμή 20 και για το 'gamma' η τιμή 0.0065 (η ακρίβεια της μεταβλητής 'gamma' ανέρχεται σε περισσότερα δεκαδικά ψηφία αλλά για λόγους απλοποίησης της καταγραφής παρατίθενται μόνο τα πρώτα 4). Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους.

- Logistic Regression (LogReg)

Για αυτή την δοκιμή, χρησιμοποιήθηκε η κλάση 'linear_model.LogisticRegression' του πακέτου 'scikit-learn' της python. Χρησιμοποιώντας το 'GridSearchCV' υλοποιήθηκε η εύρεση των καλύτερων τιμών για τις παραμέτρους 'C', 'solver' και 'max_iter' για τον ταξινομητή. Η βέλτιστη τιμή του 'C' ήταν 34, του 'solver' ήταν 'newton-cg' και του 'max_iter' ήταν 5000. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους.

- KNN

Στην μελέτη αυτή, χρησιμοποιήθηκε ο αλγόριθμος 'KNeighborsClassifier του πακέτου 'scikit-learn' της python. Χρησιμοποιώντας το 'GridSearchCV' υλοποιήθηκε η εύρεση των βέλτιστων τιμών για την παράμετρο 'n_neighbors' για τον ταξινομητή, η οποία ήταν η τιμή 2. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους.

Πίνακας 5 : Αποτελέσματα των μετρικών αξιολόγησης των ταξινομητών SVM, Logistic Regression, KNN

		Μετρικές αξιολόγησης			
		Ακρίβεια	**AUC**	**Ευαισθησία**	**Ειδικότητα**
Ταξινομητές	**SVM**	0,63	0,62	0,52	0,72
	LogReg	0,58	0,57	0,51	0,64
	KNN	0,6	0,56	0,28	0,85

Παρατηρείται ότι ο SVM παρουσιάζει τα καλύτερα αποτελέσματα ακρίβειας λήψης απόφασης και AUC με τιμές 63% και 62% αντίστοιχα.

5.5.3 Ταξινομητές με μείωση διαστατικότητας

Στα πλαίσια της ταξινόμησης χρησιμοποιήθηκαν ταξινομητές από την βιβλιοθήκη scikit-learn της python συνδυαστικά με αλγόριθμους μείωσης της

διαστατικότητας των δεδομένων εισόδου τους. Οι αλγόριθμοι μείωσης της διαστατικότητας προήλθαν από δυο βιβλιοθήκες της python, την scikit-learn και την scikit-feature. Η διαδικασία που ακολουθήθηκε σε αυτό το σημείο ήταν να δίνεται η ταξινόμηση όλων των χαρακτηριστικών στον ταξινομητή που θα διαχώριζε αυτιστικά από τυπικά αναπτυσσόμενα άτομα και να επιλέγονται τα 200 καλύτερα από αυτά για τη διαδικασία της αυτής της ταξινόμησης. Για να ελεγχθεί πόσα και ποια από αυτά τα 200 έδιναν το υψηλότερο αποτέλεσμα για την ταξινόμηση και επομένως πέρα από αυτά δεν ήταν απαραίτητο να προστεθούν άλλα για να βελτιστοποιηθεί η απόδοση του αλγόριθμου ταξινόμησης, ο αλγόριθμος έτρεξε 200 φορές, λαμβάνοντας ως είσοδο σε κάθε επανάληψη τον αριθμό των χαρακτηριστικών που θα κράταγε για την επανάληψη αυτή, δηλαδή η παράμετρος αυτή ήταν από 1 μέχρι 200.

5.5.3.1 Αλγόριθμοι για μείωση διαστατικότητας από την βιβλιοθήκη scikit-learn

Σε αυτό το τμήμα της μελέτης πραγματοποιήθηκαν δοκιμές με τους ταξινομητές ‘SVM’, ‘LDA’ και τους αλγορίθμους μείωσης διαστατικότητας ‘rfe’, ‘Select From Model’, ‘Select Percentile’ και ‘PCA’. Μερικοί από τους παραπάνω αλγορίθμους μείωσης διαστατικότητας απαιτούσαν κάποιον εκτιμητή (estimator). Ως estimator χρησιμοποιήθηκαν οι κλάσεις ‘svm.SVC’, ‘Lasso’ και ‘LogReg’.

Select from model (Lasso) - SVM

Πραγματοποιήθηκε δοκιμή του αλγορίθμου Select from model με estimator τον Lasso ενώ ως ταξινομητής για την τελική διαδικασία χρησιμοποιήθηκε ο SVM. Ήταν δυνατός ο ορισμός του αριθμού των στοιχείων – χαρακτηριστικών που θα επιλέγονταν από τον αλγόριθμο επιλογής για να εισαχθούν στην συνέχεια ως δεδομένα εισόδου στον ταξινομητή. Στην μελέτη αυτή, επιλέχθηκε να γίνει εκτέλεση του μοντέλου 200 φορές, ενώ σε κάθε επανάληψη ο αριθμός των χαρακτηριστικών που θα επιλεγόντουσαν από τον αλγόριθμο επιλογής ως δεδομένα εισόδου του ταξινομητή, θα αυξανόταν μοναδιαία ανά επανάληψη ξεκινώντας από το 1 και φτάνοντας στο 200. Ο σκοπός της παραπάνω ενέργειας είναι να βρεθεί ο ελάχιστος αριθμός χαρακτηριστικών που θα δώσει το καλύτερο αποτέλεσμα καθώς επίσης και από ποια χαρακτηριστικά προέκυψε αυτό.

Οι παράμετροι που ορίστηκαν στον estimator Lasso ήταν το tol με τιμή 1e-3, το selection (‘random’), το max_iter με τιμή 1000 και το alpha με τιμή 0.1. Για τον επιλογέα Select from model χρησιμοποιήθηκαν ως estimator ο Lasso, ως threshold το -np.inf και ως max_features η μεταβλητή feats της οποίας η τιμή αυξανόταν σε κάθε επανάληψη ώστε να επιτυγχάνεται η αύξηση κατά ένα του αριθμού των χαρακτηριστικών που θα επιλεγόντουσαν. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους. Ο ταξινομητής SVM χρησιμοποιήθηκε με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στην υποενότητα 5.5.2.

Ως τώρα στην υποενότητα 5.5 χρησιμοποιήθηκε μόνο ο άτλας ‘CC-200’, όμως σε αυτό το σημείο θα εξεταστεί αν άλλοι άτλαντες όπως ο AAL και ο HO μπορούν να δώσουν καλυτέρα αποτελέσματα για τα δεδομένα αυτής της διπλωματικής. Αυτή η διερεύνηση θα πραγματοποιηθεί μέσω αυτής της δοκιμής η οποία θα εκτελεστεί για τον κάθε άτλαντα. Ο Πίνακας 6 περιλαμβάνει το καλύτερο αποτέλεσμα των μετρικών από τις 200 επαναλήψεις του κάθε άτλαντα με κριτήριο την υψηλότερη ακρίβεια σε συνδυασμό με το υψηλότερό AUC.

Πίνακας 6: Αποτελέσματα των μετρικών αξιολόγησης του SVM με χρήση Select from model με estimator τον Lasso υπό συνθήκη των ατλάντων AAL, HO, CC-200.

		Μετρικές αξιολόγησης			
		Ακρίβεια	**AUC**	**Ευαισθησία**	**Ειδικότητα**
Άτλας	**Craddock (CC-200)**	0,64	0,62	0,52	0,73
	Harvard Oxford (HO)	0,61	0,58	0,33	0,83
	Automated Anatomical Labeling (AAL)	0,56	0,55	0,44	0,66

Χρησιμοποιώντας το GridSearchCV υλοποιήθηκε η εύρεση των καλύτερων παραμέτρων για τις παραμέτρους 'C' και 'gamma' για τον ταξινομητή. Η βέλτιστη τιμή που χρησιμοποιήθηκε του 'C' και του 'gamma' για τον AAL ήταν 5 και 0.0020 αντίστοιχα, για τον HO 5 και 0.0022 , ενώ για τον CC-200 24 και 0.0060 (στην μεταβλητή 'gamma' ανέρχεται σε περισσότερα δεκαδικά ψηφιά η ακρίβεια αλλά λογούς απλής καταγραφής παρατίθενται τα πρώτα 4). Παρατηρείται πως ο Άτλας AAL εμφανίζει τις χαμηλότερες επιδόσεις και ότι ο άτλας CC-200 παρουσιάζει τις καλύτερες από όλους με AUC 62% και ακρίβεια 64% . Επομένως δεν ήταν άστοχη η επιλογή του CC-200 στα προηγούμενα πειράματα που έγιναν στην υποενότητα 5.5. Για τα επόμενα πειράματα της υποενότητας 5.5.3 θα χρησιμοποιηθεί μόνο ο άτλας CC-200, δεδομένων των αποτελεσμάτων που λάβαμε.

RFE (SVM) - SVM

Η δοκιμή που πραγματοποιήθηκε σε αυτό το σημείο σχετίζεται με τον αλγόριθμο RFE, με estimator την κλάση SVC του SVM ενώ ως ταξινομητής χρησιμοποιήθηκε ο SVM. Ήταν δυνατός ο ορισμός του αριθμού των στοιχείων που επιλέχθηκαν από τον αλγόριθμο επιλογής για χρήση στην συνέχεια ως δεδομένα εισόδου του ταξινομητή. Το μοντέλο έτρεξε 200 φορές ώστε να παρθούν τα αποτελέσματα για τις εκπαιδεύσεις από το 1ο καλύτερο μέχρι και τα 200 καλύτερα χαρακτηριστικά από το σύνολο. Οι παράμετροι που ορίστηκαν στον estimator svm.SVC ήταν ο kernel σε 'linear', το gamma σε 0.0020 (η ακρίβεια της μεταβλητής 'gamma' περιέχει περισσότερα δεκαδικά ψηφιά αλλά λογούς απλής καταγραφής αναφέρονται τα πρώτα 4). και το C σε 5. Οι τιμές αυτές βρέθηκαν υστέρα από δοκιμή του ταξινομητή svm.SVC με kernel σε 'linear' στα αρχικά δεδομένα χρησιμοποιώντας το workflow GridSearchCV για τις μεταβλητές gamma και C. Οι παράμετροι που ορίστηκαν στον επιλογέα RFE ήταν ως estimator ο svm.SVC, το step=1, το verbose στην τιμή 0 και στο n_features_to_select η μεταβλητή feats της οποίας η τιμή αυξανόταν σε κάθε επανάληψη ώστε να επιτυγχάνεται η αύξηση κατά ένα του αριθμού των χαρακτηριστικών που θα επιλεγόντουσαν. Οι παράμετροι που δεν αναφέρθηκαν

είχαν τις προεπιλεγμένες τιμές τους. Ο ταξινομητής SVM χρησιμοποιήθηκε με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στην υποενότητα 5.5.2. Ο άτλας που χρησιμοποιήθηκε ήταν ο CC-200.

Το καλύτερο αποτέλεσμα των μετρικών αξιολόγησης από τις 200 επαναλήψεις του SVM σε συνδυασμό με τον RFE και ως estimator τον svm.SVC ήταν για την ακρίβεια / AUC / ευαισθησία / ειδικότητα 64/63/53/72 (%), αντίστοιχα. Χρησιμοποιώντας το GridSearchCV υλοποιήθηκε η εύρεση των καλύτερων παραμέτρων για τις παραμέτρους 'C' και 'gamma' για τον ταξινομητή. Η βέλτιστη τιμή που χρησιμοποιήθηκε για το 'C' ήταν 28 και για το 'gamma' ήταν 0.0070 (για την μεταβλητή 'gamma' για λογούς απλής καταγραφής παρατίθενται τα πρώτα 4 δεκαδικά ψηφία παρόλο που ανέρχεται σε περισσότερα δεκαδικά ψηφία η ακρίβεια της).

RFE(LogReg) - SVM

Στην συνέχεια πραγματοποιήθηκε ξανά το παραπάνω πείραμα της προηγούμενης υποενότητας με ακριβώς τις ίδιες παραμέτρους με μόνη διαφορά τον estimator του RFE ο οποίος αντικαταστάθηκε με τον LogReg.

Στον estimator LogReg ορίστηκε η παράμετρος C με την τιμή 34, το max_iter σε 5000 και ως solver ορίστηκε ο 'newton-cg'. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προκαθορισμένες τιμές τους. Στην μελέτη αυτή, επιλέχθηκε να τρέξει 200 φορές το μοντέλο και ο άτλας που χρησιμοποιήθηκε ήταν ο CC-200.

Το καλύτερο αποτέλεσμά ακρίβειας λήψης απόφασης από τις 200 επαναλήψεις του SVM σε συνδυασμό με τον RFE, με estimator τον LogReg είχε ποσοστά ακρίβειας / AUC / ευαισθησίας / ειδικότητας 63/62/54/69, αντίστοιχα. Χρησιμοποιώντας το GridSearchCV, υλοποιήθηκε η εύρεση των καλύτερων παραμέτρων για τις παραμέτρους 'C' και 'gamma' για τον ταξινομητή. Η βέλτιστη τιμή που χρησιμοποιήθηκε για το 'C' ήταν 30 και για το 'gamma' ήταν 0.0026 (για λογούς απλής καταγραφής για την μεταβλητή 'gamma' παρατίθενται τα πρώτα 4 δεκαδικά ψηφία ενώ περιέχει περισσότερα δεκαδικά ψηφία η ακρίβεια της). Παρατηρείται ότι η χρήση του SVM ως estimator για τον RFE απέδωσε καλύτερη ακρίβεια και AUC από την χρήση του LogReg.

Select percentile - SVM

- Select percentile (score_func=function f_classif) - SVM

Η δοκιμή που πραγματοποιήθηκε αφορούσε τον αλγόριθμο Select percentile με score_function την συνάρτηση f_classif. Ως ταξινομητής χρησιμοποιήθηκε ο SVM. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προκαθορισμένες τιμές τους. Ο ταξινομητής SVM χρησιμοποιήθηκε με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στην υποενότητα 5.5.2. Στην μελέτη αυτή, ο άτλας που χρησιμοποιήθηκε ήταν ο CC-200.

Το αποτέλεσμά των μετρικών αξιολόγησης του ταξινομητή SVM σε συνδυασμό με τον αλγόριθμο μείωσης διαστατικότητας Select percentile με score_function την συνάρτηση f_classif ήταν για την ακρίβεια / AUC / ευαισθησία / ειδικότητα ήταν 62/59/40/79 (%), αντίστοιχα. Χρησιμοποιώντας το GridSearchCV υλοποιήθηκε η εύρεση των καλύτερων παραμέτρων για τις παραμέτρους 'C' και 'gamma' για τον ταξινομητή. Η βέλτιστη τιμή που χρησιμοποιήθηκε για το 'C' ήταν 20 και για το 'gamma' ήταν 0.0126 (παρατίθενται τα πρώτα 4 δεκαδικά ψηφία της μεταβλητής

‘gamma’ αλλά περιέχει περισσότερα δεκαδικά ψηφία η ακρίβεια της που δεν παρατίθενται για λόγους απλότητας).

- Select percentile (score_func= mutual_info_classif) - SVM

Η δοκιμή που πραγματοποιήθηκε σε αυτό το σημείο αφορούσε τον αλγόριθμο Select percentile με score_function την συνάρτηση mutual_info_classif και ως ταξινομητής χρησιμοποιήθηκε ο SVM. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προκαθορισμένες τιμές τους. Ο ταξινομητής SVM χρησιμοποιήθηκε με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στην υποενότητα 5.5.2. Στην μελέτη αυτή, ο άτλας που χρησιμοποιήθηκε ήταν ο CC-200. Το αποτέλεσμα των μετρικών ακρίβειας / AUC / ευαισθησίας / ειδικότητας του SVM με Select percentile και με συνάρτηση σκορ (score_function) την συνάρτηση mutual_info_classif ήταν 59/55/25/86 (%), αντίστοιχα. Χρησιμοποιώντας το GridSearchCV υλοποιήθηκε η εύρεση των καλύτερων παραμέτρων για τις παραμέτρους ‘C’ και ‘gamma’ για τον ταξινομητή. Η βέλτιστη τιμή που χρησιμοποιήθηκε για το ‘C’ ήταν 20 και για το ‘gamma’ ήταν 0.0144 (για λόγους απλότητας παρατίθενται μόνο τα 4 πρώτα δεκαδικά ψηφία της παραμέτρου ‘gamma’).

Παρατηρείται λοιπόν ότι υπό τις ίδιες συνθήκες άτλα και δεδομένων ανάμεσα στους παραπάνω αλγορίθμους, τα καλυτέρα αποτελέσματα δόθηκαν από το μοντέλο RFE με estimator την κλάση SVC του SVM με ταξινομητή τον SVM και το μοντέλο με αλγόριθμο επιλογής χαρακτηριστικών το Select from model με estimator τον Lasso και με ταξινομητή τον SVM.

Δεδομένα - Select from model (Lasso) - SVM

Σε προηγούμενη υποενότητα χρησιμοποιήθηκε το μοντέλο αλγορίθμου μείωσης διαστατικότητας ‘Select from model’ με εκτιμητή τον αλγόριθμο Lasso και ταξινομητή τον SVM για να μελετηθεί ο άτλας με τον οποίο προκύπτουν τα καλυτέρα αποτελέσματα. Παρατηρήθηκε ότι αυτό το μοντέλο έδωσε και το καλύτερο αποτέλεσμα. Σε αυτή την υποενότητα θα χρησιμοποιηθεί αυτό το μοντέλο για να διερευνηθούν τα δεδομένα με αντίστοιχο τρόπο όπως έγινε στην υποενότητα 5.4.1. Ο σκοπός αυτής της ενέργειας είναι να εξεταστεί αν μπορεί να βελτιωθεί το αποτέλεσμα αφαιρώντας κάποια χαρακτηριστικά. Αναλυτικότερα, έγιναν δύο πειράματα τα οποία διέφεραν στις στατιστικές στιγμές που λήφθηκαν υπόψη για την δυναμική λειτουργική συνδεσιμότητα. Στο πρώτο πείραμα για τον υπολογισμό της δυναμικής λειτουργικής συνδεσιμότητας στον πίνακα των δεδομένων χρησιμοποιήθηκαν στατιστικές στιγμές της μέσης τιμής, της διακύμανσης, της λοξότητας και της κύρτωσης. Αντίθετα, στο δεύτερο πείραμα για τον υπολογισμό της δυναμικής λειτουργικής συνδεσιμότητας, χρησιμοποιήθηκαν στον πίνακα των δεδομένων στατιστικές στιγμές μέσης τιμής και διακύμανσης. Σ’ αυτή την μελέτη πραγματοποιήθηκαν οι παραπάνω δοκιμές με σκοπό την εξαγωγή συμπερασμάτων σχετικά με την επίδραση τους στην ακρίβεια λήψης απόφασης του ταξινομητή, όσον αφορά στα χαρακτηριστικά που μπορούμε να προσθέσουμε και αφορούν τη βελτίωση της απόφασης. Στο μοντέλο δοθήκαν ακριβώς οι ίδιες παράμετροι που δόθηκαν στην προηγούμενη υποενότητα και ο άτλας που χρησιμοποιήθηκε ήταν και πάλι ο CC-200. Και για τα δυο πειράματα το μοντέλο έτρεξε 200 φορές ώστε να ληφθούν σαν αποτέλεσμα για τις εκπαιδεύσεις από το 1ο καλύτερο μέχρι και τα

200 καλύτερα χαρακτηριστικά από το σύνολο. Ο Πίνακας 7 περιέχει για το κάθε πείραμα τα καλύτερα αποτελέσματά του από τις 200 επαναλήψεις του SVM ως ταξινομητή και συνδυαστικά με των αλγόριθμο επιλογής χαρακτηριστικών τον Select from model που είχε ως estimator τον Lasso. Τα καλυτέρα αποτελέσματα επιλέχθηκαν με κριτήριο τις υψηλότερες τιμές του συνδυασμού ακρίβειας και AUC.

Πίνακας 7: Αποτελέσματα των Μετρικών αξιολόγησης του SVM

		Δεδομένα	
		Μέση τιμή, διακύμανση, λοξότητα, κύρτωση	**Μέση τιμή, διακύμανση**
Μετρικές αξιολόγησης	**Ακρίβεια**	0,64	0,66
	AUC	0,62	0,64
	Ευαισθησία	0,52	0,46
	Ειδικότητα	0,73	0,81

Χρησιμοποιώντας το GridSearchCV υλοποιήθηκε η εύρεση των καλύτερων παραμέτρων για τις παραμέτρους 'C' και 'gamma' για τον ταξινομητή. Στο πρώτο πείραμα η βέλτιστη τιμή που χρησιμοποιήθηκε για το 'C' ήταν 24 και για το 'gamma' ήταν 0.0060 και στο δεύτερο ήταν 12 και 0.0020 αντίστοιχα (η μεταβλητή 'gamma' ανέρχεται σε περισσότερα δεκαδικά ψηφία η ακρίβεια αλλά για λόγους απλής καταγραφής παρατίθενται τα πρώτα 4). Όπως φαίνεται και στον παραπάνω πίνακα, το καλύτερο αποτέλεσμα προκύπτει χρησιμοποιώντας τις στατιστικές στιγμές μέσης τιμής και διακύμανσης για την δυναμική λειτουργική συνδεσιμότητα, καθώς δίνουν το υψηλότερο AUC σε συνδυασμό με την υψηλότερη ακρίβεια. Λαμβάνοντας υπόψιν τα αποτελέσματα που προέκυψαν, η συνέχεια όλων των παρακάτω μελετών πραγματοποιήθηκε χωρίς την συνεισφορά των στατιστικών τιμών της λοξότητας και της κύρτωσης στην δυναμική λειτουργική συνδεσιμότητα από τα δεδομένα.

RFE (SVM) - LDA

Η δοκιμή που πραγματοποιήθηκε ήταν αξιοποιώντας τον αλγόριθμο RFE για τη διαδικασία της επιλογής χαρακτηριστικών, με estimator την κλάση SVC του SVM ενώ ως ταξινομητής χρησιμοποιήθηκε ο LDA. Οι παράμετροι που ορίστηκαν στον estimator svm.SVC ήταν ο kernel σε 'linear', το gamma σε 0.0020 και το C σε 5. Οι τιμές αυτές βρέθηκαν υστέρα από δοκιμή χρησιμοποιώντας ως ταξινομητή του svm.SVC με kernel σε 'linear' στα αρχικά δεδομένα χρησιμοποιώντας το workflow GridSearchCV για τις μεταβλητές gamma και C. Οι παράμετροι που ορίστηκαν στον επιλογέα RFE ήταν ως estimator ο svm.SVC, το step στην τιμή 1, το verbose στην τιμή 0 και στο n_features_to_select δόθηκε η μεταβλητή feats της οποίας η τιμή αυξανόταν σε κάθε επανάληψη ώστε να επιτυγχάνεται η αύξηση κατά ένα του αριθμού των χαρακτηριστικών που θα επιλεγόντουσαν. Για τον ταξινομητή LDA χρησιμοποιήθηκε το workflow GridSearchCV για την μεταβλητή solver. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους. Ο άτλας που χρησιμοποιήθηκε ήταν ο CC-200 και η δοκιμή πραγματοποιήθηκε χωρίς την συνεισφορά των στατιστικών παραμέτρων της λοξότητας και της κύρτωσης στην δυναμική λειτουργική συνδεσιμότητα από τα δεδομένα. Το μοντέλο έτρεξε 200

φορές ώστε να ληφθούν τα αποτελέσματα για τις εκπαιδεύσεις από το 1ο καλύτερο μέχρι και τα 200 καλύτερα χαρακτηριστικά από το σύνολο.

Το καλύτερο αποτέλεσμά από τις 200 επαναλήψεις του LDA (με solver τον lsqr) σε συνδυασμό με τον RFE και με estimator τον SVM είχε ακρίβεια / AUC / ευαισθησία / ειδικότητα 57/56/49/63 (%) αντίστοιχα.

PCA – SVM

Η δοκιμή που πραγματοποιήθηκε αφορούσε τον αλγόριθμο PCA για την επιλογή των καλύτερων χαρακτηριστικών, ενώ ως ταξινομητής χρησιμοποιήθηκε ο SVM. Οι παράμετροι που ορίστηκαν στον επιλογέα PCA ήταν η μεταβλητή 'svd_solver' ως 'auto' και η μεταβλητή 'n_components' ως 'feats' της οποία η τιμή αυξανόταν σε κάθε επανάληψη ώστε να επιτυγχάνεται η αύξηση κατά ένα του αριθμού των χαρακτηριστικών που θα επιλεγόντουσαν. Οι παράμετροι που δεν αναφέρθηκαν είχαν τις προεπιλεγμένες τιμές τους. Ο ταξινομητής SVM χρησιμοποιήθηκε με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στην υποενότητα 5.5.2. Ο άτλας που χρησιμοποιήθηκε ήταν ο CC-200 και η δοκιμή πραγματοποιήθηκε χωρίς την συνεισφορά των στατιστικών παραμέτρων της λοξότητας και της κύρτωσης στην δυναμική λειτουργική συνδεσιμότητα από τα δεδομένα. Το μοντέλο έτρεξε 200 φορές ώστε να ληφθούν τα αποτελέσματα για τις εκπαιδεύσεις από το 1ο καλύτερο μέχρι και τα 200 καλύτερα χαρακτηριστικά από το σύνολο.

Το καλύτερο αποτέλεσμά από τις 200 επαναλήψεις του SVM σε συνδυασμό με τον PCA ήταν 64/63/46/80 (%). Τα καλύτερα αποτελέσματα επιλέχθηκαν με κριτήριο τις υψηλότερες τιμές του συνδυασμού ακρίβειας και AUC. Χρησιμοποιώντας το GridSearchCV οι καλύτερες τιμές που δοθήκαν στο 'C' και 'gamma' για τον ταξινομητή ήταν 5 και 0.0030 αντίστοιχα (για λογούς απλής καταγραφής παρατίθενται μόνο τα πρώτα 4 δεκαδικά ψηφιά της μεταβλητή 'gamma').

5.5.3.2 Αλγόριθμοι για μείωση διαστατικότητας από scikit-feature

Στην παραπάνω υποενότητα υπήρξε η παρατήρηση ότι ο άτλας AAL εμφανίζει τις χαμηλότερες επιδόσεις. Η παρατήρηση αυτή λήφθηκε υπόψιν στις υπόλοιπες δοκιμές των αλγορίθμων μείωσης διαστατικότητας της scikit-learn, οπότε και εστιάστηκε το ενδιαφέρον στους άτλαντες CC-200 και HO. Λαμβάνοντας υπόψιν τα συμπεράσματα που προέκυψαν από την υποενότητα «Δεδομένα - Select from model (Lasso) - SVM», στα δεδομένα που χρησιμοποιήθηκαν η συνεισφορά της δυναμικής λειτουργικής συνδεσιμότητας στη λήψη απόφασης από τον ταξινομητή περιορίστηκε στην χρήση μόνο των παραμέτρων της μέσης τιμής και της διακύμανσης, ενώ τα υπόλοιπα χαρακτηριστικά, όπως είναι η στατική λειτουργική συνδεσιμότητα, τα δημογραφικά στοιχεία αλλά και οι παράμετροι που χρησιμοποιήθηκαν στο μαγνήτη για τη λήψη, διατηρήθηκαν.

Στην παρούσα εργασία επιλέχθηκαν να χρησιμοποιηθούν οι εξής αλγόριθμοι για μείωση της διαστατικότητας: Cife, Laplacian, Cfs, Cmim, Disr, Fisher, F_score, Mim, Gini, Icap, Jmi, Mcfs, Mifs, Mrmr, Ndfs, ReliefF, Spec, T_score, Trace_ratio, Udfs, Alpha investing. Στη συνέχεια τα χαρακτηριστικά τα οποία επιλέχθηκαν από τους παραπάνω αλγόριθμους δόθηκαν ως είσοδοι στους

ταξινομητές SVM και KNN. Οι ταξινομητές SVM και KNN χρησιμοποιήθηκαν με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στην υποενότητα 5.5.2.

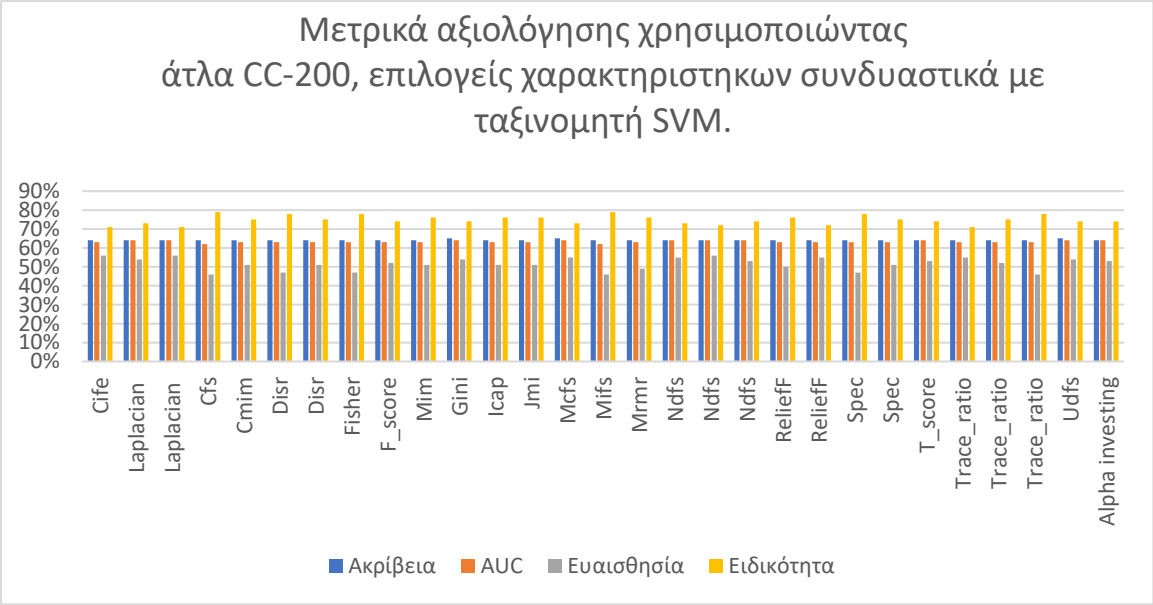
Κάθε μοντέλο έτρεξε 200 φορές ώστε να ληφθούν τα αποτελέσματα για τις εκπαιδεύσεις από το 1ο καλύτερο μέχρι και τα 200 καλύτερα χαρακτηριστικά από το σύνολο. Ο Πίνακας 8 περιλαμβάνει τα καλύτερα αποτελέσματα, με κριτήριο τον υψηλότερο συνδυασμό ακριβείας και AUC, των μετρικών αξιολόγησης απόφασης των 200 επαναλήψεων κάθε δοκιμών με τον άτλαντα CC-200 ενώ ο Πίνακας 10 περιλαμβάνει τα αποτελέσματα με τον Άτλα HO . Χρησιμοποιώντας το GridSearchCV δοθήκαν οι καλύτερες τιμές στο 'C' και στο 'gamma' για τον ταξινομητή SVM και στο 'n_neighbors' για τον ταξινομητή KNN. Οι βέλτιστες τιμές των παραμέτρων 'C' και 'gamma' του ταξινομητή SVM και 'n_neighbors' του ταξινομητή KNN που χρησιμοποιήθηκαν για τα καλύτερα αποτελέσματα των δοκιμών που έγιναν με τον άτλαντα CC-200 παρατίθενται στον Πίνακας 9. Στις δοκιμές που έγιναν με τον άτλαντα HO, ορίστηκαν για τον ταξινομητή SVM οι παράμετροι 'C' όπως εμφανίζεται στον Πίνακας 11 και 'gamma'=0,0020. Για την παράμετρο 'n_neighbors' του ταξινομητή KNN χρησιμοποιήθηκε η βέλτιστη τιμή η οποία ήταν 6. Για την μεταβλητή 'gamma' για λογούς απλής καταγραφής παρατίθενται τα πρώτα 4 δεκαδικά ψηφία παρόλο που ανέρχεται σε περισσότερα δεκαδικά ψηφία η ακρίβεια της. Στις εικόνες 8-11 γίνεται σύγκριση των αποτελεσμάτων.

Πίνακας 8: Πίνακας αποτελεσμάτων των μετρικών αξιολόγησης των ταξινομητών SVM, KNN κάνοντας χρήση επιλογέων από την scikit-feature - ATLAS CC-200

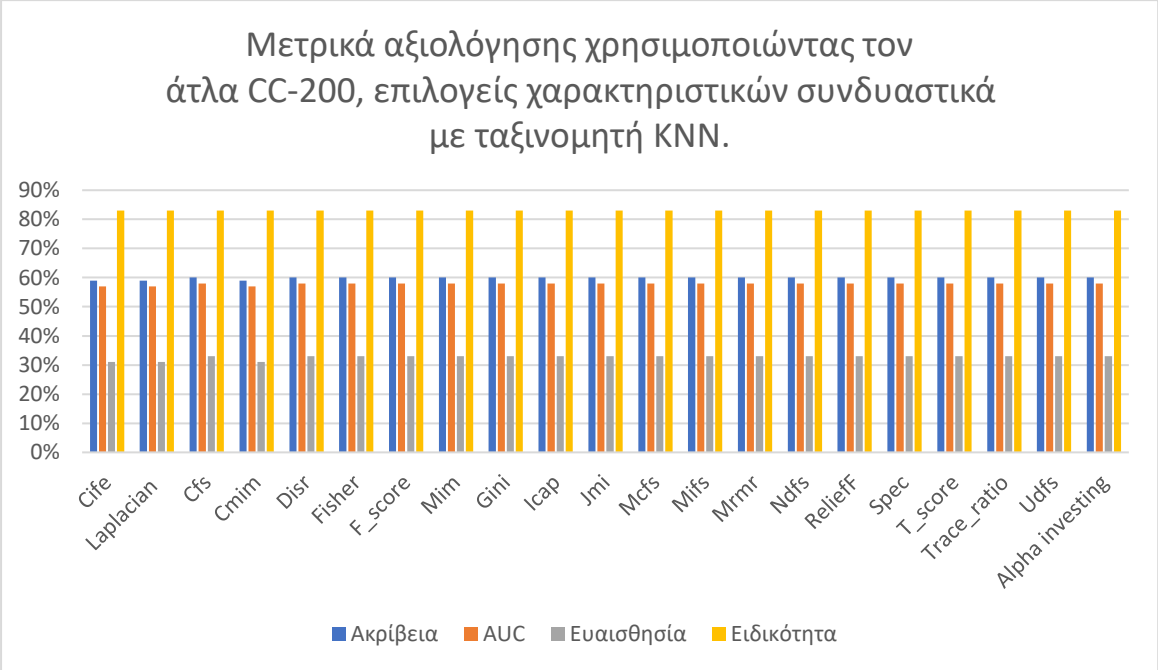
	CC-200	*Ταξινομητές*							
		SVM				KNN			
	Μετρικές αξιολόγησης	**Ακρίβ.**	**AUC**	**Ευαισθ.**	**Ειδικ.**	**Ακρίβ.**	**AUC**	**Ευαισθ.**	**Ειδικ.**
Αλγόριθμοι μείωσης διαστατικότητας	**Cife**	0,64	0,63	0,56	0,71	0,59	0,57	0,31	0,83
	Laplacian	0,64	0,64	0,54	0,73	0,59	0,57	0,31	0,83
		0,64	0,64	0,56	0,71				
	Cfs	0,64	0,62	0,46	0,79	0,6	0,58	0,33	0,83
	Cmim	0,64	0,63	0,51	0,75	0,59	0,57	0,31	0,83
	Disr	0,64	0,63	0,47	0,78	0,6	0,58	0,33	0,83
		0,64	0,63	0,51	0,75				
	Fisher	0,64	0,63	0,47	0,78	0,6	0,58	0,33	0,83
	F_score	0,64	0,63	0,52	0,74	0,6	0,58	0,33	0,83
	Mim	0,64	0,63	0,51	0,76	0,6	0,58	0,33	0,83
	Gini	0,65	0,64	0,54	0,74	0,6	0,58	0,33	0,83
	Icap	0,64	0,63	0,51	0,76	0,6	0,58	0,33	0,83
	Jmi	0,64	0,63	0,51	0,76	0,6	0,58	0,33	0,83
	Mcfs	0,65	0,64	0,55	0,73	0,6	0,58	0,33	0,83
	Mifs	0,64	0,62	0,46	0,79	0,6	0,58	0,33	0,83
	Mrmr	0,64	0,63	0,49	0,76	0,6	0,58	0,33	0,83
	Ndfs	0,64	0,64	0,55	0,73	0,6	0,58	0,33	0,83
		0,64	0,64	0,56	0,72				
		0,64	0,64	0,53	0,74				
	ReliefF	0,64	0,63	0,5	0,76	0,6	0,58	0,33	0,83
		0,64	0,63	0,55	0,72				
	Spec	0,64	0,63	0,47	0,78	0,6	0,58	0,33	0,83
		0,64	0,63	0,51	0,75				
	T_score	0,64	0,64	0,53	0,74	0,6	0,58	0,33	0,83
	Trace_ratio	0,64	0,63	0,55	0,71	0,6	0,58	0,33	0,83
		0,64	0,63	0,52	0,75				
		0,64	0,63	0,46	0,78				
	Udfs	0,65	0,64	0,54	0,74	0,6	0,58	0,33	0,83
	Alpha investing	0,64	0,64	0,53	0,74	0,6	0,58	0,33	0,83

Πίνακας 9: Τιμές παραμέτρων των ταξινομητών SVM, KNN που χρησιμοποιήθηκαν για τα αποτελέσματα του προηγούμενου πίνακα.

Παράμετροι ταξινομητή				
Αλγόριθμοι μείωσης διαστατικότητας **CC-200**	**SVM**		**KNN**	
	gamma	**c**	**n_neighbors**	**Επανάληψη SVM**
Cife	0,0022	30	4	4 & 9
Laplacian	0,0034	16	4	62
	0,0022	30		91 &173
Cfs	0,0033	5	6	168
Cmim	0,0020	14	4	150
Disr	0,0029	6	6	14
	0,0020	14		174
Fisher	0,0029	6	6	8
F_score	0,0020	16	6	3
Mim	0,0020	14	6	161 & 17
Gini	0,0037	14	6	37
Icap	0,0020	14	6	151
Jmi	0,0020	14	6	164 &170
Mcfs	0,0029	20	6	15
Mifs	0,0033	5	6	13
Mrmr	0,0022	10	6	10
Ndfs	0,0029	20	6	23
	0,0022	30		33
	0,0033	16		48
ReliefF	0,0042	7	6	27
	0,0022	27		30
Spec	0,0029	6	6	26 & 36
	0,0020	14		35 & 176
T_score	0,0033	16	6	36
Trace_ratio	0,0020	30	6	19
	0,0020	16		23
	0,0029	5		24
Udfs	0,0033	18	6	18
Alpha investing	0,0048	12	6	6



Εικόνα 8: Σύγκριση των αποτελεσμάτων του ταξινομητή SVM για κάθε μετρική αξιολόγησης υπό συνθήκη διαφορετικών αλγορίθμων μείωσης διαστατικότητας. Οι δοκιμές έγιναν με άτλα τον CC-200.



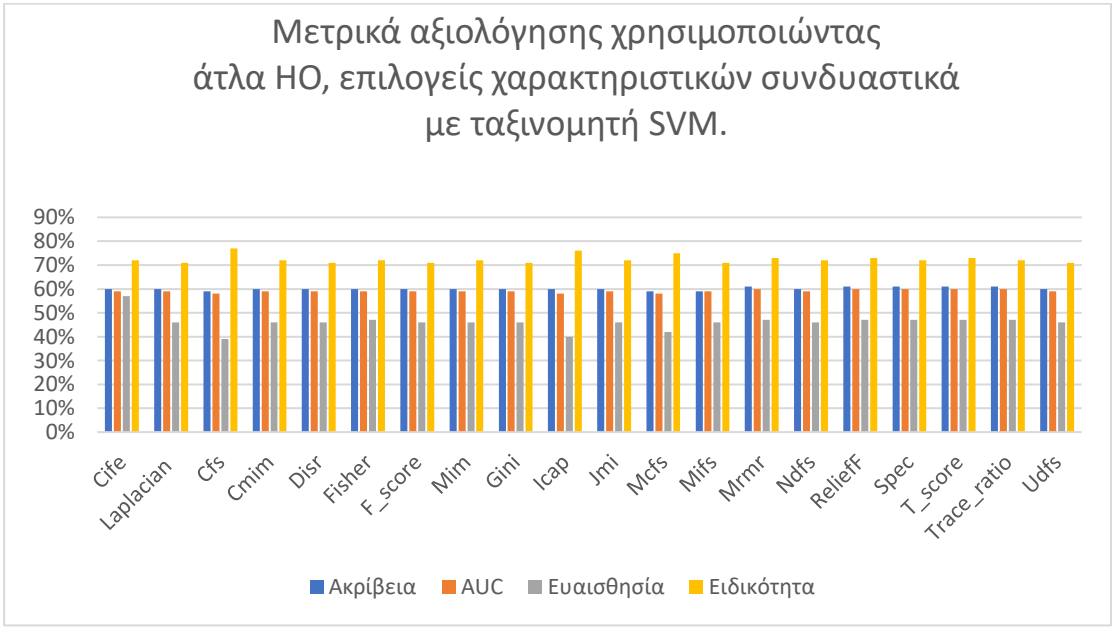
Εικόνα 9: Σύγκριση των αποτελεσμάτων του ταξινομητή KNN για κάθε μετρική αξιολόγησης υπό συνθήκη διαφορετικών αλγορίθμων μείωσης διαστατικότητας. Οι δοκιμές έγιναν με άτλα τον CC-200

Πίνακας 10: Πίνακας αποτελεσμάτων των μετρικών αξιολόγησης των ταξινομητών SVM, KNN κάνοντας χρήση επιλογέων από την scikit-feature - ATLAS HO

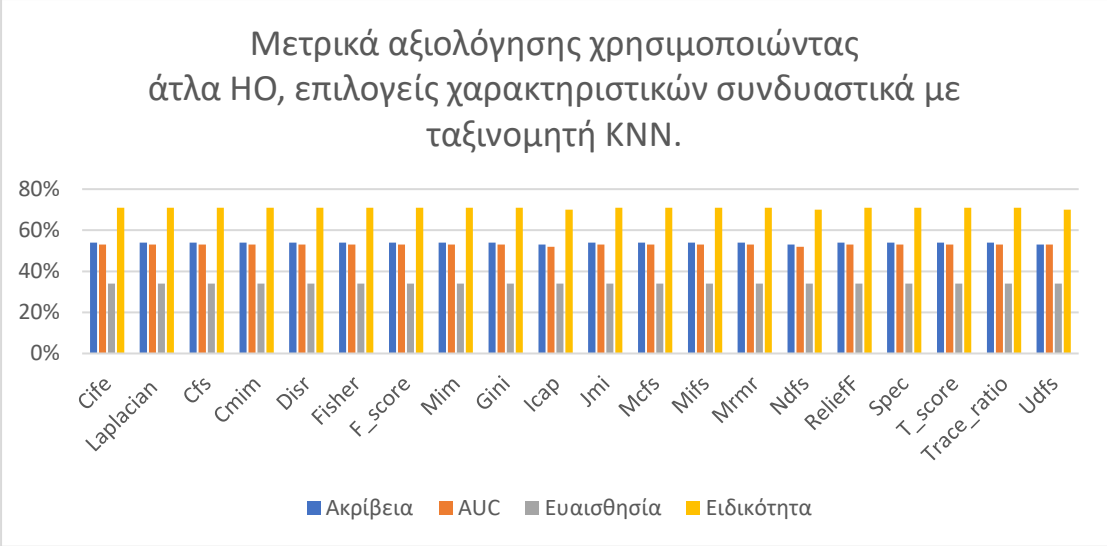
	Άτλας HO	*Ταξινομητές*							
		SVM				**KNN**			
	Μετρικές αξιολόγησης	**Ακρίβεια**	**AUC**	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**AUC**	**Ευαισθησία**	**Ειδικότητα**
Αλγόριθμοι επιλογής χαρακτηριστικών	**Cife**	0,6	0,59	0,57	0,72	0.54	0,53	0,34	0,71
	Laplacian	0,6	0,59	0,46	0,71	0.54	0,53	0,34	0,71
	Cfs	0,59	0,58	0,39	0,77	0.54	0,53	0,34	0,71
	Cmim	0,6	0,59	0,46	0,72	0,54	0,53	0,34	0,71
	Disr	0,6	0,59	0,46	0,71	0,54	0,53	0,34	0,71
	Fisher	0,6	0,59	0,47	0,72	0.54	0,53	0,34	0,71
	F_score	0,6	0,59	0,46	0,71	0.54	0,53	0,34	0,71
	Mim	0,6	0,59	0,46	0,72	0.54	0,53	0,34	0,71
	Gini	0,6	0,59	0,46	0,71	0.54	0,53	0,34	0,71
	Icap	0,6	0,58	0,4	0,76	0,53	0,52	0,34	0,7
	Jmi	0,6	0,59	0,46	0,72	0.54	0,53	0,34	0,71
	Mcfs	0,59	0,58	0,42	0,75	0.54	0,53	0,34	0,71
	Mifs	0,59	0,59	0,46	0,71	0.54	0,53	0,34	0,71
	Mrmr	0,61	0,6	0,47	0,73	0.54	0,53	0,34	0,71
	Ndfs	0,6	0,59	0,46	0,72	0,53	0,52	0,34	0,7
	ReliefF	0,61	0,6	0,47	0,73	0.54	0,53	0,34	0,71
	Spec	0,61	0,6	0,47	0,72	0.54	0,53	0,34	0,71
	T_score	0,61	0,6	0,47	0,73	0.54	0,53	0,34	0,71
	Trace_ratio	0,61	0,6	0,47	0,72	0.54	0,53	0,34	0,71
	Udfs	0,6	0,59	0,46	0,71	0,53	0,53	0,34	0,7

Πίνακας 11: Τιμές της παραμέτρου 'C' του ταξινομητή SVM που χρησιμοποιήθηκαν για τα αποτελέσματα του προηγούμενου πίνακα (άτλας HO).

Αλγόριθμοι μείωσης διαστατικότητας	C
Cife	22
Laplacian	19
Cfs	5
Cmim	22
Disr	19
Fisher	22
F_score	19
Mim	18
Gini	20
Icap	8
Jmi	18
Mcfs	10
Mifs	28
Mrmr	23
Ndfs	18
ReliefF	23
Spec	30
T_score	23
Trace_ratio	24,26
Udfs	20,23



Εικόνα 10: Σύγκριση των αποτελεσμάτων του ταξινομητή SVM για κάθε μετρική αξιολόγησης υπό συνθήκη διαφορετικών αλγορίθμων μείωσης διαστατικότητας. Οι δοκιμές έγιναν με τον άτλα HO.



Εικόνα 11: Σύγκριση των αποτελεσμάτων του ταξινομητή KNN για κάθε μετρική αξιολόγησης υπό συνθήκη διαφορετικών αλγορίθμων μείωσης διαστατικότητας. Οι δοκιμές έγιναν με άτλα τον HO.

Ο άτλας CC-200 παρουσίασε αποτελέσματα ακρίβειας λήψης απόφασης που κυμαίνονταν από 64% έως 65% με ταξινομητή τον SVM και 59% έως 60% με ταξινομητή τον KNN ενώ ο άτλας HO παρουσίασε αποτελέσματα που κυμαίνονταν μεταξύ 59% έως 61% με ταξινομητή τον SVM και 53% έως 54% με ταξινομητή τον KNN. Παρατηρείται λοιπόν ότι ο άτλας CC-200 παρουσιάζει με διαφορά περίπου 5% καλύτερα αποτελέσματα από τον άτλα HO. Για το λόγο αυτό για τις επόμενες δοκιμές δεν κρίθηκε αναγκαία η εξέταση των αποτελεσμάτων του άτλαντα HO. Ακόμα, από τα παραπάνω αποτελέσματα εξάγεται το συμπέρασμα ότι ο ταξινομητής SVM παρουσίασε περίπου 5% καλύτερα αποτελέσματα από τον ταξινομητή KNN.

6 Συμπεράσματα

Στην αρχή αυτής της εργασίας αναφέρθηκε ότι τα τελευταία χρόνια απασχολεί την ερευνητική κοινότητα το ερώτημα κατά πόσο είναι εφικτό μέσω της μηχανικής μάθησης να ληφθεί ένα ικανοποιητικό ποσοστό αυτοματοποιημένης διάγνωσης για το αν ένα άτομο ανήκει στο φάσμα του αυτισμού ή όχι. Πάνω σε αυτό βασίστηκε και η παρούσα διπλωματική, με στόχο τη μελέτη και τη βελτίωση μεθόδων ώστε να παραχθούν όσο το δυνατόν καλύτερα αποτελέσματα κατάταξης ατόμων στις δύο ομάδες. Για τον σκοπό αυτό έγιναν σειρές πειραμάτων ώστε να μελετηθεί πως επηρεάζεται το αποτέλεσμα ταξινόμησης από τον άτλαντα που θα χρησιμοποιηθεί, από την επίδραση τμημάτων των δεδομένων (στατιστικές στιγμές από μία τιμή), από τον ταξινομητή και τις παραμέτρους αυτού, καθώς και από το αν έγινε χρήση αλγορίθμων μείωσης διαστατικότητας συνδυαστικά με τον επιλεγμένο αλγόριθμο. Τα πειράματα αποσκοπούσαν στην εύρεση των παραμέτρων που θα έδιναν τα καλύτερα αποτελέσματα και την απόρριψη εκείνων που έδιναν τα ασθενέστερα ώστε να δημιουργηθεί το βέλτιστο μοντέλο για τα δεδομένα.

6.1 Συνεισφορά στατιστικών παραμέτρων στην δυναμική λειτουργική συνδεσιμότητα

Αρχικά μελετήθηκε η επίδραση τμημάτων των δεδομένων που προκύπτουν από την δυναμική λειτουργική συνδεσιμότητα στα αποτελέσματα της ταξινόμησης. Για τον προσδιορισμό της δυναμικής λειτουργικής συνδεσιμότητας χρησιμοποιήθηκαν οι στατιστικές στιγμές της μέσης τιμής, διακύμανσης, λοξότητας και κύρτωσης και πραγματοποιήθηκαν δοκιμές διαφόρων συνδυασμών των παραπάνω παραμέτρων. Στα πειράματα που πραγματοποιήθηκαν για την μελέτη της συνεισφοράς των στατιστικών στιγμών από μία τιμή στην δυναμική λειτουργική συνδεσιμότητα της παρατηρήθηκε ότι ο συνδυασμός της διακύμανσης και της μέσης τιμής παρουσίαζε τα καλύτερα αποτελέσματα ακρίβειας λήψης απόφασης. Εξάγεται λοιπόν το συμπέρασμα ότι αυτές οι παράμετροι έχουν ενισχυτικό ρολό για την σωστή κατάταξη των ατόμων ως τυπικά αναπτυσσόμενων ή μη, ενώ οι παράμετροι λοξότητας και κύρτωσης είτε συνδυαστικά με τις παραπάνω παραμέτρους είτε μεμονωμένα παρουσιάζουν πιο αδύναμα αποτελέσματα κατάταξης.

6.2 Σύγκριση των ατλάντων CC-200, HO, AAL

Πραγματοποιήθηκε επίσης, μελέτη μέσω δοκιμών της επίδρασης των διαφορετικών ατλάντων στα αποτελέσματα της ταξινόμησης. Παρατηρώντας τα αποτελέσματα των μοντέλων που δοκιμάστηκαν για την σύγκριση των ατλάντων Craddock (CC-200) , Harvard Oxford (HO) και Automated Anatomical Labeling (AAL), παρατηρείται ότι ο άτλας Craddock (CC-200) αποτελούσε σε όλες τις περιπτώσεις την καλύτερη επιλογή. Ο χάρτης AAL στην πλειοψηφία των αποτελεσμάτων που έδωσε αποτελούσε την πιο αδύναμη επιλογή συγκριτικά με τους άλλους δυο.

6.3 Σύγκριση Ταξινομητών

Ένα ακόμα ερώτημα που απαιτούταν να απαντηθεί ήταν ποιος ταξινομητής είναι ο ιδανικότερος για τα δεδομένα. Για τον σκοπό αυτό πραγματοποιήθηκαν δοκιμές με νευρωνικό δίκτυο, με ταξινομητές δέντρων και με άλλες κατηγορίες ταξινομητών χωρίς, ή και με την χρήση αλγορίθμων μείωσης διαστατικότητας. Τα συμπεράσματα των δοκιμών αναλύονται περεταίρω παρακάτω. Αξίζει να αναφερθεί ότι στην μελέτη της αποδοτικότητας, υπό τις ίδιες συνθήκες άτλαντα και στατιστικών παραμέτρων, του ταξινομητή SVM χωρίς ή με μείωση διαστατικότητας παρατηρήθηκαν τα ακόλουθα. Το μοντέλο χωρίς μείωση διαστατικότητας παρουσίαζε ακρίβεια λήψης απόφασης 63% και AUC 62%. Ενώ τα μοντέλα που χρησιμοποίησαν αλγορίθμους μείωσης διαστατικότητας παρουσίασαν ακρίβεια λήψης απόφασης που έφτασε το 64% και AUC 62%. Διεξάγεται από τα παραπάνω το συμπέρασμα ότι η χρήση αλγορίθμων μείωσης διαστατικότητας μπορεί να προσδώσει αποδοτικότερα μοντέλα.

6.3.1 Σύγκριση ταξινομητών χωρίς μείωση διαστατικότητας

Οι αλγόριθμοι δέντρων που εξετάστηκαν ήταν οι Decision tree, AdaBoost και Random Forest. Εξ αυτών ο αλγόριθμος Random Forest ήταν εκείνος που παρουσίασε τα αποδοτικότερα αποτελέσματα επιτυγχάνοντας ακρίβεια λήψης απόφασης 67% και AUC 65%. Προέκυψε ότι τα δέντρα είναι πολλά υποσχόμενα για την ταξινόμηση ατόμων στο φάσμα του αυτισμού. Οι ταξινομητές που δοκιμάστηκαν ήταν οι SVM, KNN και LogReg, και παρουσίασε το καλύτερο αποτέλεσμα ο SVM με ακρίβεια λήψης απόφασης 63% και AUC 62%.

Για τους παραπάνω ταξινομητές χρησιμοποιήθηκε ο άτλας CC-200 και οι στατιστικές στιγμές της μέσης τιμής, της διακύμανσης, της λοξότητας και κύρτωσης για τον υπολογισμό της διακύμανσης της δυναμικής λειτουργικής συνδεσιμότητας. Προς μελλοντική έρευνα θεωρείται σκόπιμο να επαναληφθούν τα πειράματα αυτής της υποενότητας χρησιμοποιώντας μόνο τις στατιστικές στιγμές της μέσης τιμής και της διακύμανσης για τον υπολογισμό της δυναμικής λειτουργικής συνδεσιμότητας καθώς φαίνονται πολλά υποσχόμενα για την βελτίωση των αποτελεσμάτων τους.

Εξετάστηκε επίσης το νευρωνικό δίκτυο MLP, για το οποίο πραγματοποιήθηκαν δοκιμές για την εύρεση του αποδοτικότερου άτλαντα, των αποδοτικότερων στατιστικών παραμέτρων για τον υπολογισμό της δυναμικής λειτουργικής συνδεσιμότητας και των αρτιότερων παραμέτρων για το πρόβλημα που εξετάζεται.

Διαπιστώθηκε ότι το καλύτερο αποτέλεσμα δόθηκε όταν ορίστηκε ως άτλας ο CC-200, ως στατιστικές στιγμές η μέση τιμή και η διακύμανση, ως επιλυτής ο ‘sgd’ και ως ρυθμός εκμάθησης το ‘constant’. Το καλύτερο αποτέλεσμα παρουσίασε ακρίβεια λήψης απόφασης 67% και AUC 64%.

6.3.2 Σύγκριση ταξινομητών με μείωση διαστατικότητας

Στην προσπάθεια περαιτέρω βελτίωσης των αποτελεσμάτων, στράφηκε το ενδιαφέρον σε δοκιμές μοντέλων ταξινομητών συνδυαστικά με αλγόριθμους μείωσης χαρακτηριστικών, δηλαδή διαστάσεων. Η πλειοψηφία των δοκίμων επικεντρώθηκε στην χρήση του SVM ως ταξινομητή καθώς αποτελεί έναν από τους δημοφιλέστερους στην βιβλιογραφία για την επίλυση τέτοιων προβλημάτων.

Υστέρα από πολλαπλές δοκιμές αλγορίθμων μείωσης διαστατικότητας το καλύτερο αποτέλεσμα που επιτεύχθηκε παρουσίαζε ακρίβεια λήψης απόφασης 66% και AUC 64% και προήλθε από το μοντέλο με αλγόριθμο μείωσης διαστατικότητας 'Select from model' που είχε εκτιμητή τον αλγόριθμο Lasso και ταξινομητή τον SVM. Το δεύτερο καλύτερο αποτέλεσμα που επιτεύχθηκε παρουσίαζε ακρίβεια λήψης απόφασης 65% και AUC 64% και προήλθε από τον ταξινομητή SVM και τους αλγορίθμους μείωσης διαστατικότητας Udfs, Gini και Mcfs.

Και τα 4 καλυτέρα αποτελέσματα που αναφέρθηκαν προέκυψαν με χρήση του άτλαντα CC-200 και την συνεισφορά των στατιστικών παραμέτρων στην δυναμική λειτουργική συνδεσιμότητα του συνδυασμού της διακύμανσης και της μέσης τιμής.

6.4 Εύρεση δυνητικών βιοδεικτών

Κρίθηκε εύλογη η περαιτέρω παρατήρηση των χαρακτηριστικών που επιλέχθηκαν από τους αλγορίθμους επιλογής χαρακτηριστικών για την πιθανή εύρεση δυνητικών βιοδεικτών που σχετίζονται άμεσα με τον αυτισμό. Μέσω του παραρτήματος, είναι δυνατή η αντιστοίχιση κάθε χαρακτηριστικού με τις περιοχές του εγκέφαλου που χρησιμοποιήθηκαν για τον υπολογισμό της λειτουργικής συνδεσιμότητας που το αποτελούν.

Για τον σκοπό αυτό εξετάστηκαν τα χαρακτηριστικά που επέλεξαν οι αλγόριθμοι που έδωσαν τα καλυτέρα αποτελέσματα, δηλαδή με ταξινομητή SVM για τους αλγορίθμους Cife, Cfs, Cmim, Disr, Fisher, F_score, Mim, Gini, Icap, Jmi, Mcfs, Mifs, Mrmr, Ndfs, ReliefF, Spec, T_score, Trace_ratio, Udfs, Alpha investing. Τα μοντέλα έκαναν την χρήση του άτλαντα CC-200, ενώ για τα δεδομένα που προήλθαν από την δυναμική λειτουργική συνδεσιμότητα χρησιμοποιήθηκαν μόνο οι στατιστικές στιγμές της μέσης τιμής και της διακύμανσης. Αναλυτικότερα, επιλέχθηκαν για κάθε αλγόριθμο τα πρώτα 20 κοινά χαρακτηριστικά από τους 10 πίνακες που δημιουργούνται για κάθε επανάληψη του αλγορίθμου επιλογής χαρακτηριστικών (10-fold – nested Cross Validation). Τα κοινά χαρακτηριστικά κάθε αλγορίθμου παρατίθενται στον Πίνακας 12. Στις περιπτώσεις που ένας αλγόριθμος παρουσίαζε πάνω από μια φορά το καλύτερο αποτέλεσμά του, έγινε εντοπισμός των κοινών χαρακτηριστικών που επιλέχθηκαν κάθε φορά που επαναλήφθηκε το καλύτερο αποτέλεσμά του και στην συνέχεια συγκεντρώθηκαν έως 20 κοινά χαρακτηριστικά που εντοπιστήκαν μεταξύ των επαναλήψεων. Όποιος αλγόριθμος δεν παρείχε επαρκή αριθμό επιλεγμένων χαρακτηριστικών ώστε ο αριθμός των κοινών χαρακτηριστικών να είναι τουλάχιστον 20, παρατίθεται στον προαναφερθέντα πίνακα με όσα κοινά χαρακτηριστικά ήταν εφικτό να εντοπιστούν και με τα υπόλοιπα κελιά του κενά.

Πίνακας 12: Κοινά χαρακτηριστικά (αναφέρονται οι κωδικοί αυτών όπως εμφανίζονται και στο παράρτημα) από τους πίνακες που δημιουργούνται σε κάθε επανάληψη του αλγορίθμου επιλογής χαρακτηριστικών

Cife	Cfs	Cmim	Disr	Fisher	F_score	Mim	Gini	Icap	Jmi	Mcfs	Mifs	Mrmr	Ndfs	ReliefF	Spec	T_score	Trace_ratio	Udfs	Alpha investing
1	1	1	1	40	40	1	462	1	1	318	1	1	153	283	380	40	283	452	1
470	2	2	2	283	193	2	463	2	2	317	469	469	306	470	287	283	40	467	2
471	3	3	3	193	283	3	461	3	3	316	468	468	160	130	134	193	193	473	4
467	4	4	4	130	130	4	467	4	4	315	463	463	239	277	241	130	130	472	40
474	5	5	5	157	157	5	460	5	5	314	462	462	177	124	368	157	157	364	
	6	6	6	4	4	6	268	6	6	313	464	464	157	362	340	4	4	423	
		7	7	42	158	7	76	7	7	312	466	466	49	122	156	158		469	
		8	8	276	37	8	395	8	8	311	465	465	140	193	267	277		436	
		9	9	158	164	9	308	9	9	310	460	460	60	4	254	42		391	
		10	10	159	276	10	35	10	10	309	461	461	108	146	409	276		438	
		11	11	277	159	11	220	11	11	308	471		295	272	234	164		386	
		12	12	164	42	12	338	12	12	307	467		5	299	232	37		267	
		13	13	195	277	13	95	13	13	306	474		179	309	46	159		457	
		14	14	55	11	14	361	14	14	305			44	143	160	5		442	
		15		274	55	15	443	15	15	304			46	61	114	11		443	
		16		5	5	16	312	16	16	303			297	356	157	274		313	
		17		37	190	17	67	17	17	302			181	157	330	167		356	
		18		123	6	18	434	18	18	300			275	1	7	190		418	
		19		11	274	19	425	19	19	319			40	355	189	195		114	
		20		6	38	20	451	20	20	320			24	40	442	6		264	

Στην συνέχεια έγινε σύγκριση μεταξύ των έως 20 επιλεγμένων κοινών χαρακτηριστικών κάθε αλγορίθμου για την εύρεση κοινών χαρακτηριστικών μεταξύ των αλγορίθμων τα οποία τελικά αποτελούν και πιθανούς βιοδείκτες οι οποίοι ενδεχομένως να συνδέονται με τον αυτισμό.

Για τον σκοπό αυτό παρατίθεται ο Πίνακας 13 ο οποίος περιέχει την συχνότητα εμφάνισης κάθε χαρακτηριστικού. Η συχνότητα εντοπισμού ενός χαρακτηριστικού υποδηλώνει τον αριθμό των αλγορίθμων που το επέλεξαν. Αξίζει να αναφερθεί ότι τα χαρακτηριστικά που αναφέρονται είναι πιθανό να έχουν μεγάλη αξία καθώς αποτελούσαν μέρος των δεδομένων που δόθηκαν στον ταξινομητή ο οποίος παρουσίασε υψηλά αποτελέσματα (ακρίβεια 64-65%). Το γεγονός ότι επιλέχθηκαν από ποικίλους αλγορίθμους μείωσης διαστατικότητας παρουσιάζοντας θετικά αποτελέσματα ενισχύει την αξία που μπορεί να έχουν για την κατάταξη των ατόμων που ανήκουν στο φάσμα.

Πίνακας 13: Συχνότητα εμφάνισης κάθε χαρακτηριστικού (αναφέρονται οι κωδικοί αυτών όπως εμφανίζονται και στο παράρτημα) του παραπάνω πίνακα.

Στατική λειτουργική συνδεσιμότητα		Δυναμική λειτουργική συνδεσιμότητα			
		Μέση τιμή		Διακύμανση	
Κωδικός Χαρακτη-ριστικού	Πλήθος αλγορίθμων που εμφανίστηκαν	Κωδικός Χαρακτη-ριστικού	Πλήθος αλγορίθμων που εμφανίστηκαν	Κωδικός Χαρακτη-ριστικού	Πλήθος αλγορίθμων που εμφανίστηκαν
1	11	156	1	308	2
2	7	157	7	309	2
3	6	158	3	310	1
4	12	159	3	311	1
5	10	160	2	312	2
6	9	164	3	313	2
7	6	167	1	314	1
8	5	177	1	315	1
9	5	179	1	316	1
10	5	181	1	317	1
11	8	189	1	318	1
12	5	190	2	319	1
13	5	193	5	320	1
14	5	195	2	330	1
15	4	220	1	338	1
16	4	232	1	340	1
17	4	234	1	355	1
18	4	239	1	356	2
19	4	241	1	361	1
20	4	254	1	362	1
24	1	264	1	364	1
35	1	267	2	368	1
37	3	268	1	380	1
38	1	272	1	386	1
40	7	274	3	391	1

42	3	275	1	395	1
44	1	276	3	409	1
46	2	277	4	418	1
49	1	283	5	423	1
55	2	287	1	425	1
60	1	295	1	434	1
61	1	297	1	436	1
67	1	299	1	438	1
76	1	300	1	442	2
95	1	302	1	443	2
108	1	303	1	451	1
114	2	304	1	452	1
122	1	305	1	457	1
123	1	306	2	460	3
124	1	307	1	461	3
130	5			462	3
134	1			463	3
140	1			464	2
143	1			465	2
146	1			466	2
153	1			467	4
				468	2
				469	3
				470	2
				471	2
				472	1
				473	1
				474	2

Τέλος, όπως αναφέρθηκε παραπάνω και μέσω του παραρτήματος, είναι δυνατή η συσχέτιση των βιοδεικτών που επιλέχθηκαν, με συγκεκριμένες περιοχές του εγκεφάλου, από την συνδεσιμότητα των οποίων προήλθαν οι βιοδείκτες. Το παράρτημα περιέχει την αντιστοίχιση κάθε χαρακτηριστικού με τις περιοχές του εγκέφαλου που χρησιμοποιήθηκαν για τον υπολογισμό της λειτουργικής συνδεσιμότητας που το αποτελούν. Στον Πίνακας 14 παρατίθεται η συσχέτιση των κοινών βιοδεικτών που επιλέχθηκαν από τους αλγόριθμους επιλογής χαρακτηριστικών με συγκεκριμένες περιοχές του εγκεφάλου για τα κοινά χαρακτηριστικά με συχνότητα εμφάνισης μεγαλύτερη του 5.

Πίνακας 14: Συσχέτιση βιοδεικτών με συγκεκριμένες περιοχές του εγκεφάλου

Χαρακτηριστικά	Συνδεσιμότητα μεταξύ περιοχών
1	106 – 133
2	95 - 133
3	95 – 106
4	91 - 133
5	91 – 106

6	91 - 95
7	109 - 133
8	109 – 106
9	109 - 95
10	109 - 91
11	51 - 133
12	51 – 106
13	51- 95
14	51 - 91
40	122 - 91
130	174
157	91
193	122
283	174

Υπενθυμίζεται ότι οι περιοχές 95 και 133 βρίσκονται στην αριστερή μετωπική περιοχή του εγκεφάλου (Frontal_L), η περιοχή 106 βρίσκεται στην δεξιά μετωπική περιοχή του εγκεφάλου (Frontal_R) και οι περιοχές 51, 91 και 109 βρίσκονται στην μεσαία μετωπική περιοχή του εγκεφάλου (Frontal_L/R). Επίσης, η περιοχή 174 βρίσκεται στην δεξιά περιοχή του προσφηνοειδούς λοβίου (Precuneus_R) και η περιοχή 122 βρίσκεται στην αριστερή πάρα-ιπποκάμπεια περιοχή (Parahippo_L).

Από τα παραπάνω 19 χαρακτηριστικά τα οποία και παρουσίασαν σχετικά υψηλό ποσοστό επιλογής από τους αλγορίθμους επιλογής χαρακτηριστικών, μόνο τα χαρακτηριστικά 157, 193 και 283 έχουν προέλθει από την δυναμική λειτουργική συνδεσιμότητα. Τα τρία αυτά χαρακτηριστικά, προέκυψαν από στατιστικές στιγμές της μέσης τιμής. Αντίθετα, τα υπόλοιπα 16 προήλθαν από την στατική λειτουργική συνδεσιμότητα. Κανένα χαρακτηριστικό από την δυναμική λειτουργική συνδεσιμότητα δεν προέκυψε από τις στατιστικές στιγμές της διακύμανσης. Παρατηρείται λοιπόν ότι η πλειοψηφία των χαρακτηριστικών που επιλέχθηκαν από τουλάχιστον 5 αλγορίθμους προήλθαν με διαφορά από την στατική λειτουργική συνδεσιμότητα.

Ακόμα, παρατηρείται από τον Πίνακας 13 ότι το των πλήθος αλγορίθμων που επέλεξαν τα χαρακτηριστικά 157, 193 και 283 (τα οποία όλα έχουν προέλθει από την δυναμική λειτουργική συνδεσιμότητα χρησιμοποιώντας στατιστικές στιγμές της μέσης τιμής) ήταν αντίστοιχα 7, 5, 5. Τα παραπάνω νούμερα είναι σχετικά χαμηλά συγκριτικά με εκείνα από την στατική λειτουργική συνδεσιμότητα για τα οποία ο μέγιστος αριθμό αλγορίθμων που τα επέλεξαν ήταν 12. Συμπεραίνεται λοιπόν, ότι από τα παραπάνω χαρακτηριστικά, εκείνα που προήλθαν από την στατική λειτουργική συνδεσιμότητα επιλέχθηκαν από μεγαλύτερο ή ίσο πλήθος αλγορίθμων σε σχέση με αυτά που προήλθαν από την δυναμική λειτουργική συνδεσιμότητα χρησιμοποιώντας στατιστικές στιγμές της μέσης τιμής.

Τα χαρακτηριστικά που προήλθαν από την δυναμική λειτουργική συνδεσιμότητα χρησιμοποιώντας στατιστικές στιγμές της διακύμανσης εμφάνισαν τον χαμηλότερο κατά μέσο όρο αριθμό αλγορίθμων που τα επέλεξαν συγκριτικά με εκείνα που προήλθαν από την στατική λειτουργική συνδεσιμότητα

και την δυναμική λειτουργική συνδεσιμότητα χρησιμοποιώντας στατιστικές στιγμές της μέσης τιμής. Από την παρατήρηση αυτή προάγεται το συμπέρασμα ότι οι στατιστικές στιγμές της διακύμανσης της δυναμικής λειτουργικής συνδεσιμότητας δεν απέδωσαν.

Η φθίνουσα σειρά αποδοτικότητας χαρακτηριστικών ανά κατηγορία είναι εκείνα που προήλθαν από την στατική λειτουργική συνδεσιμότητα, την δυναμική λειτουργική συνδεσιμότητα χρησιμοποιώντας στατιστικές στιγμές της μέσης τιμής και τέλος της διακύμανσης.

6.5 Μελλοντικές ερευνητικές κατευθύνσεις

Είναι πιθανό ότι οι λειτουργικές συνδεσιμότητες που αναφέρθηκαν στην ενότητα αυτή, κρύβουν υψηλής σημασίας πληροφορίες που θα συντελούσαν στην δημιουργία ενός ακόμα πιο αποδοτικού μοντέλου ταξινόμησης ατόμων που ανήκουν στο φάσμα του αυτισμού.

Μια μελλοντική περεταίρω διερεύνηση, για την δημιουργία ενός ακόμα πιο αποδοτικού μοντέλου ταξινόμησης ατόμων που ανήκουν στο φάσμα του αυτισμού, θα μπορούσε να αποτελέσει η δοκιμή μοντέλων ταξινομητών με ή χωρίς την χρήση αλγορίθμων μείωσης διαστατικότητας που θα λάβουν ως είσοδο τα χαρακτηρίστηκα που απομονώθηκαν και για τα οποία διαπιστώθηκε πιθανή αξία τους ως βιοδείκτες, όπως αυτά εμφανίζονται στους Πίνακες Πίνακας 12 και Πίνακας 13.

7 Βιβλιογραφία

[1] J. C. Harris, “Chapter 6 - Autism Spectrum Disorder,” in *Neurobiology of Brain Disorders*, M. J. Zigmond, L. P. Rowland, and J. T. Coyle, Eds. San Diego: Academic Press, 2015, pp. 78–97.

[2] R. G. Floyd, I. L. Woods, L. J. Singh, and H. K. Hawkins, “Chapter 10 - Use of the Woodcock–Johnson IV Tests of Cognitive Abilities in the Diagnosis of Intellectual Disability,” in *WJ IV Clinical Use and Interpretation*, D. P. Flanagan and V. C. Alfonso, Eds. San Diego: Academic Press, 2016, pp. 271–289.

[3] E. Skafidas, R. Testa, D. Zantomio, G. Chana, I. P. Everall, and C. Pantelis, “Predicting the diagnosis of autism spectrum disorder using gene pathway analysis,” *Mol Psychiatry*, vol. 19, no. 4, pp. 504–510, Apr. 2014, doi: 10.1038/mp.2012.126.

[4] S. M. Smith *et al.*, “Correspondence of the brain’s functional architecture during activation and rest,” *PNAS*, vol. 106, no. 31, pp. 13040–13045, Aug. 2009, doi: 10.1073/pnas.0905267106.

[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, Jan. 2016, doi: 10.1016/j.procs.2016.04.224.

[6] J. T. Senders *et al.*, “Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review,” *World Neurosurgery*, vol. 109, pp. 476-486.e1, Jan. 2018, doi: 10.1016/j.wneu.2017.09.149.

[7] D. A. Hashimoto, G. Rosman, D. Rus, and O. R. Meireles, “Artificial Intelligence in Surgery: Promises and Perils,” *Ann Surg*, vol. 268, no. 1, pp. 70–76, Jul. 2018, doi: 10.1097/SLA.0000000000002693.

[8] S. M. Lundberg *et al.*, “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” *Nature Biomedical Engineering*, vol. 2, no. 10, Art. no. 10, Oct. 2018, doi: 10.1038/s41551-018-0304-0.

[9] D. Bone, S. L. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, “Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion,” *Journal of Child Psychology and Psychiatry*, vol. 57, no. 8, pp. 927–937, 2016, doi: 10.1111/jcpp.12559.

[10] “2-Channel convolutional 3D deep neural network (2CC3D) for fMRI analysis: ASD classification and feature learning - IEEE Conference Publication.” https://ieeexplore.ieee.org/document/8363798 (accessed Apr. 15, 2020).

[11] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, “Identifying Autism from Resting-State fMRI Using Long Short-Term Memory Networks,” *Mach Learn Med Imaging*, vol. 10541, pp. 362–370, Sep. 2017, doi: 10.1007/978-3-319-67389-9_42.

[12] C. Wang, Z. Xiao, and J. Wu, “Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data,” *Phys Med*, vol. 65, pp. 99–105, Sep. 2019, doi: 10.1016/j.ejmp.2019.08.010.

[13] A. Lenartowicz and R. A. Poldrack, “Brain Imaging,” in *Reference Module in Neuroscience and Biobehavioral Psychology*, Elsevier, 2017.

[14] J. R. Brasic, “Neurotransmitter visualization in schizophrenia,” *Journal of Biomedical Graphics and Computing*, vol. 3, no. 2, Art. no. 2, Feb. 2013, doi: 10.5430/jbgc.v3n2p30.

[15] J. R. Brašić and M. Mohamed, “Chapter Fifteen - Human Brain Imaging of Autism Spectrum Disorders,” in *Imaging of the Human Brain in Health and Disease*, P. Seeman and B. Madras, Eds. Boston: Academic Press, 2014, pp. 373–406.

[16] S. Assili, “A Review of Tomographic Reconstruction Techniques for Computed Tomography,” *arXiv:1808.09172 [physics]*, Sep. 2018, Accessed: Jul. 07, 2020. [Online]. Available: http://arxiv.org/abs/1808.09172.

[17] J. Naqvi, K. H. Yap, G. Ahmad, and J. Ghosh, “Transcranial Doppler Ultrasound: A Review of the Physical Principles and Major Applications in Critical Care,” *International Journal of Vascular Medicine*, vol. 2013, pp. 1–13, 2013, doi: 10.1155/2013/629378.

[18] M. R. Harrigan and J. P. Deveikis, *Handbook of cerebrovascular disease and neurointerventional technique*, Second edition. Dordecht: Humana Press, 2013.

[19] D. F. Wong and J. R. Brašić, “In vivo imaging of neurotransmitter systems in neuropsychiatry,” *Clinical Neuroscience Research*, vol. 1, no. 1, pp. 35–45, Jan. 2001, doi: 10.1016/S1566-2772(00)00005-0.

[20] “Full article: Brain imaging research: Does the science serve clinical practice?” https://www.tandfonline.com/doi/full/10.1080/09540260701564849?casa_token=bqDD

IKHmlu4AAAAA%3Al-di9qiqDh6XNJDgGgtOA03vgJlb-MonlWLLBfWs6U2M8O52rB94F3tWkiy05lXqjkW-w1nF55ff (accessed Apr. 22, 2020).
[21] “Magnetic Resonance Imaging (MRI).” https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri (accessed Apr. 23, 2020).
[22] J. B. Lambert, E. P. Mazzola, and C. Ridge, *Nuclear magnetic resonance spectroscopy: an introduction to principles, applications, and experimental methods*, Second edition. Hoboken, NJ: John Wiley & Sons, 2019.
[23] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic resonance imaging: physical principles and sequence design*, Second edition. Hoboken, New Jersey: John Wiley & Sons, Inc, 2014.
[24] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*, 2nd ed. Sunderland, Mass: Sinauer Associates, 2008.
[25] D. Attwell, A. M. Buchan, S. Charpak, M. Lauritzen, B. A. MacVicar, and E. A. Newman, “Glial and neuronal control of brain blood flow,” *Nature*, vol. 468, no. 7321, pp. 232–243, Nov. 2010, doi: 10.1038/nature09613.
[26] “What are the differences among EEG, MRI and fMRI?,” *BuscaEU*, Jan. 26, 2020. https://www.brainlatam.com/blog/what-are-the-differences-among-eeg-mri-and-fmri--1014 (accessed Apr. 23, 2020).
[27] A. Shmuel, “8 - On the relationship between functional MRI signals and neuronal activity,” in *Casting Light on the Dark Side of Brain Imaging*, A. Raz and R. T. Thibault, Eds. Academic Press, 2019, pp. 49–53.
[28] D. C. Noll, “A Primer on MRI and Functional MRI,” p. 14.
[29] K. J. Friston, “Functional and Effective Connectivity: A Review,” *Brain Connectivity*, vol. 1, no. 1, pp. 13–36, Jan. 2011, doi: 10.1089/brain.2011.0008.
[30] K. Noll, D. Sabsevitz, S. Prabhu, and J. Wefel, “Chapter 9 - Neuropsychology in the Neurosurgical Management of Primary Brain Tumors,” in *Neurosurgical Neuropsychology*, C. M. Pearson, E. Ecklund-Johnson, and S. D. Gale, Eds. Academic Press, 2019, pp. 157–183.
[31] A. Karampasi, *Εφαρμογή τεχνικών κυλιόμενου παραθύρου για ταξινόμηση διαταραχής αυτιστικού φάσματος*. National Technological University of Athens, 2019.
[32] “ABIDE Preprocessed.” http://preprocessed-connectomes-project.org/abide/ (accessed Feb. 03, 2020).
[33] N. U. F. Dosenbach *et al.*, “Prediction of Individual Brain Maturity Using fMRI,” *Science*, vol. 329, no. 5997, pp. 1358–1361, Sep. 2010, doi: 10.1126/science.1194144.
[34] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, “A whole brain fMRI atlas generated via spatially constrained spectral clustering,” *Hum. Brain Mapp.*, vol. 33, no. 8, pp. 1914–1928, Aug. 2012, doi: 10.1002/hbm.21333.
[35] “ABIDE.” http://fcon_1000.projects.nitrc.org/indi/abide/ (accessed Feb. 03, 2020).
[36] T. Xu, Z. Yang, L. Jiang, X.-X. Xing, and X.-N. Zuo, “A Connectome Computation System for discovery science of brain,” *Science Bulletin*, vol. 60, no. 1, pp. 86–95, Jan. 2015, doi: 10.1007/s11434-014-0698-3.
[37] S. Giavasis *et al.*, *Fcp-Indi/C-Pac: Cpac Version 1.0.0 Beta*. Zenodo, 2016.
[38] Y. Chao-Gan and Z. Yu-Feng, “DPARSF: A MATLAB Toolbox for ‘Pipeline’ Data Analysis of Resting-State fMRI,” *Front Syst Neurosci*, vol. 4, p. 13, 2010, doi: 10.3389/fnsys.2010.00013.
[39] C. Cameron *et al.*, “The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives,” *Front. Neuroinform.*, vol. 7, 2013, doi: 10.3389/conf.fninf.2013.09.00041.
[40] P. Bellec *et al.*, “A neuroimaging analysis kit for Matlab and Octave,” presented at the International Conference on Functional Mapping of the Human Brain, Quebec, QC, Canada, 2011.
[41] R. W. Cox, “AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages,” *Computers and Biomedical Research*, vol. 29, no. 3, pp. 162–173, Jun. 1996, doi: 10.1006/cbmr.1996.0014.
[42] N. J. Tustison *et al.*, “The ANTs cortical thickness processing pipeline,” Lake Buena Vista (Orlando Area), Florida, USA, Mar. 2013, p. 86720K, doi: 10.1117/12.2007128.
[43] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, “FSL,” *NeuroImage*, vol. 62, no. 2, pp. 782–790, Aug. 2012, doi: 10.1016/j.neuroimage.2011.09.015.

[44] “fMRI Tutorial #9: ROI Analysis — Andy’s Brain Book 1.0 documentation.” https://andysbrainbook.readthedocs.io/en/latest/fMRI_Short_Course/fMRI_09_ROIAnalysis.html (accessed Feb. 03, 2020).
[45] M. J. Hawrylycz *et al.*, “An anatomically comprehensive atlas of the adult human brain transcriptome,” *Nature*, vol. 489, no. 7416, pp. 391–399, Sep. 2012, doi: 10.1038/nature11405.
[46] V. Beliveau *et al.*, “A High-Resolution *In Vivo* Atlas of the Human Brain’s Serotonin System,” *J. Neurosci.*, vol. 37, no. 1, pp. 120–128, Jan. 2017, doi: 10.1523/JNEUROSCI.2830-16.2016.
[47] S. M. Smith *et al.*, “Resting-state fMRI in the Human Connectome Project,” *Neuroimage*, vol. 80, pp. 144–168, Oct. 2013, doi: 10.1016/j.neuroimage.2013.05.039.
[48] S. M. Czerniak *et al.*, “Areas of the Brain Modulated by Single-Dose Methylphenidate Treatment in Youth with ADHD During Task-Based fMRI: A Systematic Review,” *Harvard Review of Psychiatry*, vol. 21, no. 3, pp. 151–162, 2013, doi: 10.1097/HRP.0b013e318293749e.
[49] M. P. van den Heuvel and H. E. Hulshoff Pol, “Exploring the brain network: A review on resting-state fMRI functional connectivity,” *European Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, Aug. 2010, doi: 10.1016/j.euroneuro.2010.03.008.
[50] B. B. Biswal, J. Van Kylen, and J. S. Hyde, “Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps,” *NMR Biomed*, vol. 10, no. 4–5, pp. 165–170, Aug. 1997, doi: 10.1002/(sici)1099-1492(199706/08)10:4/5<165::aid-nbm454>3.0.co;2-7.
[51] “Resting State fMRI » fmri.at | neuroimaging & stimulation,” *fmri.at | neuroimaging & stimulation*. http://www.fmri.at/research/resting-state-fmri/ (accessed Apr. 26, 2020).
[52] M. C. Reddan and T. D. Wager, “Modeling Pain Using fMRI: From Regions to Biomarkers,” *Neurosci. Bull.*, vol. 34, no. 1, pp. 208–215, Feb. 2018, doi: 10.1007/s12264-017-0150-1.
[53] L. Glass, A. L. Ware, and S. N. Mattson, “Chapter 25 - Neurobehavioral, neurologic, and neuroimaging characteristics of fetal alcohol spectrum disorders,” in *Handbook of Clinical Neurology*, vol. 125, E. V. Sullivan and A. Pfefferbaum, Eds. Elsevier, 2014, pp. 435–462.
[54] M. E. Raichle, “The Brain’s Default Mode Network,” *Annu. Rev. Neurosci.*, vol. 38, no. 1, pp. 433–447, Jul. 2015, doi: 10.1146/annurev-neuro-071013-014030.
[55] A. E. Cavanna and M. R. Trimble, “The precuneus: a review of its functional anatomy and behavioural correlates,” *Brain*, vol. 129, no. Pt 3, pp. 564–583, Mar. 2006, doi: 10.1093/brain/awl004.
[56] L. V. Gabis, “Chapter 4 - Autism spectrum disorder: A clinical path to early diagnosis, evaluation, and intervention,” in *Neuroprotection in Autism, Schizophrenia and Alzheimer’s Disease*, I. Gozes and J. Levine, Eds. Academic Press, 2020, pp. 79–100.
[57] S. L. Hyman, S. E. Levy, S. M. Myers, and S. on D. and B. P. Council on Children with Disabilities, “Identification, Evaluation, and Management of Children With Autism Spectrum Disorder,” *Pediatrics*, vol. 145, no. 1, Jan. 2020, doi: 10.1542/peds.2019-3447.
[58] C. M. MD, “How early can you — and should you — diagnose autism?,” *Harvard Health Blog*, Aug. 23, 2019. https://www.health.harvard.edu/blog/how-early-can-you-and-should-you-diagnose-autism-2019082317653 (accessed Apr. 17, 2020).
[59] X. Li *et al.*, “2-Channel convolutional 3D deep neural network (2CC3D) for fMRI analysis: ASD classification and feature learning,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, Apr. 2018, pp. 1252–1255, doi: 10.1109/ISBI.2018.8363798.
[60] E. Alpaydin, *Introduction to Machine Learning*, Fourth edition. Cambridge, Massachusetts: The MIT Press, 2020.
[61] X.-Y. Zhou, Y. Guo, M. Shen, and G.-Z. Yang, “Artificial Intelligence in Surgery,” *arXiv:2001.00627 [physics]*, Dec. 2019, Accessed: Aug. 01, 2020. [Online]. Available: http://arxiv.org/abs/2001.00627.
[62] D. L. Robins, K. Casagrande, M. Barton, C.-M. A. Chen, T. Dumont-Mathieu, and D. Fein, “Validation of the Modified Checklist for Autism in Toddlers, Revised With Follow-up (M-CHAT-R/F),” *Pediatrics*, vol. 133, no. 1, pp. 37–45, Jan. 2014, doi: 10.1542/peds.2013-1813.

[63] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, “The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians,” *J Autism Dev Disord*, vol. 31, no. 1, pp. 5–17, Feb. 2001, doi: 10.1023/a:1005653411471.

[64] S. L. Barrett, M. Uljarević, E. K. Baker, A. L. Richdale, C. R. G. Jones, and S. R. Leekam, “The Adult Repetitive Behaviours Questionnaire-2 (RBQ-2A): A Self-Report Measure of Restricted and Repetitive Behaviours,” *J Autism Dev Disord*, vol. 45, no. 11, pp. 3680–3692, 2015, doi: 10.1007/s10803-015-2514-6.

[65] Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators and Centers for Disease Control and Prevention (CDC), “Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010,” *MMWR Surveill Summ*, vol. 63, no. 2, pp. 1–21, Mar. 2014.

[66] C. Lord *et al.*, “The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism,” *J Autism Dev Disord*, vol. 30, no. 3, pp. 205–223, Jun. 2000, doi: 10.1023/A:1005592401947.

[67] K. Strimbu and J. A. Tavel, “What are Biomarkers?,” *Curr Opin HIV AIDS*, vol. 5, no. 6, pp. 463–466, Nov. 2010, doi: 10.1097/COH.0b013e32833ed177.

[68] L. Shen *et al.*, “Advances in Biomarker Studies in Autism Spectrum Disorders,” in *Reviews on Biomarker Studies in Psychiatric and Neurodegenerative Disorders*, P. C. Guest, Ed. Cham: Springer International Publishing, 2019, pp. 207–233.

[69] P. Walsh, M. Elsabbagh, P. Bolton, and I. Singh, “In search of biomarkers for autism: scientific, social and ethical challenges,” *Nature Reviews Neuroscience*, vol. 12, no. 10, Art. no. 10, Oct. 2011, doi: 10.1038/nrn3113.

[70] K. E. Stephan and K. J. Friston, “Functional Connectivity,” in *Encyclopedia of Neuroscience*, L. R. Squire, Ed. Oxford: Academic Press, 2009, pp. 391–397.

[71] Y. Du, Z. Fu, and V. D. Calhoun, “Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging,” *Front. Neurosci.*, vol. 12, p. 525, Aug. 2018, doi: 10.3389/fnins.2018.00525.

[72] T. Yamada *et al.*, “Resting-State Functional Connectivity-Based Biomarkers and Functional MRI-Based Neurofeedback for Psychiatric Disorders: A Challenge for Developing Theranostic Biomarkers,” *Int J Neuropsychopharmacol*, vol. 20, no. 10, pp. 769–781, Jul. 2017, doi: 10.1093/ijnp/pyx059.

[73] Y. Li *et al.*, “Dynamic Functional Connectivity Reveals Abnormal Variability and Hyper-connected Pattern in Autism Spectrum Disorder,” *Autism Research*, vol. 13, no. 2, pp. 230–243, 2020, doi: 10.1002/aur.2212.

[74] R. A. Fisher, “Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population,” *Biometrika*, vol. 10, no. 4, p. 507, May 1915, doi: 10.2307/2331838.

[75] A. Savva, G. Mitsis, and G. Matsopoulos, “Assessment of dynamic functional connectivity in resting-state fMRI using the sliding window technique,” *Brain and Behavior*, Mar. 2019, doi: 10.1002/brb3.1255.

[76] L. Buitinck *et al.*, “API design for machine learning software: experiences from the scikit-learn project,” *arXiv:1309.0238 [cs]*, Sep. 2013, Accessed: Jul. 04, 2020. [Online]. Available: http://arxiv.org/abs/1309.0238.

[77] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, p. 2825−2830, Oct. 2011.

[78] K. Hsu, H. V. Gupta, and S. Sorooshian, “Artificial Neural Network Modeling of the Rainfall-Runoff Process,” *Water Resources Research*, vol. 31, no. 10, pp. 2517–2530, 1995, doi: 10.1029/95WR01955.

[79] “K1.pdf.” Accessed: Mar. 14, 2020. [Online]. Available: http://kelifos.physics.auth.gr/COURSES/neural/K1.pdf.

[80] S. S. Haykin and S. S. Haykin, *Neural networks and learning machines*, 3rd ed. New York: Prentice Hall, 2009.

[81] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, “Activation Functions: Comparison of trends in Practice and Research for Deep Learning,” *arXiv:1811.03378 [cs]*, Nov. 2018, Accessed: Mar. 15, 2020. [Online]. Available: http://arxiv.org/abs/1811.03378.

[82] “7 Types of Activation Functions in Neural Networks: How to Choose?,” *MissingLink.ai*. https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/ (accessed Mar. 15, 2020).
[83] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.
[84] “Explained: Neural networks,” *MIT News*. http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 (accessed Mar. 15, 2020).
[85] J. Nocedal, “Updating quasi-Newton matrices with limited storage,” *Math. Comp.*, vol. 35, no. 151, pp. 773–773, Sep. 1980, doi: 10.1090/S0025-5718-1980-0572855-7.
[86] S. Sra, S. Nowozin, and S. J. Wright, Eds., *Optimization for machine learning*. Cambridge, Mass: MIT Press, 2012.
[87] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, Accessed: Aug. 01, 2020. [Online]. Available: http://arxiv.org/abs/1412.6980.
[88] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[89] R. Berwick, “An Idiot’s guide to Support vector machines (SVMs),” p. 27.
[90] D. Freitag and A. K. Mccallum, “Information Extraction with HMMs and Shrinkage,” in *In Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999, pp. 31–36.
[91] M.-C. Wu, S.-Y. Lin, and C.-H. Lin, “An effective application of decision tree to stock trading,” *Expert Systems with Applications*, vol. 31, no. 2, pp. 270–274, Aug. 2006, doi: 10.1016/j.eswa.2005.09.026.
[92] Y. X. Gu, Q. R. Wang, and C. Y. Suen, “Application of a Multilayer Decision Tree in Computer Recognition of Chinese Characters,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 1, pp. 83–89, Jan. 1983, doi: 10.1109/TPAMI.1983.4767349.
[93] A. Elnaggar and J. Noller, “Application of Remote-sensing Data and Decision-Tree Analysis to Mapping Salt-Affected Soils over Large Areas,” *Remote Sensing*, vol. 2, no. 1, pp. 151–165, Dec. 2009, doi: 10.3390/rs2010151.
[94] X. E. Pantazi, D. Moshou, and D. Bochtis, “Chapter 2 - Artificial intelligence in agriculture,” in *Intelligent Data Mining and Fusion Systems in Agriculture*, X. E. Pantazi, D. Moshou, and D. Bochtis, Eds. Academic Press, 2020, pp. 17–101.
[95] N. B. Amor, S. Benferhat, and Z. Elouedi, “Qualitative Classification with Possibilistic Decision Trees,” in *Modern Information Processing*, B. Bouchon-Meunier, G. Coletti, and R. R. Yager, Eds. Amsterdam: Elsevier Science, 2006, pp. 159–169.
[96] R. Nisbet, J. Elder, and G. Miner, “Chapter 13 - Model Evaluation and Enhancement,” in *Handbook of Statistical Analysis and Data Mining Applications*, R. Nisbet, J. Elder, and G. Miner, Eds. Boston: Academic Press, 2009, pp. 285–312.
[97] R. E. Schapire, “Explaining AdaBoost,” in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, B. Schölkopf, Z. Luo, and V. Vovk, Eds. Berlin, Heidelberg: Springer, 2013, pp. 37–52.
[98] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[99] A. Meyer-Baese and V. Schmid, “Chapter 2 - Feature Selection and Extraction,” in *Pattern Recognition and Signal Analysis in Medical Imaging (Second Edition)*, A. Meyer-Baese and V. Schmid, Eds. Oxford: Academic Press, 2014, pp. 21–69.
[100] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, doi: 10.1093/bioinformatics/btm344.
[101] “Estimator (Statistics),” *DeepAI*, May 17, 2019. https://deepai.org/machine-learning-glossary-and-terms/estimator (accessed May 14, 2020).
[102] J. Li *et al.*, “Feature Selection: A Data Perspective,” *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Dec. 2017, doi: 10.1145/3136625.
[103] Y.-W. Chen and C.-J. Lin, “Combining SVMs with Various Feature Selection Strategies,” in *Feature Extraction: Foundations and Applications*, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Heidelberg: Springer, 2006, pp. 315–324.
[104] X. He, D. Cai, and P. Niyogi, “Laplacian Score for Feature Selection,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds. MIT Press, 2006, pp. 507–514.

[105] Z. Zhao and H. Liu, “Spectral feature selection for supervised and unsupervised learning,” in *Proceedings of the 24th international conference on Machine learning*, Corvalis, Oregon, USA, Jun. 2007, pp. 1151–1157, doi: 10.1145/1273496.1273641.
[106] M. Robnik-Šikonja and I. Kononenko, “Theoretical and Empirical Analysis of ReliefF and RReliefF,” *Machine Learning*, vol. 53, no. 1, pp. 23–69, Oct. 2003, doi: 10.1023/A:1025667309714.
[107] D. Lin and X. Tang, “Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion,” in *Computer Vision – ECCV 2006*, Berlin, Heidelberg, 2006, pp. 68–82, doi: 10.1007/11744023_6.
[108] D. Koller and M. Sahami, “Toward Optimal Feature Selection,” 1996.
[109] F. Fleuret, “Fast Binary Feature Selection with Conditional Mutual Information,” *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Dec. 2004.
[110] P. E. Meyer, C. Schretter, and G. Bontempi, *Information-Theoretic Feature Selection in Microarray Data using Variable Complementarity*. 2009.
[111] A. Adjimi, A. H. Gharbi, P. Ravier, and M. Mostefai, “Extraction and selection of binarised statistical image features for fingerprint recognition,” *IJBM*, vol. 9, no. 1, p. 67, 2017, doi: 10.1504/IJBM.2017.084133.
[112] D. S. Sisodia and A. Shukla, “Investigation of Feature Selection Techniques on Performance of Automatic Text Categorization,” in *Data, Engineering and Applications*, R. K. Shukla, J. Agrawal, S. Sharma, and G. Singh Tomer, Eds. Singapore: Springer Singapore, 2019, pp. 71–82.
[113] Hanchuan Peng, Fuhui Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.
[114] J. Zhou, D. Foster, R. Stine, and L. Ungar, “Streaming feature selection using alpha-investing,” in *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD ’05*, Chicago, Illinois, USA, 2005, p. 384, doi: 10.1145/1081870.1081914.
[115] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, “2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning,” p. 6.
[116] J. Tang, X. Hu, H. Gao, and H. Liu, “Discriminant Analysis for Unsupervised Feature Selection,” in *Proceedings of the 2014 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2014, pp. 938–946.
[117] D. Cai, C. Zhang, and X. He, “Unsupervised feature selection for multi-cluster data,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10*, Washington, DC, USA, 2010, p. 333, doi: 10.1145/1835804.1835848.
[118] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, “Unsupervised Feature Selection Using Nonnegative Spectral Analysis,” presented at the AAAI Conference on Artificial Intelligence, 2012.
[119] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan. 2002, doi: 10.1023/A:1012487302797.
[120] A. Samantaray and S. R. Dash, “Feature Selection Techniques to Predict the Religion of a Country from Its Flag,” in *Smart Intelligent Computing and Applications*, Singapore, 2020, pp. 191–201, doi: 10.1007/978-981-13-9282-5_18.
[121] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Phys. Rev. E*, vol. 69, no. 6, p. 066138, Jun. 2004, doi: 10.1103/PhysRevE.69.066138.
[122] B. C. Ross, “Mutual Information between Discrete and Continuous Data Sets,” *PLoS ONE*, vol. 9, no. 2, p. e87357, Feb. 2014, doi: 10.1371/journal.pone.0087357.
[123] N. Halko, P.-G. Martinsson, and J. A. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” *arXiv:0909.4061 [math]*, Dec. 2010, Accessed: Sep. 04, 2020. [Online]. Available: http://arxiv.org/abs/0909.4061.
[124] B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee, “The Insight ToolKit image registration framework,” *Front Neuroinform*, vol. 8, p. 44, 2014, doi: 10.3389/fninf.2014.00044.

[125] A. Zijdenbos, R. Forghani, and A. Evans, “Automatic ‘pipeline’ analysis of 3-D MRI data for clinical trials,” *IEEE transactions on medical imaging*, vol. 21, pp. 1280–91, Nov. 2002, doi: 10.1109/TMI.2002.806283.
[126] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, “Within-subject template estimation for unbiased longitudinal image analysis,” *Neuroimage*, vol. 61, no. 4, pp. 1402–1418, Jul. 2012, doi: 10.1016/j.neuroimage.2012.02.084.
[127] S. M. Smith *et al.*, “Advances in functional and structural MR image analysis and implementation as FSL,” *NeuroImage*, vol. 23, pp. S208–S219, Jan. 2004, doi: 10.1016/j.neuroimage.2004.07.051.
[128] S. M. Smith *et al.*, “Correspondence of the brain’s functional architecture during activation and rest,” *PNAS*, vol. 106, no. 31, pp. 13040–13045, Aug. 2009, doi: 10.1073/pnas.0905267106.
[129] B. Rashid *et al.*, “Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity,” *NeuroImage*, vol. 134, pp. 645–657, Jul. 2016, doi: 10.1016/j.neuroimage.2016.04.051.
[130] B. Christian and T. Griffiths, “Chapter 7: Overfitting,” in *Algorithms To Live By: The computer science of human decisions*, William Collins, 2017, pp. 149–168.
[131] Krishni, “An introduction to Grid search,” *Medium*, Jan. 05, 2019. https://medium.com/datadriveninvestor/an-introduction-to-grid-search-ff57adcc0998 (accessed Mar. 23, 2020).

8 Παράρτημα

Η λειτουργική συνδεσιμότητα (είτε στατική είτε δυναμική) μεταξύ δυο περιοχών του εγκέφαλου αντιστοιχεί σε ένα χαρακτηριστικό. Στους παρακάτω πίνακες παρουσιάζεται ένας χάρτης που προσδιορίζει από ποιες περιοχές του εγκέφαλου έχει προέλθει η λειτουργική συνδεσιμότητα που αντιστοιχεί σε κάθε χαρακτηριστικό. Ο Πίνακας 15 αφορά την στατική λειτουργική συνδεσιμότητα περιοχών του άτλαντα Craddock (CC-200). Ο Πίνακας 16 αφορά την δυναμική λειτουργική συνδεσιμότητά περιοχών του άτλαντα Craddock (CC-200), η οποία υπολογίστηκε ως προς τη μέση τιμή στον άξονα του χρόνου. Ο Πίνακας 17 αφορά την δυναμική λειτουργική συνδεσιμότητά περιοχών του άτλαντα Craddock (CC-200), η οποία υπολογίστηκε ως προς τη διακύμανση στον άξονα του χρόνου. Οι πίνακες εμφανίζονται στις επόμενες σελίδες.

Στους παρακάτω πίνακες σε κάθε ανοιχτού γκρι χρώματος κελί περιέχεται η περιοχή εγκέφαλου με την κωδική αριθμητική ονομασία της, σε κάθε σκούρου γκρι χρώματος κελί περιέχεται η κωδική λεκτική ονομασία της περιοχής του εγκέφαλου που περιέχεται στο ανοιχτού γκρι χρώματος δεξιό του κελί και σε κάθε λευκού χρώματος κελί περιέχεται η κωδική ονομασία χαρακτηριστικού. Δίνεται ένα παράδειγμα για την καλύτερη κατανόηση της διάταξης των στοιχείων του πίνακα: το χαρακτηριστικό με κωδικό 1 του άτλαντα Craddock (CC-200) εντοπίζεται στον Πίνακας 15 και προέρχεται από την στατική λειτουργική συνδεσιμότητα των περιοχών 106 και 133 δηλαδή των περιοχών Frontal_R και Frontal_L.

Πίνακας 15: Ο Πίνακας προσδιορίζει από ποιες περιοχές του άτλαντα CC-200 έχει προέλθει η στατική λειτουργική συνδεσιμότητα που αντιστοιχεί σε κάθε χαρακτηριστικό.

Static		Περιοχές του εγκεφάλου (άτλας CC-200)																	
			133	**106**	**95**	**91**	**109**	**51**	**22**	**5**	**62**	**122**	**170**	**166**	**97**	**136**	**163**	**197**	**174**	**58**
Περιοχές του εγκεφάλου (άτλας CC-200)	Frontal_L	**133**	0																
	Frontal_R	**106**	1	0															
	Frontal_L	**95**	2	3	0														
	Frontal_L/R	**91**	4	5	6	0													
	Frontal_L/R	**109**	7	8	9	10	0												
	Frontal_L/R	**51**	11	12	13	14	15	0											
	Frontal_R	**22**	16	17	18	19	20	21	0										
	Frontal_L	**5**	22	23	24	25	26	27	28	0									
	Parahippo_R	**62**	29	30	31	32	33	34	35	36	0								
	Parahippo_L	**122**	37	38	39	40	41	42	43	44	45	0							
	Occipital_R	**170**	46	47	48	49	50	51	52	53	54	55	0						
	Occipital_R	**166**	56	57	58	59	60	61	62	63	64	65	66	0					
	Occipital_L	**97**	67	68	69	70	71	72	73	74	75	76	77	78	0				
	Precuneus_L	**136**	79	80	81	82	83	84	85	86	87	88	89	90	91	0			
	Precuneus_R	**163**	92	93	94	95	96	97	98	99	100	101	102	103	104	105	0		
	Precuneus_L	**197**	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	0	
	Precuneus_R	**174**	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	0
	Paracingulate	**58**	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	0

Πίνακας 16: Ο Πίνακας προσδιορίζει από ποιες περιοχές του Άτλα CC-200 έχει προέλθει η δυναμική λειτουργική συνδεσιμότητα από την μέση τιμή, που αντιστοιχεί σε κάθε χαρακτηριστικό.

Mean dynamic functional connectivity		Περιοχές του εγκεφάλου (άτλας CC-200)																		
			133	**106**	**95**	**91**	**109**	**51**	**22**	**5**	**62**	**122**	**170**	**166**	**97**	**136**	**163**	**197**	**174**	**58**
Περιοχές του εγκεφάλου (άτλας CC-200)	Frontal_L	**133**	0																	
	Frontal_R	**106**	154	0																
	Frontal_L	**95**	155	156	0															
	Frontal_L/R	**91**	157	158	159	0														
	Frontal_L/R	**109**	160	161	162	163	0													
	Frontal_L/R	**51**	164	165	166	167	168	0												
	Frontal_R	**22**	169	170	171	172	173	174	0											
	Frontal_L	**5**	175	176	177	178	179	180	181	0										
	Parahippo_R	**62**	182	183	184	185	186	187	188	189	0									
	Parahippo_L	**122**	190	191	192	193	194	195	196	197	198	0								
	Occipital_R	**170**	199	200	201	202	203	204	205	206	207	208	0							
	Occipital_R	**166**	209	210	211	212	213	214	215	216	217	218	219	0						
	Occipital_L	**97**	220	221	222	223	224	225	226	227	228	229	230	231	0					
	Precuneus_L	**136**	232	233	234	235	236	237	238	239	240	241	242	243	244	0				
	Precuneus_R	**163**	245	246	247	248	249	250	251	252	253	254	255	256	257	258	0			
	Precuneus_L	**197**	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	0		
	Precuneus_R	**174**	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	0	
	Paracingulate	**58**	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	0

Πίνακας 17: Ο Πίνακας προσδιορίζει από ποιες περιοχές του άτλαντα CC-200 έχει προέλθει η δυναμική λειτουργική συνδεσιμότητα από την διακύμανση, που αντιστοιχεί σε κάθε χαρακτηριστικό.

Var dynamic functional connectivity		Περιοχές του εγκεφάλου (άτλας CC-200)																		
			133	**106**	**95**	**91**	**109**	**51**	**22**	**5**	**62**	**122**	**170**	**166**	**97**	**136**	**163**	**197**	**174**	**58**
Περιοχές του εγκεφάλου (άτλας CC-200)	Frontal_L	**133**	0																	
	Frontal_R	**106**	307	0																
	Frontal_L	**95**	308	309	0															
	Frontal_L/R	**91**	310	311	312	0														
	Frontal_L/R	**109**	313	314	315	316	0													
	Frontal_L/R	**51**	317	318	319	320	321	0												
	Frontal_R	**22**	322	323	324	325	326	327	0											
	Frontal_L	**5**	328	329	330	331	332	333	334	0										
	Parahippo_R	**62**	335	336	337	338	339	340	341	342	0									
	Parahippo_L	**122**	343	344	345	346	347	348	349	350	351	0								
	Occipital_R	**170**	352	353	354	355	356	357	358	359	360	361	0							
	Occipital_R	**166**	362	363	364	365	366	367	368	369	370	371	372	0						
	Occipital_L	**97**	373	374	375	376	377	378	379	380	381	382	383	384	0					
	Precuneus_L	**136**	385	386	387	388	389	390	391	392	393	394	395	396	397	0				
	Precuneus_R	**163**	398	399	400	401	402	403	404	405	406	407	408	409	410	411	0			
	Precuneus_L	**197**	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	0		
	Precuneus_R	**174**	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	0	
	Paracingulate	**58**	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	0