Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
ΔΠΜΣ Επιστημη Δεδομενων και Μηχανικη Μαθηση

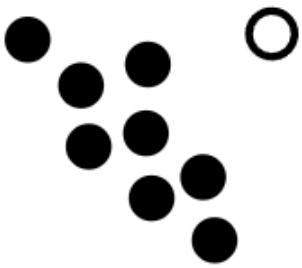
Εύρεση Ανωμαλιών

Μελέτη και Σύγκριση Μεθόδων

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΗ ΠΑΠΑΜΑΥΡΟΥ



Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Παρασκευή Τζούβελη
ΕΔΙΠ Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
ΔΠΜΣ Επιστημη Δεδομενων και Μηχανικη Μαθηση

Εύρεση Ανωμαλιών

Μελέτη και Σύγκριση Μεθόδων

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΗ ΠΑΠΑΜΑΥΡΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Παρασκευή Τζούβελη
ΕΔΙΠ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 5η Οκτωβρίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας	Αθανάσιος Βουλόδημος	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
ΔΠΜΣ Επιστημη Δεδομενων και Μηχανικη Μαθηση

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Δημήτρης Παπαμαύρος

5 Οκτωβρίου 2022

Περίληψη

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, μελετήθηκε η συμπεριφορά μιας πληθώρας τεχνικών, τόσο κλασσικής όσο και βαθιάς μηχανικής μάθησης, με σκοπό την Εύρεση Ανωμαλιών εστιάζοντας στην ερευνητική περιοχή της Ανίχνευσης Εισβολής σε Δεδομένα Δικτύου. Για τον σκοπό αυτό, εκπαιδεύτηκαν Μοντέλα μιας Κλάσης αλλά και Επιβλεπόμενης και Ημι-Επιβλεπόμενης μηχανικής μάθησης χρησιμοποιώντας τη συλλογή δεδομένων CSE-CIC-IDS2018. Σε αυτά τα πλαίσια τα μοντέλα επιβλεπόμενης μάθησης κατάφεραν να επιτύχουν βέλτιστα αποτελέσματα με το επιλεγέν μοντέλο να παρουσιάζει στο σύνολο ελέγχου Ακρίβεια, Average Precision και ROC-AUC ίσα με 98.65%, 98.44% και 99.28% αντίστοιχα.

Πρώτο βήμα αποτέλεσε η εκπαίδευση Μοντέλων μιας Κλάσης βασισμένα αποκλειστικά στις φυσιολογικές παρατηρήσεις με σκοπό την ταυτοποίηση επιθέσεων. Σε αυτό το σημείο ιδιαίτερη σημασία δόθηκε στην εκπαίδευση διαφόρων αρχιτεκτονικών για Αυτοκωδικοποιητές (Undercomplete AE, Stacked AE, Sparse AE, Denoising AE). Η εκπαίδευση των Αυτοκωδικοποιητών έλαβε χώρα με σκοπό την εκμάθηση χρήσιμων απεικονίσεων στο λανθάνων χώρο για τη κλάση των φυσιολογικών παρατηρήσεων. Ως αποτέλεσμα κατέστη εφικτή η αναγνώριση ανωμαλιών αξιοποιώντας το υψηλό παρατηρηθέν σφάλμα ανακατασκευής.

Στη συνέχεια, χρησιμοποιήθηκαν Μοντέλα Ενισχυτικής Κλίσης (XGBoost, LightGBM, CatBoost) συνδυάζοντάς Μάθηση με Ευαισθησία στο Κόστος και Τεχνικές Επαύξησης Δεδομένων στα πλαίσια της Μάθησης με Μη Ισορροπημένα Δεδομένα. Συγκεκριμένα, χρησιμοποιήθηκε η τεχνική SMOTE καθώς και η χρήση ενός Παραγωγικού Ανταγωνιστικού Δικτύου (CTGAN) με σκοπό τη παραγωγή συνθετικών παρατηρήσεων. Επιπλέον, στα πλαίσια των πειραμάτων, επιλέχθηκε η χρήση ενός Αυτοκωδικοποιητή με σκοπό τον εμπλουτισμό του συνόλου δεδομένων με το διάνυσμα του σφάλματος ανακατασκευής. Για το καθορισμό των βέλτιστων υπερπαραμέτρων των μοντέλων Επιβλεπόμενης και Ημι-Επιβλεπόμενης Μάθησης επιλέχθηκε η χρήση Μπεϋζιανής Βελτιστοποίησης.

Τέλος, το επιλεγέν μοντέλο αξιολογήθηκε και στο σύνολο δεδομένων CIC-IDS2017. Η συγκεκριμένη συλλογή αποτελείται από παρόμοιου τύπου επιθέσεις με σημαντικά διαφορετική κατανομή. Με αυτόν το τρόπο καταφέραμε να εξάγουμε χρήσιμα συμπεράσματα σχετικά με προβλήματα μετατόπισης θεματικής (concept drift).

Λέξεις Κλειδιά

Εύρεση Ανωμαλιών, Ανίχνευση Εισβολής σε Δεδομένα Δικτύου, Μοντέλα μιας Κλάσης, Αυτοκωδικοποιητές, Μοντέλα Ενισχυτικής Κλίσης, Μάθηση με Μη Ισορροπημένα Δεδομένα, Μπεϋζιανή Βελτιστοποίηση

Abstract

In the context of this thesis, the behavior of a multitude of, both classical and deep machine learning techniques was studied with the purpose of Anomaly Detection, focusing on the research area of Network Intrusion Detection. To this end, One Class as well as Supervised and Semi-Supervised machine learning models were trained using the CSE-CIC-IDS2018 dataset. In these contexts the supervised learning models managed to achieve optimal results with the selected model presenting, in the Test Set, Accuracy, Average Precision and ROC-AUC equal to 98.65%, 98.44% and 99.28% respectively.

The first step was training of One-Class Models based exclusively on normal observations in order to identify attacks. At this point, our effort was focuced on training various architectures for Autoencoders (Undercomplete AE, Stacked AE, Sparse AE, Denoising AE). The training of the Autoencoders took place in a way that would assit in learning useful representations in latent space for the normal class. As a result, anomalies were identified through exploiting the high observed reconstruction error.

Following, Gradient Boosting Models (XGBoost, LightGBM, CatBoost) were used in combination with Cost-Sensitive Learning and Data Augmentation Techniques in the framework of Learning with Imbalanced Data. In more detail, the SMOTE method as well as a Generative Adversarial Network (CTGAN) were used in order to produce synthetic observations. Furthermore, in the context of the experiments, the use of an Autoencoder was chosen in order to enrich the dataset with the reconstruction error vector. To determine the optimal hyperparameters of the Supervised and Semi-Supervised Learning models, the use of Bayesian Optimization was chosen.

Finally, the selected model was also evaluated on the CIC-IDS2017 dataset. This particular collection consists of similar types of attacks with significantly different distributions. In this way we were able to draw useful conclusions regarding problems related to concept drifts.

Keywords

Anomaly Detection, Network Intrusion Detection, One Class Models, Autoencoders, Gradient Boosting Models, Imbalanced Learning, Bayesian Optimization

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Στέφανο Κόλλια για την ευκαιρία που έδωσε να ασχοληθώ με ένα επίκαιρο θέμα.
Επιπλέον, θα ήθελα να ευχαριστήσω τα μέλη της επιτροπής τον Αν. Καθηγητή κ. Γεώργιο Στάμου και τον Επίκουρο Καθηγητή κ. Αθανάσιο Βουλόδημο για την τιμή που μου έκαναν να συμμετάσχουν στην επιτροπή εξέτασης της εργασίας.
Ιδιαίτερα όμως, θα ήθελα να ευχαριστήσω και την Διδάκτορα και επιβλέπουσα μου κα. Παρασκευή Τζούβελη. Χωρίς την στήριξη και τις συμβουλές της η ολοκλήρωση της παρούσας διπλωματικής εργασίας δεν θα ήταν εφικτή.
Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου που ήταν δίπλα μου σε όλη αυτή τη πορεία. Η στήριξή τους ήταν πάντα αμέριστη.

Αθήνα, Οκτώβριος 2020

Δημήτρης Παπαμαύρος

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Μέρος I

Εισαγωγή

Κεφάλαιο 1

Ορισμός Προβλήματος

Το πρόβλημα της Ανίχνευσης Ανωμαλιών αποτελεί ένα ενεργό πεδίο έρευνας το οποίο θα μπορούσε να χαρακτηριστεί ως ένα πρόβλημα ταξινόμησης το οποίο χαρακτηρίζεται από σημαντική μη ισορροπία ανάμεσα στις δύο κλάσης ενδιαφέροντος (Φυσιολογικές Παρατηρήσεις και Ανωμαλίες). Οι ανωμαλίες αποτελούν, στα πλαίσια διαφόρων εφαρμογών, σημεία υψίστης σημασίας καθώς κρύβουν πολλές φορές σημαντική και χρήσιμη πληροφορία.

Σε μια εποχή όπου η καθημερινότητα πολλών συνανθρώπων μας είναι συμβιβασμένη με τη καθημερινή αλληλεπίδραση με προσφερόμενες υπηρεσίες στο Διαδικτύου, η βελτίωση των επιπέδων Κυβερνοασφάλειας κρίνεται αναγκαία. Τα τελευταία χρόνια οι κυβερνοεπιθέσεις έχουν αναχθεί σε ένα από τα σημαντικότερα ζητήματα ασφάλειας καθώς με τη μεταφορά σημαντικών υπηρεσιών, ακόμη και κρατικών, στο διαδίκτυο οι επιτιθέμενοι είναι στη θέση να υποκλέψουν ευαίσθητες πληροφορίες καθώς και να αποκλείσουν χρήστες από τη πρόσβαση σε σημαντικές υπηρεσίες. Η αποτροπή αυτών των επιθέσεων ορίζει το πρόβλημα της Ανίχνευσης Εισβολής σε Δεδομένα Δικτύου το οποίο με τη σειρά του αποτελεί υποπεριοχή της ευρύτερης περιοχής της Ανίχνευσης Ανωμαλιών και σε γενικότερες γραμμές έχει απασχολήσει τους ερευνητές στα πλαίσια της ανάπτυξης συστημάτων Κυβερνοασφάλειας.

Στα πλαίσια της Ανίχνευσης Εισβολής οι ερευνητές καλούνται να δημιουργήσουν μοντέλα μηχανικής μάθησης ικανά να εντοπίσουν έγκαιρα τους επίδοξους επιτιθέμενους τόσο σε μεμονωμένους υπολογιστές όσο και σε δίκτυα υπολογιστών. Η δημιουργία μοντέλων ικανών να εντοπίσουν κινήσεις δικτύου οι οποίες αποκλίνουν από αντίστοιχες μη επιθετικές κινήσεις δικτύου αποτελεί ένα ενεργό πεδίο. Ειδικά η δημιουργία μοντέλων ικανά να εντοπίσουν πληθώρα διαφορετικών επιθέσεων, χωρίς να εστιάζουν σε συγκεκριμένα ήδη επιθέσεων, θα βελτιώσει σημαντικά τις αποδόσεις των εκάστοτε συστημάτων στα πλαίσια της εφαρμογής τους σε πραγματικά συστήματα καθώς η δημιουργία ενός γενικού μοντέλου εύρεσης ανωμαλιών θα βοηθήσει ως προς τη προστασία έναντι σε κάθε είδους πιθανόν επιθέσεων. Παρόλο αυτά το παρόν δεν αποτελεί ένα εύκολο προς επίλυση πρόβλημα. Κύριο λόγο για αυτό αποτελεί η έλλειψη δεδομένων ικανά να περιγράψουν με επιτυχία μοτίβα επιθέσεις. Η έλλειψη δεδομένων οφείλεται στη ευαισθησία του συγκεκριμένου είδους δεδομένων, γεγονός το οποίο αποτρέπει το δημόσιο διαμοιρασμό τους, αλλά και την ικανότητα των επιτιθεμένων να αναπροσαρμόζουν ανά τακτά χρονικά διαστήματα τα μοτίβα επίθεσης με αποτέλεσμα τη δυσχέρανση της κατασκευής αντίστοιχων συστημάτων.

Κεφάλαιο 2

Δομή Διπλωματικής Εργασίας

Στο 1ο Μέρος παρουσιάζεται συνοπτικά το έναυσμα για την εκπόνηση της διπλωματικής καθώς και ονομαστικά οι τεχνικές που χρησιμοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας και τέλος η δομή της εργασίας.

Στο 2ο Μέρος παρουσιάζονται αναλυτικά όλες οι μέθοδοι που χρησιμοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας με έμφαση στο θεωρητικό υπόβαθρο και τρόπο λειτουργίας τους.

Στο 3ο Μέρος παραθέτουμε τόσο πληροφορίες σχετικά με τα πειραματικά δεδομένα, την μεθοδολογία που χρησιμοποιήθηκε για την υλοποίηση των διάφορων μοντέλων και αρχιτεκτονικών καθώς και τα τελικά ευρήματα.

Στο 4ο Μέρος πραγματοποιείται μια σύνοψη της διαδικασίας που ακολουθήθηκε και παρουσιάζονται μελλοντικές προτάσεις.

Κεφάλαιο 3

Συγγενείς Εργασίες

Το πρόβλημα της Ανίχνευσης Ανωμαλιών αποτελεί ένα ενεργό πεδίο έρευνας με αποτέλεσμα η διεθνής βιβλιογραφία να περιέχει πληθώρα σχετικών άρθρων. Στα πλαίσια του κεφαλαίου αυτού θα αναφερθούμε σε ορισμένα από τα πιο σχετικά στο πλαίσιο της παρούσας διπλωματικής.

Αρχικά το 2020 ο Hua συνδύασε τη χρήση του μοντέλου LightGBM με υποδειγματοληψία της πλειοψηφικής κλάσης καθώς και ενσωματωμένων τεχνικών επιλογής χαρακτηριστικών με σκοπό την ανίχνευσης ανωμαλιών [1]. Συγκρίνοντας την απόδοση του μοντέλου με άλλα πέντε απλούστερα μοντέλα καθώς και ένα Συνελικτικό Δίκτυο ο Hua συμπέρανε ότι η χρήση του LightGBM υπέρισχυε τον υπολοίπων. Πιο συγκεκριμένα, για την επιλογή των χαρακτηριστικών επιλέχθηκε η χρήση του αλγορίθμου XGBoost ενώ τα υπόλοιπα υπό εξέταση μοντέλα αποτελούνταν από Support Vector Machine, Random Forest (RF), AdaBoost, Multilayer Perceptron (MLP) και Naive Bayes. Χρησιμοποιώντας τα δέκα σημαντικότερα χαρακτηριστικά κατάφερε να επιτύχει ακρίβεια, ευστοχία και ανάκληση ίση με 98.37%, 98.14% και 98.37% αντίστοιχα.

Στα πλαίσια της δημοσίευσης [2] οι Fitni και Ramli έχοντας πραγματοποιήσει επιλογή χαρακτηριστικών, βασισμένη στο Spearman's Rank Correlation και τον X^2-Έλεγχο, επέλεξαν τη χρήση ενός ταξινομητή συνόλου (Ensemble Classifier) αποτελούμενο από τα τρία μοντέλα (Gradient Boosting Machine, Logistic Regression, Decision Tree). Μέσω αυτής της μεθόδου κατέληξαν σε αποτελέσματα της τάξης των 98.80%, 98.80% και 97.10% για την ακρίβεια, ευστοχία και ανάκληση αντίστοιχα.

Το 2020 οι Ferrag et al. [3] πραγματοποίησαν μια εξονυχιστική μελέτη σε σχέση με την επίδοση τεχνικών βαθιάς μάθησης. Συγκεκριμένα χρησιμοποιήθηκαν Βαθιά Νευρωνικά Δίκτυα, RNNs, Συνελικτικά Νευρωνικά Δίκτυα (CNNs), βαθιές και περιορισμένες μηχανές Boltzmann καθώς και Βαθιά Δίκτυα Αυτοκωδικοποίησης. Η υψηλότερη ακρίβεια (97.38%) επιτεύχθηκε μέσω της χρήση ενός RNN ενώ η υψηλότερη ανάκληση (98.18%) επιτεύχθηκε μέσω χρήσης ενός CNN. Αντίθετα στο [4] οι Basnet et al. ανέφεραν επιδόσεις, ως προς την ακρίβεια, της τάξης του 99% για την ανίχνευση ανωμαλιών μέσω της χρήσης MLP.

Για τον καθορισμό βέλτιστων υπερπαραμέτρων οι συγγραφείς του [5] χρησιμοποιούν GridSearchCV για ένα Multilayer Perceptron με δύο κρυφά επίπεδα. Στα πλαίσια της μελέτης οι ερευνητές εστιάζουν αποκλειστικά στην εύρεση επιθέσεων botnet κάτι το οποίο επιτυγχάνεται με τις μετρικές ακρίβειας, ανάκλησης και ευστοχίας να επιτυγχάνουν τιμές της τάξης του 100%.

Η χρήση της τεχνικής SMOTE με σκοπό την ενίσχυση του συνόλου δεδομένων χρησιμοποιήθηκε από τους Karatas et al. [6] με σκοπό την αντιμετώπιση προβλημάτων ανισορροπίας των κλάσεων στα πλαίσια της δημιουργίας μοντέλων με σκοπό την ανίχνευση εισβολής. Στο συγκεκριμένο άρθρο μελετήθηκε η επίδραση της μεθόδου σε μοντέλα μηχανικής μάθησης (k-NN, RF, Gradient Boosting, AdaBoost, Decision Tree, Linear Discriminant Analysis) με το AdaBoost να παρουσιάζει, μέσω της χρήσης SMOTE, ακρίβεια ίση με 99.69%, καθώς και τιμές ευστοχίας και ανάκλησης ίσες με 99.70% και 99.69% αντίστοιχα.

Αξίζει να σημειωθεί ότι όλες οι προαναφερθείσες μελέτες αφορούν το σύνολο δεδομένων CSE-CIC-IDS2018 και χαρακτηρίζονται από διαφορετικές μεθοδολογίες προ-επεξεργασίας αυτού. Ερευνώντας τη δυνατότητα γενίκευσης διαφόρων μοντέλων μηχανικής μάθησης σε διαφορετικά σύνολα δεδομένων οι D'hooge et al. το 2020 χρησιμοποίησαν από κοινού τα σύνολα δεδομένων CSE-CIC-IDS2018 και CIC-IDS2017 με σκοπό την εκπαίδευση επιβλεπόμενων μοντέλων μηχανικής μάθησης. Στα πλαίσια αυτής της έρευνας οι συγγραφείς χρησιμοποίησαν 12 μοντέλα κλασσικής μηχανικής μάθησης, αποφεύγοντας τη χρήση βαθιάς μάθησης, με το XGBoost να πετυχαίνει τα καλύτερα αποτελέσματα (για το CSE-CIC-IDS2018 ακρίβεια, ευστοχία, ανάκληση ίσα με 96%, 99% και 79% αντίστοιχα) ενώ αξίζει να σημειωθεί ότι οι συγγραφείς κατέληξαν στο συμπέρασμα της μη εφικτής γενίκευσης για μοντέλα εκπαιδευμένα αποκλειστικά στο ένα εκ των δύο συνόλων δεδομένων.

Στα πλαίσια της εργασίας και ανάπτυξης των προτεινόμενων μοντέλων μελετήθηκαν οι παρακάτω εργασίες οι οποίες εστιάζουν στην ανάπτυξη μοντέλων ικανά να ανιχνεύσουν ανωμαλίες στα πλαίσια της Ανίχνευσης Εισβολής. Συγκεκριμένα, μελετήθηκαν δημοσιεύσεις σχετικά με την ανάπτυξη μοντέλων βασισμένα σε Ensemble αρχιτεκτονικές [7],[8] καθώς και τη χρήση τεχνικών εξελικτικής αναζήτησης (Genetic Algorithms, Particle Swarm Optimization, Colony Optimization) στα πλαίσια της επιλογής χαρακτηριστικών και την εκ των υστέρων δημιουργία Ensemble αρχιτεκτονικών [9]. Επιπροσθέτως, ερευνήθηκαν μελέτες σχετικές με τη χρήση υβριδικών μεθόδων βασισμένες στο συνδυασμό μη επιβλεπόμενων και επιβλεπόμενων τεχνικών μάθησης ερευνήθηκαν καθώς και η δημιουργία πολυεπίπεδων ταξινομητών [10], [11], [12], τη χρήση τεχνικών επαύξησης δεδομένων στα πλαίσια του προβλήματος [13] καθώς και τη χρήση Αυτοκωδικοποιήτων με σκοπό την εξαγωγή χαρακτηριστικών και τη περαιτέρω δημιουργία ταξινομητών τόσο στα πλαίσια επιβλεπόμενης μάθησης αλλά και υβριδικών μεθόδων ημι-επιβλεπόμενης μάθησης [14], [15], [16], [17]. Τέλος, μελετήθηκαν μεθοδολογίες βασισμένες σε βαθιά μάθηση όπως η χρήση δικτύων RNN, δικτύων CNN με σκοπό τον εμπλουτισμό των δεδομένων και την εκμάθηση σημασιολογικών απεικονίσεων αλλά και η εφαρμογή BLSTM δικτύων σε συνδυασμό με τεχνικές προσοχής (attention) [18], [19], [20], [21], [22].

Τέλος, αρχιτεκτονικές βαθιών νευρωνικών δικτύων έχουν υλοποιηθεί και χρησιμοποιηθεί σε διάφορες εφαρμογές από μέλη του Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης του ΕΜΠ. Ειδικότερα τεχνικές CNN και CNN-RNN έχουν εφαρμοστεί για ιατρική διάγνωση νευροεκφυλιστικών ασθενειών, όπως της νόσου του Πάρκινσον [23], [24], [25], [26], [27] ή της Covid-19 [28], [29], [30], βασισμένες σε 2-D ή 3-D εικόνες. Έμφαση έχει δοθεί στη διαφάνεια και στη προσαρμογή των μοντέλων [31], [32], [33] αλλά και στην ανάπτυξη πλέον σύνθετων αρχιτεκτονικών, μπεϋεσιανών, με κάψουλες και αβεβαιότητα [34], [35], [36], [37]. Βαθιές 3-D νευρωνικές αρχιτεκτονικές έχουν εφαρμοστεί στην ανίχνευση βλαβών σε

πυρηνικούς αντιδραστήρες [38], [39], στην πρόβλεψη της παραγωγής στον αγροτικό τομέα [40], [41] και στην αναγνώριση και σύνθεση συναισθήματος [42], [43], [44], [45], ενώ άλλες εφαρμόζονται σε προβλήματα ανάλυσης εικόνων και αλληλεπίδρασης ανθρώπου-υπολογιστή [46], [47], [48].

Μέρος II

Θεωρητικό Υπόβαθρο

Κεφάλαιο 4

Εύρεση Ανωμαλιών

4.1 Ανάλυση Ακραίων Σημείων και Ανωμαλιών

Ως ακραίο σημείο ή ανωμαλία ορίζουμε μια παρατήρηση η οποία αποκλίνει σημαντικά από τις ¨φυσιολογικές¨ παρατηρήσεις, η δημιουργία της δηλαδή έχει προέλθει από διαφορετικό μηχανισμό σε σχέση με τα ¨φυσιολογικά¨ δεδομένα [49]. Τα ακραία σημεία διαφέρουν από τα δεδομένα θορύβου καθώς ο θόρυβος αποτελεί τυχαίο σφάλμα ή διακύμανση σε κάποια μετρήσιμη μεταβλητή και πρέπει να αφαιρεθεί προτού ξεκινήσει η διαδικασία ανίχνευσης ακραίων τιμών.

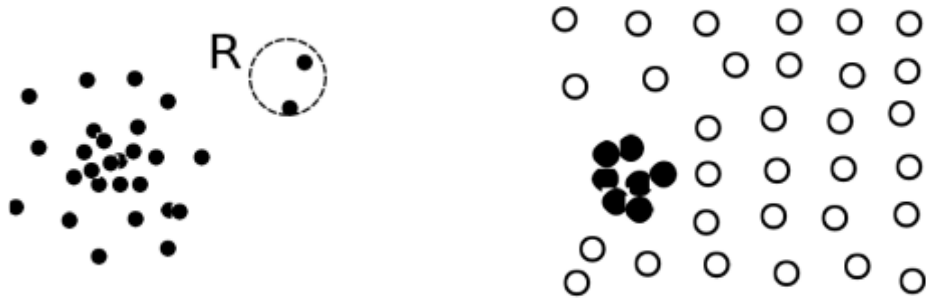
Τα ακραία σημεία διακρίνονται σε τρεις κατηγορίες με βάση τα διακριτά χαρακτηριστικά τους [50].

- **Ακραίο Σημείο Υπό Όρους (Contextual Outlier)**
  Ακραία σημεία υπό όρους θεωρούνται παρατηρήσεις οι οποίες αποκλίνουν σημαντικά δεδομένου ενός συγκεκριμένου πλαισίου ή ενός επιλεγμένου περιβάλλοντος. Για παράδειγμα μια μέτρηση θερμοκρασίας της τάξεως των $20^{\circ}C$ μπορεί να θεωρείται φυσιολογική εκτός εάν έχει καταγραφεί στην Ανταρκτική.

- **Ανωμαλία Σημείου (Global Outlier/Point Anomaly)**
  Ως Ανωμαλία Σημείου θεωρούνται παρατηρήσεις οι οποίες αποκλίνουν σημαντικά από το υπόλοιπο σύνολο δεδομένων (Σχήμα 4.1α΄). Οι συγκεκριμένες αποτελούν το πιο κοινό είδος ακραίων σημείων με τους περισσότερους αλγορίθμους να στοχεύουν στην ανίχνευση τέτοιου είδους ακραίων σημείων. Για παράδειγμα, σε προβλήματα ανίχνευσης εισβολής σε δίκτυα υπολογιστών μπορούμε να θεωρήσουμε ως ανωμαλίες σημείου επικοινωνιακές συμπεριφορές ενός υπολογιστή οι οποίες διαφέρουν σημαντικά από τα κανονικά μοτίβα.

- **Συλλογικά Ακραία Σημεία (Collective Outlier)**
  Ως συλλογικά ακραία σημεία ορίζουμε μια συλλογή από παρατηρήσεις η οποία αποκλίνει από το υπόλοιπο σύνολο δεδομένων ακόμα και εάν οι παρατηρήσεις από μόνες τους δεν αποτελούν ακραίες τιμές. Στο Σχήμα 4.1β΄ παρατηρούμε ότι οι μαύρες παρατηρήσεις σχηματίζουν ένα συλλογικό ακραίο σημείο, παρόλο που τα ξεχωριστά σημεία δεν αποτελούν ακραίες τιμές, καθώς η πυκνότητα τους είναι σημαντικά υψηλότερη σε σχέση με το υπόλοιπο σύνολο δεδομένων. Στα πλαίσια του παραδείγματος για την ανίχνευση εισβολής σε δίκτυα υπολογιστών, μπορούμε να θεωρήσουμε ως συλλογικά

ακραία σημεία ένα σύνολο υπολογιστών το οποίο στέλνει συνεχώς πακέτα άρνησης υπηρεσίας στο υπόλοιπο δίκτυο.



(α') *Ανωμαλία Σημείου* (β') *Συλλογικά Ακραία Σημεία*

Σχήμα 4.1: *Είδη Ακραίων Σημείων*

Φυσικά, ένα σύνολο δεδομένων δύναται να αποτελείται από μια πληθώρα ειδών ακραίων σημείων, ενώ μια παρατήρηση είναι πιθανόν να ανήκει σε περισσότερα από έναν τύπους ακραίων τιμών. Συνεπώς, κρίνεται υψίστης σημασίας, ο καθορισμός των μεθόδων για την ανίχνευση των ακραίων τιμών.

Οι μέθοδοι για εύρεση ακραίων τιμών διαχωρίζονται με βάση το εάν υπάρχουν διαθέσιμα δεδομένα με ετικέτες με σκοπό τη ταυτοποίηση μια τιμής ως ακραία ή όχι [51].

Στη περίπτωση που δεν υπάρχουν ετικέτες, οι μέθοδοι αποτελούνται, κατά κύριο λόγο, από μη εποπτευόμενες μεθόδους, οι οποίες προβαίνουν σε παραδοχές για την κατανομή τόσο των φυσιολογικών όσο και των ακραίων τιμών. Τέτοιες μέθοδοι αποτελούν στατιστικές προσεγγίσεις (παραμετρικές και μη παραμετρικές) όπως η μοντελοποίηση του συνόλου δεδομένων από μείγματα κανονικών κατανομών, η χρήση της απόστασης Mahalanobis ή η χρήση ιστογραμμάτων. Επιπροσθέτως, στα πλαίσια της ανίχνευσης ακραίων τιμών χωρίς τη βοήθεια ετικετών, μέθοδοι βασισμένοι στην πυκνότητα, όπως ο αλγόριθμος Local Outlier Factor (LOF), ή μέθοδοι βασισμένοι στην ομαδοποίηση (DBSCAN, CBLOF, K-Means, κ.λπ) μπορούν να χρησιμοποιηθούν.

Βασική υπόθεση και των δύο τύπων μεθόδων αποτελεί ότι τα ακραία σημεία τοποθετούνται σε περιοχές χαμηλής πυκνότητας. Συνέπεια αυτού αποτελεί το γεγονός ότι σε μια προσπάθεια ομαδοποίησης, τέτοιου είδους σημεία είτε αποτυγχάνουν να ομαδοποιηθούν σε κάποια ομάδα είτε δημιουργούν ομάδες οι οποίες μπορεί να είναι αραιές ή να χαρακτηρίζονται από χαμηλή πληθικότητα. Αντίθετα, στη περίπτωση όπου διαθέτουμε δεδομένα με ετικέτες έχουμε τη δυνατότητα να κατασκευάσουμε μοντέλα εποπτευόμενης μάθησης, είτε μοντελοποιώντας και τις δύο κλάσης ή μια εκ των δύο κλάσεων (συνήθως λόγο έλλειψης δεδομένων μοντελοποιείται η κλάση των φυσιολογικών παρατηρήσεων) και θεωρώντας ότι όσες παρατηρήσεις ξεφεύγουν από το μοντέλο προέρχονται από την εναλλακτική κλάση.

Τέλος, σε εφαρμογές όπου ο κύριος όγκος δεδομένων είναι χωρίς ετικέτα και ο αριθμός των δεδομένων με ετικέτες είναι σχετικά μικρός, μπορούμε να χρησιμοποιήσουμε τεχνικές ημι-εποπτευόμενης μάθησης. Σε αυτή τη περίπτωση μπορούμε να χρησιμοποιήσουμε τα ¨φυσιολογικά¨ δεδομένα σε συνδυασμό με κάποια δεδομένα χωρίς ετικέτα, τα οποία είναι ¨κοντά¨ στα φυσιολογικά, για την εκπαίδευση ενός μοντέλου μιας κλάσης. Εναλλακτικά,

μπορούμε να εμπλουτίσουμε το σύνολο των ακραίων τιμών χρησιμοποιώντας μη εποπτευόμενες τεχνικές μάθησης. Στη συνέχεια, περιγράφουμε τα θεωρητικά μοντέλα τα οποία χρησιμοποιήθηκαν με σκοπό τη δημιουργία μοντέλων μιας κλάσης για τα φυσιολογικά δεδομένα.

4.2 Μηχανές Διανυσμάτων Υποστήριξης Μιας Κλάσης (One Class-SVM)

Στο [52] οι Schölkopf et al. προτείνουν μια επέκταση των κλασσικών Μηχανών Διανυσμάτων Υποστήριξης με σκοπό τη χρήση τους για εύρεση ανωμαλιών. Το μοντέλο προσπαθεί να περικλείσει φυσιολογικές παρατηρήσεις με αποτέλεσμα να θεωρεί τις υπόλοιπες ως μη φυσιολογικές. Πιο συγκεκριμένα, έχοντας μετασχηματίσει τα δεδομένα σε ένα χώρο χαρακτηριστικών F (όπου μπορεί να υπάρχει ένα υπερεπίπεδο που διαχωρίζει τις παρατηρήσεις σύμφωνα με την αντίστοιχη κλάση) διαχωρίζουμε όλα τα σημεία στον αρχικό χώρο μεγιστοποιώντας την απόσταση του υπερεπιπέδου από τον αρχικό χώρο. Αυτό έχει ως αποτέλεσμα μια δυαδική συνάρτηση η οποία καταγράφει περιοχές στον χώρο εισόδου, της πυκνότητας πιθανότητας των δεδομένων. Η παραπάνω συνάρτηση επιστρέφει +1 σε μια ¨μικρή¨ περιοχή η οποία θεωρείται περιοχή μη ακραίων τιμών και −1 στον υπόλοιπο χώρο. Είναι προφανές ότι για την εκπαίδευση του μοντέλου αρκεί η επίλυση ενός προβλήματος ελαχιστοποίησης (Τετραγωνικός Προγραμματισμός), το οποίο ορίζεται ως εξής:

$$\min_{w,\xi_i,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

υπό τις ακόλουθες συνθήκες:

$$\begin{aligned}(w \cdot \phi(x_i)) &\geq \rho - \xi_i, \quad \forall i = 1, \ldots, n\\ \xi_i &\geq 0, \qquad\quad \forall i = 1, \ldots, n\end{aligned}$$

Σημαντική παράμετρο αποτελεί η τιμή του ν καθώς καθορίζει τα αποτελέσματα του μοντέλου. Πρακτικά η προαναφερθείσα παράμετρος θέτει, κατά τη διαδικασία της εκπαίδευσης του μοντέλου, τόσο ένα άνω φράγμα για το ποσοστό των παρατηρήσεων που μπορούν να θεωρηθούν ως ακραίες τιμές καθώς και ένα κάτω φράγμα για τον αριθμό από παρατηρήσεις που μπορούν να χρησιμοποιηθούν ως σημεία support vectors.

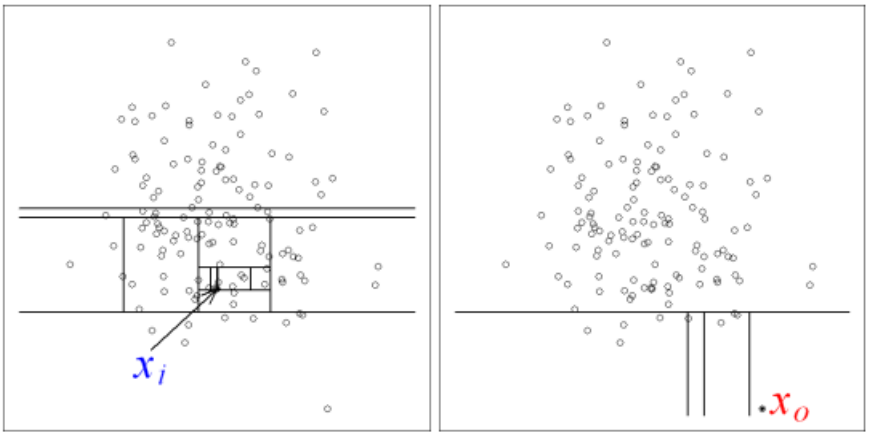
Τέλος αντίστοιχα με τις συνήθεις Μηχανές Διανυσμάτων Υποστήριξης για την επίλυση του παραπάνω προβλήματος βελτιστοποίησης χρησιμοποιείται η τεχνική Lagrange με αποτέλεσμα η συνάρτηση απόφασης να έχει την ακόλουθη μορφή:

$$f(x) = \operatorname{sgn}\left((w \cdot \phi(x_i)) - \rho\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} a_i K(x, x_i) - \rho\right)$$

Σε αυτό το σημείο αξίζει να σημειωθεί ότι οι Μηχανές Διανυσμάτων Υποστήριξης Μιας Κλάσης δύναται να εκπαιδευτούν τόσο με όσο και χωρίς επίβλεψη.

4.3 Δάση Απομόνωσης (Isolation Forest)

Τα Δάση Απομόνωσης προτάθηκαν από τον Fei Tony Liu [53], [54] και αποτελούν έναν αλγόριθμό εύρεσης ανωμαλιών ο οποίος βασίζεται στη λογική της απομόνωσης για τον εντοπισμό των ανωμαλιών. Αντίστοιχα με τα Σύνολα Δέντρων Ταξινόμησης το Δάσος Απομόνωσης επιλέγει τυχαία ένα χαρακτηριστικό για το οποίο στη συνέχεια επιλέγεται τυχαία μια τιμή διαχωρισμού στο εύρος των παρατηρηθέντων τιμών για το χαρακτηριστικό. Χρησιμοποιώντας επαναληπτικά ένα τέτοιο τυχαίο διαχωρισμό περιμένουμε οι μη φυσιολογικές τιμές να τοποθετούνται πιο κοντά στη ρίζα του δέντρου (μικρότερα μέσα μήκη μονοπατιών) και να απαιτούν λιγότερους διαχωρισμούς σε σχέση με τις φυσιολογικές τιμές (μικρότερο αριθμό από αναγκαίους διαχωρισμούς μέχρι ότου να απομονωθεί το σημείο) καθώς αυτές οι τιμές εμφανίζουν χαμηλότερη συχνότητα και αναμένουμε να τοποθετούνται σε διαχωρίσιμα, από τις φυσιολογικές τιμές, πεδία του χώρου. Στο Σχήμα 4.2 παρουσιάζουμε την προαναφερθείσα υπόθεση. Παρατηρούμε ότι για την απομόνωση του μη φυσιολογικό σημείο (δεξιά) απαιτούνται σημαντικά λιγότεροι διαχωρισμοί σε σχέση με το φυσιολογικό σημείο (αριστερά).



Σχήμα 4.2: *Δάσος Απομόνωσης - Προσδιορισμός φυσιολογικών (αριστερά) έναντι μη φυσιολογικών παρατηρήσεων (δεξιά)*

Αντίστοιχα με άλλες μεθόδους εύρεσης ακραίων τιμών, τα Δάση Απομόνωσης υπολογίζουν ένα σκορ για κάθε παρατήρηση με βάση το οποίο καθορίζεται εάν η παρατήρηση είναι φυσιολογική ή όχι. Σε ότι αφορά τα Δάση Απομόνωσης αυτό ορίζεται ως εξής:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}},$$

όπου ως $h(x)$ ορίζουμε το μήκος του μονοπατιού για τη παρατήρηση x από τον αρχικό κόμβο μέχρι την απομόνωση του σε κάποιο τερματικό κόμβο και ως $E(h(x))$ την εκτίμηση της μέση τιμή του $h(x)$ με βάση ένα σύνολο από Δέντρα Απομόνωσης. Σχετικά με τον όρο $c(n)$, αυτός αποτελεί τη μέση τιμή του $h(x)$ και χρησιμοποιείται με σκοπό τη κανονικοποίηση του κλάσματος. Συγκεκριμένα ορίζεται ως το μέσο μήκος του μονοπατιού για μη επιτυχημένες αναζητήσεις σε ένα Δέντρο Δυαδικής Αναζήτησης (καθώς τα Δέντρα Απομόνωσης έχουν

αντίστοιχη δομή με τα Δέντρα Δυαδικής Αναζήτησης)

$$c(n) = 2H(n-1) - (2(n-1)/n),$$

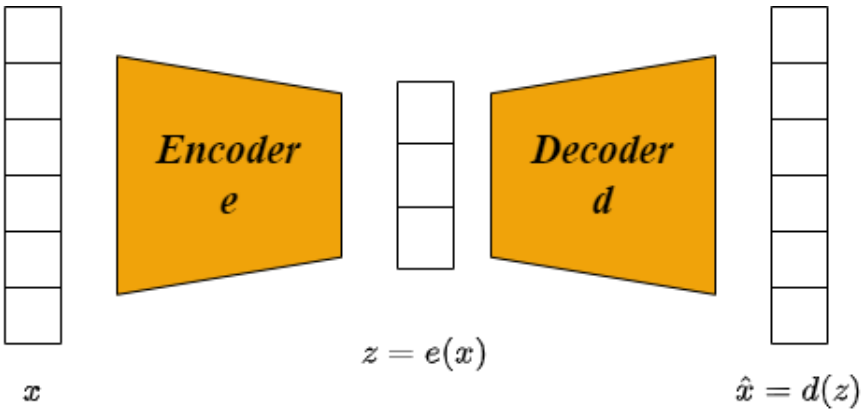
όπου το $H(i)$ αποτελεί τον αρμονικό αριθμό και μπορεί να εκτιμηθεί ως $ln(i)$+Euler's constant. Τέλος, έχοντας καθορίσει τον παραπάνω δείκτη για κάθε παρατήρηση, έχουμε τη δυνατότητα να κατατάξουμε τις παρατηρήσεις ως φυσιολογικές ή μη φυσιολογικές με βάση τους ακόλουθους κανόνες:

- Παρατηρήσεις με **s πολύ κοντά στη μονάδα** θεωρούνται ως **μη φυσιολογικές**.
- Παρατηρήσεις με **s μικρότερο από 0.5** θεωρούνται ως **φυσιολογικές**.
- Εάν **όλες οι παρατηρήσεις επιστρέφουν s πολύ κοντά στο 0.5** θεωρούμε ότι **το σύνολο δεδομένων δεν αποτελείται από μη φυσιολογικές παρατηρήσεις**.

Αντίστοιχα με τις Μηχανές Διανυσμάτων Υποστήριξης Μιας Κλάσης, τα Δάση Απομόνωσης μπορούν να εκπαιδευτούν τόσο με, όσο και χωρίς, επίβλεψη.

4.4 Αυτοκωδικοποιητές

Οι αυτοκωδικοποιητές προτάθηκαν το μακρινό 1986 από τους Rumelhart et al. [55]. Αποτελούν μια μορφή νευρωνικών δικτύων τα οποία μαθαίνουν μέσω μη επιβλεπόμενης μάθησης χρησιμοποιώντας δεδομένα χωρίς ετικέτα. Παραδοσιακά, οι αυτοκωδικοποιητές έχουν χρησιμοποιηθεί σε προβλήματα μείωσης διάστασης και εξαγωγής χαρακτηριστικών καθώς και ως παραγωγικά μοντέλα δεδομένων [56], [57], [58]. Σκοπό των συγκεκριμένων μοντέλων αποτελεί αρχικά η εκμάθηση κάποιας απεικόνισης μειωμένης διάστασης για ένα σύνολο από αρχικά δεδομένα και στη συνέχεια η επιτυχής ανακατασκευή των αρχικών δεδομένων με τρόπο τέτοιο ώστε να ελαχιστοποιείται το σφάλμα ανακατασκευής των αρχικών δεδομένων.



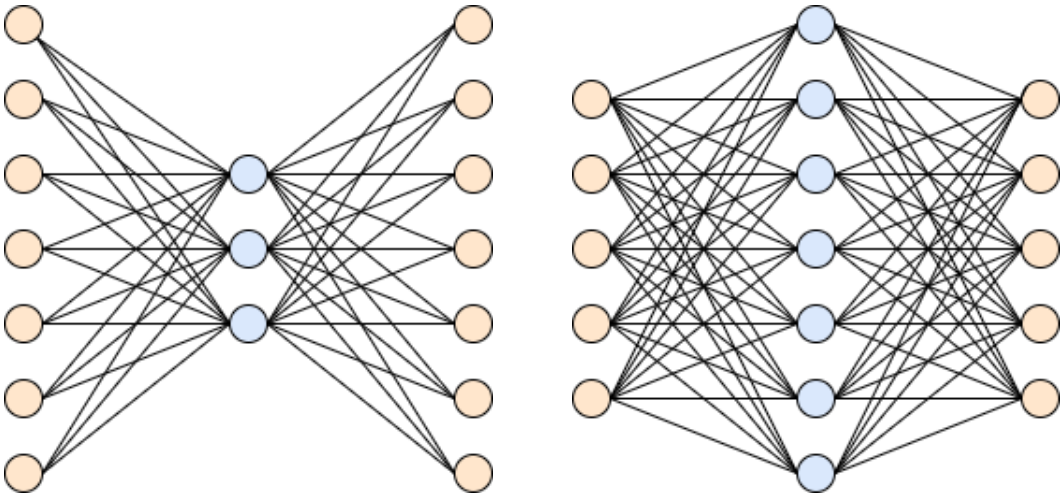
Σχήμα 4.3: *Γενική Μορφή Αυτοκωδικοποητών*

Είναι λοιπόν προφανές ότι οι αυτοκωδικοποιητές (Σχήμα 4.3) αποτελούνται από δύο ευδιάκριτα μέρη, ένα κωδικοποιητή (encoder) και έναν αποκωδικοποιητή (decoder). Πιο

συγκεκριμένα, έστω X υποσύνολο του $\mathbb{R}^p$ και $x \in X$ κάθε στοιχείο του συνόλου X. Αντίστοιχα, ο κωδικοποιητής και ο αποκωδικοποιητής ορίζονται ως συναρτήσεις μετάβασης e, d τέτοιες ώστε:

$$e : X \subseteq \mathbb{R}^p \rightarrow Z \subseteq \mathbb{R}^l$$
$$d : Z \subseteq \mathbb{R}^l \rightarrow X \subseteq \mathbb{R}^p$$
$$e, d = \arg\min_{e,d} ||X - (d \circ e)X||^2$$

Στη πιο συνήθης περίπτωση το l είναι σημαντικά μικρότερο από το p και οι αυτοκωδικοποιητές αποκαλούνται υποπλήρης (Σχήμα 4.4α') ενώ σε περιπτώσεις όπου το l είναι μεγαλύτερο από το p, υπερπλήρεις (Σχήμα 4.4β')



(α') *Υποπλήρης Αυτοκωδικοποιητής* (β') *Υπερπλήρης Αυτοκωδικοποιητής*

Σχήμα 4.4: *Παραδείγματα Αυτοκωδικοποιητών*

Στην απλούστερη περίπτωση όπου έχουμε μόνο ένα κρυφό επίπεδο, ο Αυτοκωδικοποιητής δέχεται ως εισαχθέν δεδομένα $x \in \mathbb{R}^p$ το οποίο στη συνέχεια περνάει από τον κωδικοποιητή $z = e(x) \in \mathbb{R}^l$ το οποίο μπορεί να γραφτεί και ως:

$$x = \sigma(Wh + b),$$

όπου το σ αποτελεί τη συνάρτηση ενεργοποίησης (π.χ. $tanh, ReLU, SELU$ κ.λπ), W το πίνακα βαρών και b το διάνυσμα της μεροληψίας. Με Z συμβολίζουμε το λανθάνων χώρο απεικόνισης και z τη λανθάνουσα (κωδικοποιημένη) απεικόνιση. Αντίστοιχα, μέσω της χρήσης του αποκωδικοποιητή εφαρμόζεται ο αντίστροφος μετασχηματισμός με σκοπό την ανακατασκευή των δεδομένων εισόδου με βάση τη λανθάνουσα απεικόνιση z. Συνεπώς για τη περίπτωση με ένα κρυμμένο επίπεδο το ανακατασκευασμένο σημείο $\hat{x}$ υπολογίζεται ως:

$$\hat{x} = \tilde{\sigma}(\tilde{W}h + \tilde{b}),$$

όπου τα $\tilde{\sigma}, \tilde{W}$ και $\tilde{b}$ δύναται να διαφέρουν από τα αντίστοιχα του κωδικοποιητή.

Σε αυτό το σημείο αξίζει να σημειωθεί ότι στη περίπτωση κατά την οποία διαθέτουμε ένα Αυτοκωδικοποιητή με κρυφά επίπεδα στα οποία χρησιμοποιούνται αποκλειστικά γραμμικές συναρτήσεις ενεργοποίησης τότε καταλήγουμε σε μοντέλο σχεδόν αντίστοιχο με την Ανάλυση

Κυρίων Συνιστωσών. Παρόλο αυτά εάν έχουμε Αυτοκωδικοποιητές, οι οποίοι αποτελούνται μόνο από ένα κρυφό επίπεδο, τότε ανεξαρτήτως του είδους της συνάρτησης ενεργοποίησης, ο βέλτιστος Αυτοκωδικοποιητής θα ταυτίζεται με τη λύση από την Ανάλυση σε Κύριες Συνιστώσες. Στη πραγματικότητα ο πίνακας W που υπολογίζει ο Αυτοκωδικοποιητής δεν είναι απαραίτητα ίδιος με τον αντίστοιχο πίνακα από την Ανάλυση Κυριών Συνιστωσών όμως ο υπόχωρος που εκτείνεται από τα αντίστοιχα W θα είναι ίδιος [59].

Επιπροσθέτως, οι επιλεγμένες αρχιτεκτονικές για τον κωδικοποιητή και των αποκωδικοποιητή δύναται είτε να είναι ίδιες (συμμετρικοί) είτε να διαφέρουν (ασύμμετροι) καθώς και να διαθέτουν περισσότερα από ένα κρυφά επίπεδα, στη περίπτωση αυτή οι Αυτοκωδικοποιητές ονομάζονται Βαθιά Στοιβαγμένοι Αυτοκωδικοποιητές. Τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής μπορούν να αποτελούνται από τα κλασικά νευρωνικά δίκτυα εμπρόσθια κίνησης καθώς και από συνελικτικά και επαναλαμβανόμενα νευρωνικά δίκτυα ακόμα και από παραγωγικά αντιμαχόμενα δίκτυα.

Για την εκπαίδευση των μοντέλων ελαχιστοποιούμε το λάθος ανακατασκευής μεταξύ των σημείων εισόδου x_i και των αντίστοιχων δεδομένων εξόδου $\hat{x}_i$.

$$\text{loss} = \|x - \hat{x}\|^2 = \|x - d(z)\|^2 = \|x - d(e(x))\|^2,$$

Το παραπάνω κριτήριο συνήθως εκφράζεται μέσω του μέσου τετραγωνικού σφάλματος :

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2$$

όπου με $\|\cdot\|_2$ συμβολίζεται η L^2-νόρμα και με N το πλήθος δεδομένων στο σύνολο εκπαίδευσης, είτε μέσω της δυαδικής διασταυρούμενης εντροπίας. Τέλος η ελαχιστοποίηση των κριτηρίων γίνεται εφικτή μέσω της μεθόδου οπισθοδιάδοσης για όλες τις παραμέτρους του δικτύου.

Στο πλαίσιο του προβλήματος της εύρεσης ανωμαλιών, οι Αυτοκωδικοποιητές μπορούν να εκπαιδευτούν χρησιμοποιώντας αποκλειστικά ¨φυσιολογικά¨ δεδομένα. Με αυτό το τρόπο μοντελοποιείται η κατανομή των ¨φυσιολογικών¨ παρατηρήσεων. Παρόλο αυτά ο διανυσματικός χώρος παρουσιάζει καλά χαρακτηριστικά για περιοχές στις οποίες έχει παρατηρήσει δεδομένα κατά τη διαδικασία της εκπαίδευσης. Ως αποτέλεσμα σε περιοχές μακριά από τη κατανομή των φυσιολογικών δεδομένων το σφάλμα ανακατασκευής παρουσιάζει συγκριτικά υψηλότερες τιμές και συνεπώς διαφορετική κατανομή κάτι το οποίο έρχεται σε συμφωνία με την αρχική υπόθεση σχετικά με τη παραγωγή των ακραίων τιμών, ότι δηλαδή ο μηχανισμός παραγωγής τους διαφέρει σημαντικά από αυτόν για τις φυσιολογικές παρατηρήσεις. Συμπερασματικά, κατά τη διαδικασία της επικύρωσης - ελέγχου του μοντέλου ταξινομούνται ως ανωμαλίες, οι παρατηρήσεις για τις οποίες το σφάλμα ανακατασκευής είναι υψηλότερο από κάποιο επιβληθέν κατώφλι.

Στους κλασικούς Αυτοκωδικοποιητές έχουν προταθεί διάφορες παραλλαγές στοχεύοντας στο να ωθήσουν τα μοντέλα να μάθουν απεικονίσεις στο λανθάνων χώρο με χρήσιμες ιδιότητες. Τέτοια είδη αποτελούν οι Κανονικοποιημένοι Αυτοκωδικοποιητές, όπως οι Αραιοί Αυτοκωδικοποιητές και οι Αυτοκωδικοποιητές Αφαίρεσης Θορύβου, καθώς και οι Μεταβλητοί Αυτοκωδικοποιητές οι οποίοι μπορούν να χρησιμοποιηθούν ως παραγωγικά μοντέλα.

4.4.1 Αραιοί Αυτοκωδικοποιητές

Οι Αραιοί Αυτοκωδικοποιητές έχουν την δυνατότητα να δώσουν καλύτερα αποτελέσματα σε προβλήματα ταξινόμησης, ωθώντας σε αραιότερες απεικονίσεις στο χώρο [60], [61]. Πιο συγκεκριμένα, οι Αραιοί Αυτοκωδικοποιητές περιλαμβάνουν στη συνάρτηση απώλειας τους κάποιον όρο ποινής $\Omega(z)$ ικανό να επιβάλει την συνθήκη της αραιότητας. Στη γενική μορφή της η συνάρτηση απώλειας έχει την ακόλουθη μορφή:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) + \Omega(z)$$

Με αυτό τον τρόπο το νευρωνικό δίκτυο ωθείται προς την ενεργοποίηση συγκεκριμένων νευρώνων με βάση τα δεδομένα εισόδου και την απενεργοποίηση διαφορετικών περιοχών του δικτύου. Το παραπάνω γίνεται εφικτό μέσω της χρήσης διαφόρων μεθόδων όπως η χρήση της απόκλισης Kullback-Leibler (KL) [62] και με την επιβολή L1 ή L2 Κανονικοποίησης. Υποθέτοντας δείγμα εισόδου $X = \{x_1, ..., x_n\}$ μεγέθους n και a_j συνάρτηση ενεργοποίησης για το κρυφό επίπεδό j, τότε μπορούμε να ορίσουμε τη μέση τιμή του j-οστού επιπέδου ως:

$$\hat{\rho}_j = \frac{1}{n} \sum_{i=1}^{n} \left[a_j(x_i) \right]$$

Οι τιμές του $\hat{\rho}_j$ θέλουμε να είναι κοντά στη τιμή του ρ, όπου ρ αποτελεί παράμετρο αραιότητας με τη τιμή της να καθορίζεται κοντά στο μηδέν. Θέτοντας τον προηγούμενο περιορισμό οι ενεργοποιήσεις των νευρώνων ωθούνται κοντά στο μηδέν επιτυγχάνοντας με αυτό το τρόπο την αραιότητα στο επίπεδο των νευρώνων. Για να επιβάλλουμε το παραπάνω περιορισμό ελαχιστοποιούμε την απώλεια του δικτύου, προσαυξημένη με έναν όρο ικανό να ποινικοποιήσει $\hat{\rho}_j$ τα οποία αποκλίνουν σημαντικά από το ρ σαν τον ακόλουθο:

$$KL\left(\rho || \hat{\rho}_j\right) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j}$$

Ο παραπάνω όρος αποτελεί την απόκλιση KL ανάμεσα σε μια τυχαία μεταβλητή Bernoulli με μέση τιμή ρ και σε μια τυχαία μεταβλητή Bernoulli με μέση τιμή $\hat{\rho}_j$. Τέτοιες συναρτήσεις δύναται να ποσοτικοποιήσουν τη διαφοροποίησης ανάμεσα σε δύο κατανομές πιθανότητας. Σημαντική ιδιότητα του παραπάνω όρου ποινής αποτελεί το γεγονός ότι η συνθήκη $KL\left(\rho || \hat{\rho}_j\right) = 0$ ικανοποιείται εάν και μόνο εάν $\rho = \hat{\rho}_j$. Σε διαφορετική περίπτωση η τιμή της ποινής αυξάνεται μονοτονικά όσο το $\hat{\rho}_j$ αποκλίνει από το ρ. Συνεπώς αρκεί να προσθέσουμε τον παραπάνω όρο στη συνάρτηση απώλειας, αθροίζοντας φυσικά ως προς το πλήθος των κρυφών νευρώνων s το οποίο με τη σειρά του κανονικοποιείται από μια υπερπαράμετρο β.

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) + \beta \sum_{j=1}^{s} KL\left(\rho || \hat{\rho}_j\right)$$

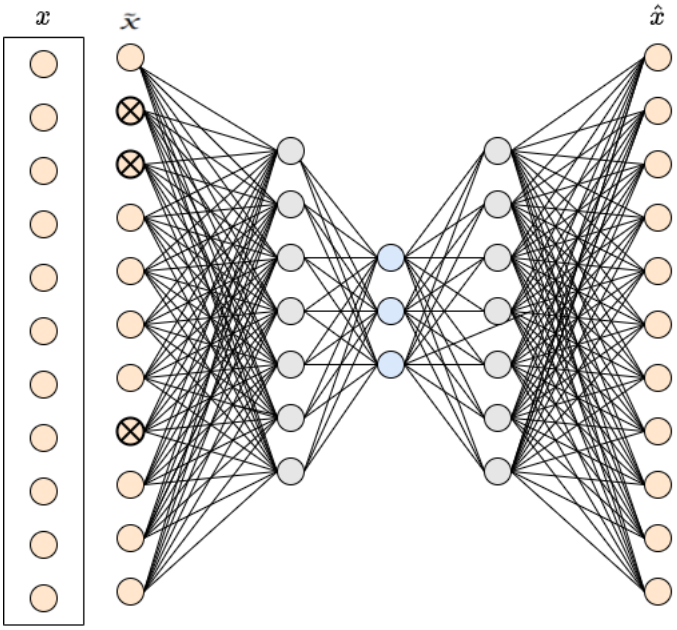
Αντίθετα, μέσω της εφαρμογής της L1 Κανονικοποίησης, προσθέτουμε στη συνάρτηση απώλειας ως ποινή την απόλυτη τιμή του διανύσματος από τις ενεργοποιήσεις στο λανθάνων χώρο Z το οποίο κανονικοποιείται από μια υπερπαράμετρο λ:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \sum_{i} |z_i|$$

4.4.2 Αυτοκωδικοποιητές Αφαίρεσης Θορύβου

Ένα από τα συχνότερα προβλήματα το οποίο εμφανίζεται σε overcompleted αυτοκωδικοποιητές είναι η εκμάθηση της ταυτοτικής συνάρτησης με αποτέλεσμα την υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης. Οι Αυτοκωδικοποιητές Αφαίρεσης Θορύβου έχουν την δυνατότητα να επιλύσουν το προαναφερθέν πρόβλημα [63].

Κάτι τέτοιο επιτυγχάνεται μέσω της εκ προμελέτης διαφθοράς των δεδομένων εισόδου x κατά τη διαδικασία εκπαίδευσης σε $\tilde{x}$ μέσω κάποιας στοχαστικής διαδικασίας $\tilde{x} \sim q(\tilde{x} \mid x)$. Συνεπώς για κάθε καινούργιο σημείο x το οποίο παρουσιάζεται στο μοντέλο, ένα καινούργιο μολυσμένο σημείο $\tilde{x}$ παράγεται στοχαστικά μέσω της $q(\tilde{x} \mid x)$.



Σχήμα 4.5: *Γενική Μορφή Βαθιών Αυτοκωδικοποιητών Αφαίρεσης Θορύβου*

Στη συνέχεια, το μοντέλο (Σχήμα 4.5) εκπαιδεύεται έτσι ώστε να μπορεί να ανακατασκευάζει τα μολυσμένα δεδομένα μεταφέροντας τα μολυσμένα δεδομένα σε κρυφές απεικονίσεις σύμφωνα με τις κλασικές και προαναφερθείσες αρχές των μοντέλων αυτοκωδικοποίησης [64], [38]. Κατά τον υπολογισμό της απώλειας του δικτύου συγκρίνονται οι πραγματικές τιμές των δεδομένων με τα ανακατασκευασμένα δεδομένα, μετά την μόλυνση τους, με αποτέλεσμα την επιτυχή επίλυση του προβλήματος της εκμάθησης της ταυτοτικής συνάρτησης [64]. Συνέπεια των παραπάνω αποτελεί η επίτευξη ¨καλών¨ απεικονίσεων και ανθεκτικών φίλτρων μειώνοντας ταυτόχρονα το κίνδυνο υπερπροσαρμογής του μοντέλου.

Σε ότι αφορά τη διαδικασία μόλυνσης των δεδομένων μπορεί να χρησιμοποιηθεί οποιαδήποτε γνωστή μέθοδος. Ορισμένες από τις πιο σύνηθες μεθόδους μόλυνσης των δεδομένων περιγράφονται παρακάτω.

- **Masking Noise**
  Ένα ποσοστό από τα χαρακτηριστικά εισόδου επιλέγεται τυχαία και οι τιμές των επιλεγέντων χαρακτηριστικών αντικαθιστώνται από μηδενικές τιμές.

- **Salt & Pepper Noise**
  Ένα ποσοστό από τα χαρακτηριστικά εισόδου επιλέγεται τυχαία και οι τιμές των επιλεγέντων χαρακτηριστικών αντικαθιστώνται από την ελάχιστη ή τη μέγιστη τιμή με ομοιόμορφη πιθανότητα.

- **Gaussian Noise**
  Θόρυβος από Κανονικές Κατανομές προστίθενται στα δεδομένα εισόδου.

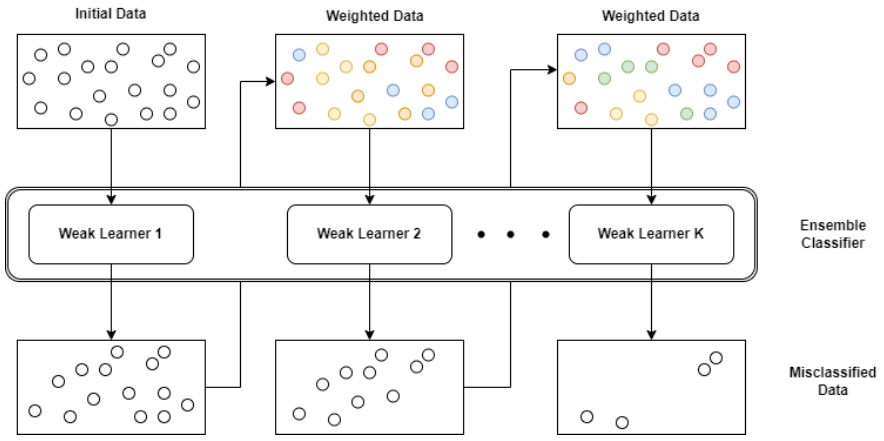
Κεφάλαιο 5

Επιβλεπόμενη Εύρεση Ανωμαλιών

5.1 Μοντέλα Ενισχυτικής Κλίσης

Τα μοντέλα βαθιάς μάθησης έχουν καταφέρει να παράξουν ενθαρρυντικά αποτελέσματα σε ερευνητικά πεδία όπως η υπολογιστική όραση και η επεξεργασία φυσικής γλώσσας. Παρόλο αυτά, σε προβλήματα τα οποία αποτελούνται από δομημένα σύνολα δεδομένων σε μορφή πίνακα, τέτοιου είδος μοντέλα (π.χ. TabNet [65], VIME [66]) δεν παρουσιάζουν ικανοποιητικές αποδόσεις σε σχέση με απλούστερα και λιγότερα απαιτητικά μοντέλα, λαμβάνοντας υπόψη τόσο το υπολογιστικό κόστος όσο και τον όγκο των δεδομένων [67].

Αντιθέτως, μοντέλα τα οποία χρησιμοποιούν Ενισχυτική Κλίση (Gradient Boosting) καταλήγουν σε καλύτερα αποτελέσματα, ειδικά σε μη ισορροπημένα προβλήματα ταξινόμησης, διατηρώντας ένα σχετικά μικρό κόστος για την εκπαίδευση των μοντέλων καθώς και ορισμένες δυνατότητες ερμηνευσιμότητας των τελικών αποτελεσμάτων. Το Gradient Boosting (GB) [68] αποτελεί μια τεχνική στατιστικής μάθησης κατά την οποία συνδυάζονται απλά Μοντέλα Βάσης (Weak Learners) σε έναν Ταξινομητή Συνόλου (Ensemble Classifier). Στην περίπτωση της κλασικής Ενίσχυσης (Boosting), τα Μοντέλα Βάσης εκπαιδεύονται ακολουθιακά με διαφορετικά βάρη για κάθε παρατήρηση από το σύνολο δεδομένων εκπαίδευσης και συνδυάζοντας με διαφορετικά βάρη το σύνολο των μοντέλων βάσης με τρόπο τέτοιο ώστε το επόμενο μοντέλο να επηρεάζεται από το προηγούμενο, σε αντίθεση με τη τεχνική του Bagging όπου τα μοντέλα βάσης εκπαιδεύονται παράλληλα σε bootstrapped δείγματα με κάθε μοντέλο να παρουσιάζει την ίδια βαρύτητα στη τελική απόφαση [69].



Σχήμα 5.1: *Λειτουργία της μεθόδου Boosting για προβλήματα ταξινόμησης*

Πιο συγκεκριμένα για την εκπαίδευση ενός μοντέλου μέσω της τεχνικής Boosting (Σχήμα 5.1) ακολουθούνται τα παρακάτω βήματα:

1. Τα δεδομένα από το σύνολο εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου βάσης, χρησιμοποιώντας ίσα βάρη για όλα τα δεδομένα.

2. Το βάρος κάθε παρατήρησης μεταβάλλεται ανάλογα με το σφάλμα πρόβλεψης. Σε παρατηρήσεις με μεγάλο σφάλμα ανατίθενται μεγαλύτερα βάρη με σκοπό την βελτίωση των αποτελεσμάτων στην επόμενη φάση εκπαίδευσης.

3. Ένα βάρος ανατίθεται στο μοντέλο βάσης ανάλογα την απόδοση του.

4. Τα δεδομένα εκπαίδευσης σε συνδυασμό με τα βάρη τους χρησιμοποιούνται για την εκπαίδευση του εκ των υστέρων μοντέλου και επαναλαμβάνονται τα βήματα 2-3.

5. Το Βήμα 4 επαναλαμβάνεται για έναν προκαθορισμένο αριθμό επαναλήψεων ή μέχρις ότου επιτευχθεί κάποιο κατώφλι για το σφάλμα πρόβλεψης του μοντέλου.

Το GB διατηρεί στοιχεία από το κλασικό Boosting, αντιθέτως όμως, τα δεδομένα δεν διαθέτουν βάρη και το μοντέλο χρησιμοποιεί τα κατάλοιπα στις προβλέψεις των προηγούμενων μοντέλων βάσης. Επιπλέον, βελτιστοποιούν μια συνάρτηση απώλειας, χρησιμοποιώντας σταθερό learning rate για όλα τα απλά μοντέλα βάσης. Πιο συγκεκριμένα, ο σκοπός είναι η εύρεση μιας συνάρτησης $\hat{F}(x)$ τέτοιας ώστε να μπορεί να προσεγγίσει ικανοποιητικά τις μεταβλητές εξόδου δεδομένων των μεταβλητών εισόδου. Αυτό γίνεται εφικτό μέσω της ελαχιστοποίησης της αναμενόμενης απώλειας στα πλαίσια της ελαχιστοποίησης του εμπειρικού ρίσκου:

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))].$$

Η εκτιμώμενη συνάρτηση $\hat{F}(x)$ προσεγγίζεται ως ένα σταθμισμένο σύνολο από K συναρτήσεις $h_k(x) \epsilon \mathcal{H}$ οι οποίες ανήκουν σε μια κλάση συναρτήσεων $\mathcal{H}$. Στη πράξη, αυτές αποτελούν τα προαναφερθεί μοντέλα βάσης.

$$\hat{F}(x) = \sum_{k=1}^{K} \gamma_k h_k(x) + \text{const.}$$

Για την εκτίμηση της $\hat{F}(x)$ χρησιμοποιείται μια ακολουθιακή διαδικασία, ξεκινώντας από ένα σταθερό μοντέλο $F_0(x)$ και προσθέτοντας πιο σύνθετα μοντέλα στη συνέχεια:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right)$$

$$F_k(x) = F_{k-1}(x) + \arg\min_{h_k \in \mathcal{H}} \left[\sum_{i=1}^{n} L\left(y_i, F_{k-1}\left(x_i\right) + h_k\left(x_i\right)\right) \right]$$

Η εύρεση της βέλτιστης συνάρτησης h_k σε κάθε βήμα αποτελεί ένα πρόβλημα ελαχιστοποίησης το οποίο δεν έχει λύση για υπολογιστικούς λόγους. Συνεπώς χρησιμοποιείται μια προσέγγιση αυτού χρησιμοποιώντας τη μέθοδο καθόδου κλίσης με σκοπό την εύρεση ενός τοπικού ελάχιστου για μια συνάρτηση απώλειας λαμβάνοντας υπόψη τις προηγούμενες

$F_{k-1}(x)$. Ως αποτέλεσμα, η εξίσωση για την εύρεση των $F_k(x)$ λαμβάνει την ακόλουθη μορφή:

$$F_k(x) = F_{k-1}(x) - \gamma_k \sum_{i=1}^{n} \nabla_{F_{k-1}} L\left(y_i, F_{k-1}\left(x_i\right)\right),$$

όπου $\gamma > 0$ αποτελεί μια μικρή τιμή τέτοια ώστε να ισχύει η γραμμική προσέγγιση. Το γ βελτιστοποιείται σε κάθε βήμα:

$$\gamma_k = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_k\left(x_i\right)\right) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{k-1}\left(x_i\right) - \gamma \nabla_{F_{k-1}} L\left(y_i, F_{k-1}\left(x_i\right)\right)\right)$$

Στη συνέχεια παρουσιάζεται, λεπτομερώς, ο γενικός αλγόριθμος για ένα σύνολο δεδομένων εκπαίδευσης $\{(x_i, y_i)\}_{i=1}^{n}$ και μια διαφορίσιμη συνάρτηση απώλειας $L(y, F(x))$, όπου K ο αριθμός από βήματα ενίσχυσης.

Αλγοριθμος 5.1: *Γενικός Αλγόριθμος Gradient Boosting*

1: Αρχικοποίηση ενός σταθερού μοντέλου: $F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right)$

2: Για $k = 1$ έως K:

3: i. Υπολόγισε τα ψευδό-κατάλοιπα:

$$r_{ik} = -\left[\frac{\partial L\left(y_i, F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right]_{F(x)=F_{k-1}(x)} \quad \text{για } i = 1, \ldots, n$$

4: ii. Εκπαίδευσε μοντέλο βάσης χρησιμοποιώντας τα ψευδό-κατάλοιπα στο $\{(x_i, r_{ik})\}_{i=1}^{n}$

5: iii. Υπολόγισε το γ_k λύνοντας το ακόλουθο πρόβλημα ελαχιστοποίησης:

$$\gamma_k = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{k-1}\left(x_i\right) + \gamma h_k\left(x_i\right)\right)$$

6: iv. Ανανέωσε το μοντέλο

$$F_k(x) = F_{k-1}(x) + \gamma_k h_k(x)$$

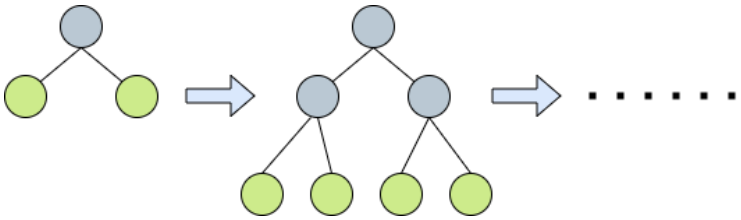
7: Επέστρεψε την $F_K(x)$

Συνήθως ως Μοντέλα Βάσης επιλέγονται απλά Δέντρα Απόφασης (CART), με λίγα φύλλα και σχετικά μικρό βάθος. Ορισμένα από τα πιο σημαντικά Μοντέλα Ενισχυτικής Κλίσης αποτελούν οι Μηχανές Ενισχυτικής Κλίσης (Gradient Boosting Machines - GBM) [70], οι οποίες συνδυάζουν τις προβλέψεις από πολλαπλά δέντρα απόφασης, τα οποία μαθαίνουν από τα σφάλματα των προγενέστερων. Κάθε κόμβος από τα δέντρα χρησιμοποιεί διαφορετικό υποσύνολο από χαρακτηριστικά για τον καθορισμό του βέλτιστου σημείου διαχωρισμού. Πέραν αυτών, αρκετά διαδεδομένα είναι μοντέλα όπως τα XGBoost, LightGBM, CatBoost (λεπτομέρειες παρατίθενται παρακάτω), τα οποία επεξεργάζονται παράλληλα τα δεδομένα και σε αντίθεση με τα GBM κανονικοποιούν τις συναρτήσεις απώλειας και ¨κλαδεύουν¨ τα δέντρα απόφασης με αποτέλεσμα να μειώνεται η πιθανότητα υπερπροσαρμογής του μοντέλου.

5.1.1 XGBoost

Ο αλγόριθμος XGBoost (Extreme Gradient Boosting) αναπτύχθηκε το 2014 από τους Tianqi Chen and Carlos Guestrin στα πλαίσια ενός ερευνητικού προγράμματος στο Πανεπιστήμιο της Ουάσιγκτον [71]. Σαν μοντέλο υποστηρίζει την κατανεμημένη εκπαίδευση καθώς και την αξιοποίηση μονάδων γραφικής επεξεργασίας με κύρια χαρακτηριστικά του να αποτελούν οι πολύ καλές επιδόσεις σε συνδυασμό με το μειωμένο κόστος εκπαίδευσης, από άποψη χρόνου και υπολογιστικών πόρων. Όπως και τα GBM έτσι και ο αλγόριθμος XGBoost αποτελούν μοντέλα συνόλου τα οποία εφαρμόζουν την τεχνική Boosting μέσω τη μεθόδου καθόδου κλίσης. Ο αλγόριθμος XGBoost καινοτομεί βελτιστοποιώντας τόσο το σύστημα όσο και το τρόπο λειτουργίας προγενέστερων GB μοντέλων. Αρχικά σχετικά με τη βελτιστοποίηση του συστήματος, τα δεντρά απόφασης δημιουργούνται παράλληλα και όχι ακολουθιακά χρησιμοποιώντας όλους τους διαθέσιμους πυρήνες της CPU. Επιπλέον αξιοποιείται το διαθέσιμο hardware στο μέγιστο βελτιστοποιώντας το cache, διανέμοντας κατάλληλα την κατανομή εσωτερικών buffer σε κάθε νήμα για την αποθήκευση στατιστικών κλίσης. Επίσης υλοποιούνται υπολογισμοί ¨εκτός πυρήνα¨ για σύνολα δεδομένων τα οποία δεν χωράνε στην μνήμη.

Συνεχίζοντας με τις βελτιώσεις στο επίπεδο του αλγόριθμου, τα φύλλα των δέντρα απόφασης αναπτύσσονται χρησιμοποιώντας τη μέθοδο Depth-First (Level-Wise) κατά την οποία ο επόμενος κόμβος διαχωρισμού επιλέγεται με βάση τη βελτίωση στο προγενέστερο κόμβο επιλέγοντας δηλαδή το κόμβο ο οποίος αυξάνει τη καθαρότητα των παρατηρήσεων, αυξάνοντας τη καθαρότητα των κόμβων (Σχήμα 5.2). Τα δέντρα αναπτύσσονται μέχρι ένα προκαθορισμένο μέγιστο βάθος και στη συνέχεια περικόπτονται αφαιρώντας κόμβους διαχωρισμούς για τους οποίους δεν υπάρχει κάποιο θετικό κέρδος με σκοπό την αποφυγή προβλημάτων υπερπροσαρμογής.



Σχήμα 5.2: *Ανάπτυξη Δέντρων - Depth-First (Level-Wise)*

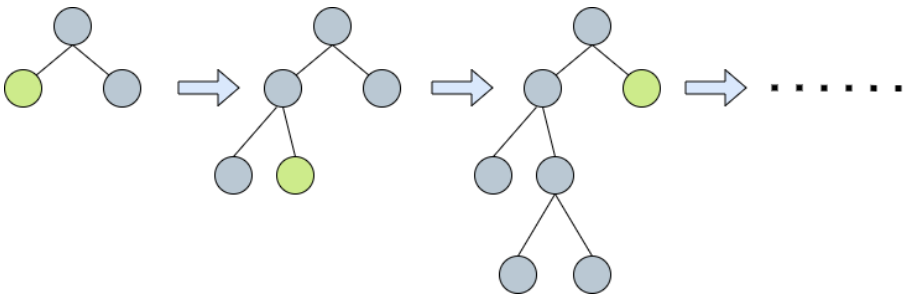
Για την εύρεση του εκάστοτε σημείου διαχωρισμού ο αλγόριθμος χρησιμοποιεί μια άπληστη (greedy) στρατηγική κατά την οποία λαμβάνει υπόψη όλα τα δεδομένα εκπαίδευσης χρησιμοποιώντας μια πρωταρχική διάταξη αυτών. Πιο συγκεκριμένα για κάθε κόμβο απαριθμούνται όλα τα χαρακτηριστικά και για κάθε χαρακτηριστικό διατάσσονται οι τιμές. Στη συνέχεια χρησιμοποιείται γραμμική αναζήτηση με σκοπό το καθορισμό των επιμέρους σημείων διαχωρισμού σύμφωνα με το επιλεγέν κριτήριο για τη ποσοτικοποίηση της πληροφορίας κέρδους. Τέλος, αρκεί να επιλεχθεί το βέλτιστο σημείο διαχωρισμού λαμβάνοντας υπόψη όλα τα χαρακτηριστικά. Είναι προφανές ότι η προαναφερθείσα διαδικασία με βεβαιότητα προκαλεί προβλήματα ειδικά για μεγάλα σύνολα δεδομένων καθώς ο χρόνος εκπαίδευσης αυξάνεται σημαντικά. Για αυτό το λόγο τα σημεία διαχωρισμού μπορούν να προσεγγιστούν χρησιμοποιώντας στρατηγικές βασισμένες στα ποσοστημόρια των χαρακτηριστικών. Ζυγισμένα Ποσοστημοριακά Σχήματα (Weighted Quantile Sketches) χρησιμοποιούνται σε κάθε

βήμα με σκοπό την εύρεση του βέλτιστου σημείου διαχωρισμού για δεδομένα, τα οποία είναι πιθανόν να διαθέτουν ακόμη και βάρη. Τα σταθμισμένα δεδομένα μπορούν να θεωρηθούν ως σταθμισμένα ποσοστημόρια των οποίων οι δειγματικές κατανομές διαθέτουν τα ίδια στατιστικά χαρακτηριστικά με τα πραγματικά ποσοστημόρια. Επιπροσθέτως, παρόλο που ο αλγόριθμος δεν διαθέτει κάποιο μηχανισμό για τη μοντελοποίηση κατηγορικών μεταβλητών ή ελλιπών τιμών, διαθέτει μηχανισμό κατάλληλο για την αντιμετώπιση αραιών χαρακτηριστικών. Σε κάθε δενδρικό κόμβο ο αλγόριθμος προσδίδει μια προκαθορισμένη κατεύθυνση για το κόμβο διαχωρισμού με αποτέλεσμα το μοντέλο να συμπεριφέρεται σε κάθε αραιή παρατήρηση ως μοναδική εφόσον έχει μάθει τα μοτίβα αραιότητας.

Τέλος, για την επίλυση προβλημάτων υπερπροσαρμογής, όροι L1-Lasso και L2-Ridge κανονικοποίησης προστίθενται στην αντικειμενική συνάρτηση, τεχνικές συρρίκνωσης χρησιμοποιούνται στα βάρη κάθε boosting δέντρου με αποτέλεσμα να μειώνεται η επιρροή του κάθε δέντρου στα μελλοντικά και τεχνικές υποδειγματοληψίας ανά χαρακτηριστικό εφαρμόζονται ανά δέντρο, επίπεδο ή κόμβο δέντρου.

5.1.2 LightGBM

Ο αλγόριθμος LightGBM αναπτύχθηκε το 2016 από τον Guolin Ke και την ερευνητική ομάδα της Microsoft [72]. Αποτελεί με τη σειρά του ένα βελτιστοποιημένο μοντέλο Ενισχυτικής Κλίσης ικανό να επεξεργαστεί δεδομένα μεγάλης κλίμακας τόσο κατανεμημένα, παράλληλα αλλά και αξιοποιώντας κάρτες γραφικών (GPUs), παρουσιάζοντας εξαιρετικά χαμηλούς χρόνος εκπαίδευσης με ταυτόχρονη χαμηλή χρήση μνήμης RAM. Ο αλγόριθμος διατηρεί βελτιστοποιήσεις που εισήγαγε ο προηγούμενος αλγόριθμος σχετικά με την αντιμετώπιση προβλημάτων υπερπροσαρμογής των μοντέλων καθώς και σχετικά με τη παραλληλοποίηση των βημάτων εκπαίδευσης και την αντιμετώπιση των ελλιπών τιμών. Αντίθετα όμως με τον αλγόριθμο XGBoost, ο συγκεκριμένος αλγόριθμος χρησιμοποιεί μια προσέγγιση Best-First (Leaf-Wise) για την ανάπτυξη των φύλλων των δέντρων απόφασης (Σχήμα 5.3).



Σχήμα 5.3: *Ανάπτυξη Δέντρων - Best-First (Leaf-Wise)*

Στη περίπτωση δημιουργίας δύο δέντρων απόφασης στο ίδιο σύνολο δεδομένων τότε οι δύο διαφορετικές μέθοδοι ανάπτυξης των φύλλων του εκάστοτε δέντρου, εάν τα δέντρα δεν περικοπούν ή χρησιμοποιηθεί κάποιο κριτήριο για πρόωρο τερματισμό της εκπαίδευσης, αναμένεται να καταλήξουν σε δέντρα ίδιας μορφής. Η διαφορά των δύο μεθόδων έγκειται λοιπόν στη σειρά με την οποία αναπτύσσονται τα δέντρα και συγκεκριμένα οι κόμβοι αυτών. Μέσω της μέθοδο Leaf-Wise τα επόμενα σημεία διαχωρισμού επιλέγονται συνυπολογίζοντας τη συνεισφοράς τους στη συνολική συνάρτηση απώλειας και όχι στην απώλεια για το συγκε-

κριμένο κλαδί. Ως αποτέλεσμα η συγκεκριμένη προσέγγιση καταλήγει συνήθως σε δέντρα χαμηλότερης απώλειας γρηγορότερα σε σχέση με τη προσέγγιση Level-Wise. Αντίστοιχα με τον XGBoost ορίζονται κατώφλια μέγιστου βάθους για τα δέντρα με σκοπό την αποφυγή υπερπροσαρμογής του αλγορίθμου.

Σχετικά με το καθορισμό των βέλτιστων σημείων διαχωρισμού, χρησιμοποιείται μέθοδος βασισμένη στο Ιστόγραμμα κατά την οποία τα συνεχή χαρακτηριστικά διαχωρίζονται σε διακριτά κομμάτια (discrete bins) τα οποία χρησιμοποιούνται έναντι των πραγματικών δεδομένων. Αξίζει να σημειωθεί ότι παρόμοιες μέθοδοι, για τη δημιουργία των δέντρων απόφασης, βασισμένοι στη μέθοδο του Ιστογράμματος υλοποιήθηκαν μεταγενέστερα και στα πλαίσια του αλγορίθμου XGBoost και είναι αυτές οι οποίες χρησιμοποιούνται κατά κόρον λόγο των υψηλών αποδόσεων που παρουσιάζουν. Το υπολογιστικό κόστος για τη κατασκευή των ιστογραμμάτων καθορίζεται ως O(#Δεδομένα $*$ #Χαρακτηριστικά) και O(#Κομματιών $*$ #Χαρακτηριστικά) για την εύρεση των σημείων διαχωρισμού. Συνεπώς, η υπολογιστική πολυπλοκότητα καθορίζεται κυρίως από το μέρος το οποίο αφορά τη κατασκευή ιστογραμμάτων καθώς το πλήθος των διακριτών κομματιών είναι σημαντικά μικρότερο από το πλήθος των δεδομένων. Σε αυτό το σημείο ο αλγόριθμος καινοτομεί προτείνοντας δύο προσεγγίσεις για τη μείωση της πολυπλοκότητας κατά τη κατασκευή των ιστογραμμάτων.

Η πρώτη προσέγγιση Gradient Based One Side Sampling (GOSS) έγκειται στη μείωση των δεδομένων μέσω της χρήσης υποδειγματοληψίας βασισμένη στους συντελεστές κλίσης. Βασιζόμενη στην αρχή ότι σημεία για τα οποία παρατηρούνται χαμηλοί συντελεστές κλίσης, κατά απόλυτη τιμή, αποτελούν ένα σύνολο για το οποίο το μοντέλο έχει εκπαιδευτεί επαρκώς, η μέθοδος πραγματοποιεί τυχαία υποδειγματοληψία στο προαναφερθέν σύνολο διατηρώντας σταθερό το σύνολο των σημείων με υψηλούς συντελεστές κλίσης για το οποίο το μοντέλο δεν έχει εκπαιδευτεί επαρκώς. Με αυτό το τρόπο επιτυγχάνεται η μείωση των δεδομένων επηρεάζοντας όσο το δυνατόν λιγότερο τη στατιστική κατανομή αυτών.

Συμπληρωματικά με τη προηγούμενη μέθοδο, η δεύτερη προσέγγιση Exclusive Feature Bundling (EFB) χρησιμοποιείται με σκοπό τη μείωση της διάστασης των χαρακτηριστικών του συνόλου εκπαίδευσης. Κάτι τέτοιο επιτυγχάνεται μέσω του εντοπισμού χαρακτηριστικών τα οποία είναι ικανά να συνενωθούν σε μια δέσμη χαρακτηριστικών, όντας αμοιβαία διαχωρίσιμα, διατηρώντας την αντιστρεψιμότητα της πράξης. Ο εντοπισμός χαρακτηριστικών τα οποία ικανοποιούν τη προαναφερθείσα συνθήκη πραγματοποιείται δημιουργώντας ένα γράφο με βάρη στα άκρα (edges) του. Τα βάρη καθορίζονται σύμφωνα με την αναλογία από αμοιβαία διαχωρίσιμα χαρακτηριστικά τα οποία παρουσιάζουν επικαλυπτόμενες τιμές. Στη συνέχεια τα χαρακτηριστικά ταξινομούνται σε φθίνουσα σειρά σύμφωνα με το πλήθος από μη μηδενικές παρατηρήσεις. Τέλος αρκεί να προσπελάσουμε τη ταξινομημένη λίστα από χαρακτηριστικά και να αναθέσουμε το εκάστοτε χαρακτηριστικό σε μια καινούργια ή προϋπάρχουσα δέσμη. Δοθέντος των χαρακτηριστικών τα οποία μπορούν να συνενωθούν πραγματοποιείται η πράξη της συνένωσης ως εξής:

1. Υπολόγισε τη ποσότητα αντιστάθμισης (offset) η οποία θα προστεθεί σε κάθε χαρακτηριστικό.

2. Αρχικοποίησε με μηδέν τις παρατηρήσεις των χαρακτηριστικών δέσμης για τα οποία οι αντίστοιχες παρατηρήσεις των αρχικών χαρακτηριστικών είναι όλες μηδέν.

3. Υπολόγισε τη νέα δέσμη για κάθε μη μηδενική παρατήρηση προσθέτοντας τη ποσότητα αντιστάθμισης στην αρχική δέσμη του χαρακτηριστικού.

Στο Πίνακα 5.1 παρατίθεται ένα απλό παράδειγμα στο οποίο παρουσιάζεται ο τρόπος λειτουργίας της μεθόδου. Τα χαρακτηριστικά Α, Β αποτελούν δύο αμοιβαίως διαχωρίσιμα χαρακτηριστικά τα οποία δύναται να συνδυαστούν (Χαρακτηριστικό Γ). Για να επιτύχουμε μη επικαλυπτόμενα buckets προσθέτουμε το μέγεθος της δέσμης του Χαρακτηριστικού Α στο Χαρακτηριστικό Β. Ως αποτέλεσμα το Χαρακτηριστικό Γ αντιπροσωπεύει τα μη μηδενικά στοιχεία του Χαρακτηριστικού Β στα buckets 1 έως 4 ενώ τα buckets 5 έως 6 αντιπροσωπεύουν τα μη μηδενικά στοιχεία του χαρακτηριστικού Α.

Χαρακτηριστικό Α	**Χαρακτηριστικό Β**	**Χαρακτηριστικό Γ**
0	1	1
0	4	4
0	3	3
0	2	2
2	0	6
2	0	6
1	0	5

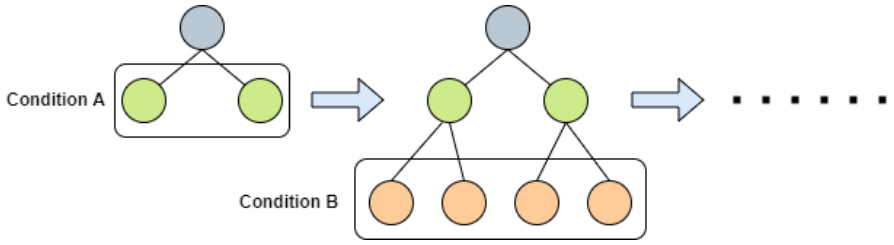
Πίνακας 5.1: *Παράδειγμα Λειτουργίας EFB*

Τέλος ο αλγόριθμός LightGBM είναι σε θέση να χειριστεί κατηγορικά δεδομένα, χωρίς να χρειαστεί αυτά να μετασχηματιστούν μέσω one-hot encoding. Αντιθέτως για την αποφυγή δημιουργίας αραιών χώρων για τους οποίους το δέντρο θα έπρεπε να χαρακτηρίζεται από αρκετά μεγάλο βάθος έτσι ώστε να επιτύχει ικανοποιητικά αποτελέσματα, επιλέγεται ο διαχωρισμός των κατηγοριών σε μόνο δύο υποσύνολα. Συνεπώς για μια μεταβλητή με k επίπεδα το βέλτιστο σημείο διαχωρισμού επιλέγεται από ένα σύνολο $2^{k-1} - 1$ δυνατών διαχωρισμών.

5.1.3 CatBoost

Ο αλγόριθμος CatBoost αναπτύχθηκε το 2017 από τους ερευνητές της εταιρείας Yandex [73]. Αποτελεί και αυτός με τη σειρά του ένα βελτιστοποιημένο ένα βελτιστοποιημένο μοντέλο Ενισχυτικής Κλίσης ικανό να επεξεργαστεί δεδομένα μεγάλης κλίμακας τόσο κατανεμημένα, παράλληλα αλλά και αξιοποιώντας κάρτες γραφικών (GPUs), με κύρια χαρακτηριστικά να αποτελούν την ικανότητα διαχείρισης κατηγορικών μεταβλητών στα πλαίσια του gradient boosting, χωρίς να προϋποθέτει τη μετατροπή αυτών σε κάποια αραιή απεικόνιση (π.χ. one-hot encoding) καθώς και τις αρκετά καλές επιδόσεις μέσω της χρήσης των προκαθορισμένων τιμών για τις υπερπαραμέτρους του.

Ο αλγόριθμος χρησιμοποιεί βελτιστοποιήσεις προγενέστερων μοντέλων διαφοροποιώντας όμως τη διαδικασία ανάπτυξης των φύλλων των δέντρων απόφασης (Σχήμα 5.4). Σε αντίθεση με τα προαναφερθείσα μοντέλα ο CatBoost αναπτύσσει συμμετρικά δεντρά απόφασης. Σε κάθε επίπεδο του κάθε δέντρου, το ίδιο χαρακτηριστικό μαζί με το επιλεχθέν κατώφλι, για τα οποία παρατηρείται το ελάχιστο σφάλμα στη συνάρτηση απώλειας, επιλέγεται για όλους τους κόμβους του συγκεκριμένου επιπέδου.



Σχήμα 5.4: *Ανάπτυξη Δέντρων - Συμμετρικά*

Για την κωδικοποίηση των επίπεδων των κατηγορικών μεταβλητών με λίγα επίπεδα χρησιμοποιείται one-hot encoding ενώ για τις υπόλοιπες με υψηλή πληθικότητα χρησιμοποιούνται Target Statistics (TS) τα οποία εκτιμούν τη μέση τιμή των τιμών στόχου για κάθε κατηγορία. Παρόλο αυτά η μέθοδος παρουσιάζει πρόβλημα target leakage καθώς τα νέα χαρακτηριστικά υπολογίζονται με βάση τα προηγούμενα με αποτέλεσμα η κατανομή του συνόλου εκπαίδευσης να είναι μετατοπισμένη σε σχέση με την κατανομή του συνόλου επικύρωσης. Για αυτό το λόγο και επιλέγεται η χρήση holdout TS ή leave-one out TS οι οποίες παρόλο αυτά δεν καταφέρνουν να λύσουν πλήρως το πρόβλημα. Παρόλο αυτά ο αλγόριθμος CatBoost χρησιμοποιεί μια αποτελεσματική στρατηγική, Ordered Target Statistics, η οποία βασίζεται στην επίδραση των επιπέδων στη μεταβλητή στόχο λαμβάνοντας υπόψη και μια prior για το εκάστοτε χαρακτηριστικό. Οι τιμές των Target Statistics για κάθε παράδειγμα βασίζονται στο παρατηρηθέν ιστορικό. Για να το επιτύχει αυτό εισάγουν τεχνητό χρόνο, τυχαία μετάθεση κατά σ του συνόλου εκπαίδευσης. Στη περίπτωση όπου χρησιμοποιηθεί μόνο μια τυχαία μετάθεση τότε τα Target Statistics παρουσιάζουν υψηλή διακύμανση με αποτέλεσμα να επιλέγονται διαφορετικές μεταθέσεις στα διαφορετικά βήματα του gradient boosting. Σε αυτή τη περίπτωση για αποφυγή προβλημάτων leakage εκπαιδεύονται πολλαπλά μοντέλα ταυτόχρονα τα οποία στη συνέχεια συνενώνονται.

Σε ότι αφορά τη βελτιστοποίηση για την εύρεση των σημείων διαχωρισμού, για τη μέθοδο του Ιστογράμματος, αντί του GOOS που χρησιμοποιεί ο LightGBM χρησιμοποιείται Δειγματοληψία Ελάχιστης Διακύμανσης (Minimal Variance Sampling - MVS). Η σταθμισμένη δειγματοληψία πραγματοποιείται στο επίπεδο των δένδρων και όχι στο επίπεδο των κόμβων, βοηθώντας στη μεγιστοποίησης της ακρίβειας των κόμβων.

Επιπροσθέτως, αντιμετωπίζονται προβλήματα υπερπροσαρμογής, σχετιζόμενα με το γεγονός ότι κατά την εκτίμηση των τιμών των φύλλων λαμβάνονται υπόψη όλα τα gradients που θα είναι στο συγκεκριμένο φύλλο. Η εκτίμηση αυτή δεν είναι αμερόληπτη καθώς εκτιμάμε τη τιμή των φύλλων χρησιμοποιώντας τα ίδια αντικείμενα με τα οποία δημιουργήσαμε το μοντέλο. Για την αντιμετώπιση του προαναφερθέν προβλήματος ο αλγόριθμός χρησιμοποιεί την ίδια τεχνική τυχαίας μετάθεσης, όπως αυτή περιγράφηκε στα πλαίσια των Ordered Target Statistics. Πιο συγκεκριμένα, για κάθε αντικείμενο η εκτίμηση των τιμών των φύλλων πραγματοποιείται με βάση το μοντέλο που δεν έχει εκπαιδευτεί στο παρών αντικείμενο.

Τέλος σε αντίθεση με τα άλλα δύο μοντέλα όπου οι ελλιπείς τιμές τοποθετούνται στο αντίστοιχο κλαδί με τρόπο τέτοιο ώστε να μειώνεται η απώλεια σε κάθε κόμβο, ο CatBoost αντιμετωπίζει τις ελλιπείς τιμές μέσω της χρήσης Αντικατάστασης Ελάχιστης ή Μέγιστης τιμής. Στη περίπτωση επιλογής της Μέγιστης τιμής οι ελλιπείς τιμές αντικαθιστώνται από

κάποια τιμή μεγαλύτερη από τη μέγιστη τιμή του εκάστοτε χαρακτηριστικού. Με αυτό το τρόπο έχουμε εγγυήσεις ότι, κατά την επιλογή των βέλτιστων σημείων διαχωρισμού, θα εξεταστεί και κάποιο σημείο διαχωρισμού το οποίο θα διαχωρίζει τις ελλιπείς τιμές από όλα τα υπόλοιπα σημεία του δείγματος. Επιλέγοντας τη λύση της Ελάχιστης τιμής η υλοποίηση παραμένει αντίστοιχη.

5.2 Ανάλυση Κυρίων Συνιστωσών - PCA

Η Ανάλυση Κυρίων Συνιστωσών [74], [75], [76] αποτελεί ένα γραμμικό μετασχηματισμό του συνόλου δεδομένων μέσω του οποίου αλλάζουμε το σύστημα συνταγμένων. Με αυτό το τρόπο καταφέρνουμε να παράξουμε ασυσχέτιστους γραμμικούς συνδυασμούς οι οποίοι θα περιέχουν όσο το δυνατό μεγαλύτερο μέρος της συνολικής διακύμανσης. Το παραπάνω γίνεται εφικτό μέσω της χρήσης φασματικής ανάλυσης για τον εκτιμώμενο, από το σύνολο των δεδομένων, Πίνακα Συνδιακύμανσης (ή Συσχέτισης). Πιο συγκεκριμένα αναλύουμε το Πίνακα Συνδιακύμανσης ($\Sigma_{(pxp)}$) στην εξής μορφή:

$$\Sigma_{(pxp)} = P\Lambda P'$$

όπου $\Lambda_{(pxp)}$ αποτελεί διαγώνιο πίνακα με τις ιδιοτιμές $\lambda_i \geq 0$ (Σ θετικά ορισμένος) σε φθίνουσα σειρά και P ορθογώνιος πίνακας ($P'P = I$) με τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Η διακύμανση των αρχικών μεταβλητών συσσωρεύεται στις πρώτες κύριες συνιστώσες γεγονός το οποίο καθιστά εφικτή την μείωση διαστάσεων καθώς και τον καθαρισμό των δεδομένων από θόρυβο. Διατηρώντας όλες τις Κύριες Συνιστώσες καταφέρνουμε να διατηρήσουμε τη συνολική διακύμανση των δεδομένων $tr(\Sigma) = tr(\Lambda) = \sum_{i=i}^{p} \lambda_i$.

Για την επιλογή του πλήθους των πρώτων k κυρίων συνιστωσών έχει προταθεί μια πληθώρα μεθόδων χωρίς παρόλο αυτά να υπάρχει κάποια μέθοδος η οποία να εγγυάται βέλτιστα αποτελέσματα. Ενδεικτικά αναφέρουμε μερικές από αυτές:

- ***Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες***
  Επιλογή *min* k τέτοιου ώστε $\Sigma \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$ μεγαλύτερο από όριο ερμηνευόμενης διακύμανσης
- ***Μέθοδος σπασμένου ραβδίου (Broken Stick)***
  Επιλογή k τέτοιου ώστε $\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} > g_k$ όπου $g_k = \frac{1}{p} \sum_{i=k}^{p} \left(\frac{1}{i}\right)$
- ***Προσομοίωση από την Εμπειρική Κατανομή (Bootstrap)*** [77]
  Υπολογισμός 95% ΔΕ για τα λ_i και σύγκριση αυτών με το $\overline{\lambda}$ με σκοπό την επιλογή του βέλτιστου αριθμού κυρίων συνιστωσών.
- ***Διασταυρούμενη Επικύρωση***

5.3 Μάθηση με μη Ισορροπημένα Δεδομένα

Σε πολλά προβλήματα ταξινόμησης τα δεδομένα είναι από τη φύση τους μη ισορροπημένα (Ανίχνευση Εισβολής, Ανίχνευση Ακραίων Στοιχείων, Ανίχνευση Απάτης, Ανίχνευση Ασθενειών κ.λπ). Το παρών λαμβάνει χώρα καθώς πολλές φορές είτε κοστίζει η εύρεση δεδομένων

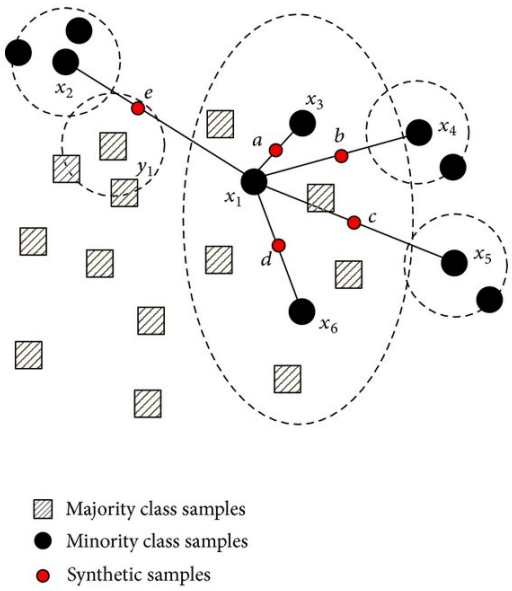
από κάποια κλάση είτε τα δεδομένα αυτά καθεαυτά είναι δυσεύρετα. Ως αποτέλεσμα τα μοντέλα μηχανικής μάθησης πολλές φορές μαθαίνουν αρκετά καλά τη πλειοψηφική κλάση και αποτυγχάνουν να μάθουν τη μειοψηφική, η οποία λόγο της ανισορροπίας ενδεχομένως αγνοείται. Σε μια προσπάθεια να ξεπεραστεί το προαναφερθέν ζήτημα έχουν προταθεί αρκετές μέθοδοι, τόσο στο επίπεδο των δεδομένων όσο και στο επίπεδο της εκπαίδευσης των μοντέλων [78]. Στη συνέχεια, παρουσιάζονται αναλυτικά οι μέθοδοι που χρησιμοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας.

5.3.1 Τυχαία Υποδειγματοληψία Πλειοψηφίας

Η πιο απλή μέθοδος για την εξουδετέρωση της ανισορροπίας των κλάσεων είναι η τυχαία υποδειγματοληψία. Σε αυτή τη περίπτωση αφαιρούμε τυχαία παρατηρήσεις από τη πλειοψηφική κλάση μέχρις ότου να επιτευχθεί η επιθυμητή αναλογία μεταξύ των κλάσεων.

5.3.2 Τεχνική Συνθετικής Υπερδειγματοληψίας Μειονότητας - SMOTE

Στο Chawla et al.[79] οι συγγραφείς του άρθρου προτείνουν μια μέθοδο δειγματοληψίας με σκοπό την αντιμετώπιση προβλημάτων μη ισορροπημένων συνόλων δεδομένων. Σε αντίθεση με άλλες μεθόδους υπερδειγματοληψίας η παρούσα τεχνική δεν παράγει δείγματα με επανάθεση αλλά δημιουργεί καινούργιες παρατηρήσεις από τη μειοψηφική κλάση, οι οποίες διαφέρουν ελάχιστα από τα αρχικά δεδομένα. Στο Σχήμα 5.5 παραθέτουμε μια γραφική απεικόνιση του μηχανισμού που χρησιμοποιείται για τη παραγωγή συνθετικών δειγμάτων. Τα συνθετικά δείγματα τοποθετούνται τυχαία κατά μήκος κάθε γραμμής.



Σχήμα 5.5: *Λειτουργία της μεθόδου SMOTE*

Πιο συγκεκριμένα για τη παραγωγή μιας συνθετικής παρατήρησης ακολουθούνται τα ακόλουθα βήματα:

1. Τυχαία δειγματοληψία δείγματος μεγέθους n από τη μειοψηφική κλάση

2. Υπολογισμός των κ-κοντινότερων γειτόνων για το τυχαίο δείγμα
3. Υπολογισμός ενός διανύσματος μεταξύ μιας επιλεγμένης παρατήρησης και μιας επιλεγμένης γειτονιάς
4. Το διάνυσμα πολλαπλασιάζεται με ένα τυχαίο αριθμό στο εύρος από μηδέν έως ένα.
5. Για τη παραγωγή της συνθετικής παρατήρησης, αθροίζουμε τη παραπάνω τιμή με την επιλεγμένη παρατήρηση

Στα πλαίσια της παρούσας διπλωματικής η μέθοδος Τεχνικής Υπερδειγματοληψίας της Συνθετικής Μειονότητας χρησιμοποιήθηκε τόσο σε συνδυασμό με τη Τυχαία Υποδειγματοληψία όσο και με την Ανάλυση σε Κύριες Συνιστώσες.

Εντούτοις αξίζει να σημειωθεί ότι η αποτελεσματικότητα της μεθόδου έχει αμφισβητηθεί τόσο για σύνολα δεδομένων υψηλής διάστασης [80] όσο και σε σχέση με την απόδοση SOTA μοντέλων [81], όπως τα XGBoost, LightGBM, CatBoost, γεγονός το οποίο μας οδηγεί στην αναζήτηση αντίστοιχων τεχνικών στο χώρο της Βαθιάς Μάθησης.

5.3.3 Τεχνικές Υπερδειγματοληψίας βασισμένες σε Βαθιά Μάθηση

Οι τεχνικές υπερδειγματοληψίας βασισμένες σε βαθιά μάθηση βασίζονται στη χρήση βαθιών παραγωγικών μοντέλων μέσω των οποίων καθίσταται δυνατή η εκπαίδευση μοντέλων με την ικανότητα να παράγουν καινούργια συνθετικά δεδομένα σύμφωνα με τις ιδιότητες του εκάστοτε δείγματος εκπαίδευσης. Στα πλαίσια της παρούσας διπλωματικής εργασίας έμφαση δόθηκε στη χρήση τέτοιου είδος μοντέλων με σκοπό τη μοντελοποίηση δεδομένα μορφής πίνακα. Δύο κύριες κατηγορίες αυτών των μοντέλων αποτελούν τα Παραγωγικά Ανταγωνιστικά Δίκτυα (Generative Adversarial Networks, GANs) και οι Μεταβλητοί Αυτοκωδικοποιητές (VAE) οι οποίοι όμως δεν χρησιμοποιήθηκαν στα πλαίσια αυτής της διπλωματικής εργασίας.

Προτού όμως περιγράψουμε το μοντέλο το οποίο χρησιμοποιήθηκε, για λόγους πληρότητας, παραθέτουμε μια σύντομη περιγραφή για τα Παραγωγικά Ανταγωνιστικά Δίκτυα.

5.3.3.1 Παραγωγικά Ανταγωνιστικά Δίκτυα

Τα Παραγωγικά Ανταγωνιστικά Δίκτυα (GANs) αποτελούνται από δύο ευδιάκριτα μέρη τον Παράγωγο (Generator) και το Διευκρινιστή (Discriminator). Ο Παράγωγος λαμβάνει ως είσοδο θόρυβο και παράγει δείγματα της μορφής $x = G(z; \theta_g)$ ενώ ο Διευκρινιστής προσπαθεί να ξεχωρίσει, με όσο το δυνατόν μεγαλύτερη ακρίβεια, πραγματικά δεδομένα και δεδομένα παραχθέντα από τον Παράγωγο μέσω της χρήσης της πιθανότητας $D(x; \theta_d)$ [58]. Η βελτιστοποίηση της προαναφερθείσας διαδικασίας επιτυγχάνεται όταν ο Παράγωγος καταφέρει να μάθει την κατανομή των πραγματικών δεδομένων με αποτέλεσμα τα συνθετικά δεδομένα να μην διαφέρουν σημαντικά από τα πραγματικά και ο Διευκρινιστής να μην καταφέρνει την επιτυχή διάκριση ανάμεσα στις δύο κλάσεις. Σε μαθηματικούς όρους θα μπορούσαμε να ορίσουμε την προαναφερθείσα διαδικασία ως τη μεγιστοποίηση από το Παράγωγο της πιθανότητας ο Διευκρινιστής να κάνει λάθος:

$$\min_G \max_D J\left(\theta_g, \theta_d\right),$$

όπου $J\left(\theta_g, \theta_d\right) = \mathbb{E}_{x\sim p_{\text{data}}\,(x)}\left[\log D\left(x;\theta_d\right)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\left(G\left(z;\theta_g\right)\theta_d\right)\right)\right]$ το οποίο μπορεί να γραφτεί στη γενικότερη μορφή, αγνοώντας τις παραμέτρους θ_g, θ_d, ως εξής:

$$J(G, D) = \mathbb{E}_{x\sim p_{\text{data}}\,(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

Τα GANs αποτελούν ένα μη συνεργατικό παιχνίδι μηδενικού αθροίσματος (zero-sum non-cooperative game), συνεπώς στη περίπτωση όπου ένα εκ των δύο κερδίσει το άλλο χάνει. Αποτέλεσμα αυτού είναι η επιδίωξη μιας η ιδανικής κατάστασης εκπαίδευσης μέσω της ισορροπίας κάτι το οποίο κατά την εκπαίδευση των GANs επιτυγχάνεται όταν τα δύο αντίπαλα δίκτυα φτάνουν στην Ισορροπία του Νας (Nash Equilibrium). Σύμφωνα με τον Goodfellow [82] το σημείο ισορροπίας επιτυγχάνεται εάν και μόνο εάν έχουμε $p_g(x) = p(x)$ με τη τιμή εξόδου του Διευκρινιστή να είναι το 0.5 ανεξάρτητα της εισόδου. Ο βέλτιστος Διευκρινιστής για γνωστό Παράγωγο δίνεται από τον εξής τύπο:

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)}$$

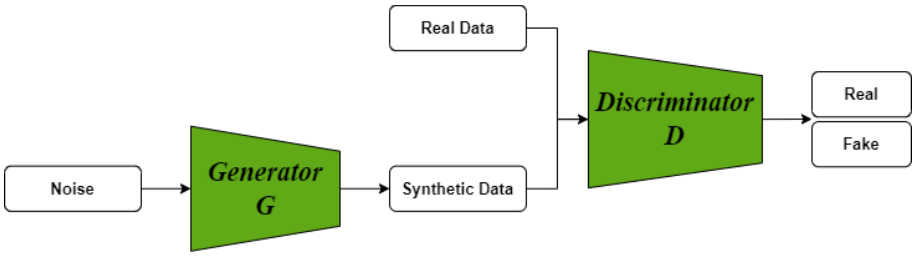
Η εκπαίδευση των GANs μπορεί να καταλήξει μια εξαιρετικά πολύπλοκη διαδικασία καθώς τα εκπαιδευμένα μοντέλα είναι αρκετά ασταθή. Ένα από τα προβλήματα το οποίο εμφανίζεται κατά την εκπαίδευση είναι αυτό της Κατάρρευσης Λειτουργίας (Mode Collapse). Πιο συγκεκριμένα, ο Παράγωγος παρόλο που καταφέρνει να παράξει δεδομένα ικανά να μπερδέψουν το Διευκρινιστή ως αληθινά, δεν καταφέρνει να μάθει ικανοποιητικά την κατανομή των δεδομένων με αποτέλεσμα οι παραγόμενες έξοδοι να παρουσιάζουν χαμηλή μεταβλητότητα καθώς και χαμηλή ποικιλομορφία. Επιπροσθέτως, η σύγκλιση για το μοντέλο είναι πολύ πιθανόν να μην είναι σταθερή καθώς η απώλεια του Παραγώγου βελτιώνεται όταν βελτιώνεται η απώλεια του Διευκρινιστή, και αντίστροφα. Ως αποτέλεσμα αυτού η χρήση της συνάρτησης απώλειας δεν είναι ασφαλές να χρησιμοποιηθεί με σκοπό την εξαγωγή συμπερασμάτων σχετικά με τη σύγκλιση ή όχι του μοντέλου. Σε περιπτώσεις όπου ο Παράγωγος επιτύχει την πλήρη εκμάθηση της κατανομής των πραγματικών δεδομένων, οδηγεί τον Διευκρινιστή να επιλέγει μεταξύ συνθετικών ή όχι παρατηρήσεων με πιθανότητα 0.5. Συνεπώς είναι επιτακτικό η εκπαίδευση του μοντέλου να διακοπεί σε εκείνο το χρονικό σημείο καθώς πέραν αυτού η ανατροφοδότηση προς το Παράγωγο είναι στη πραγματικότητα τυχαία με αποτέλεσμα να αυξάνεται η πιθανότητα κατάρρευσης του. Επιπλέον ο Διευκρινιστής βάση ορισμού ανατροφοδοτεί τον Παράγωγο με τρόπο τέτοιο ώστε να βελτιώνεται και αυτός παρόλο αυτά η παραπάνω λειτουργία δύναται να δημιουργήσει ορισμένα προβλήματα σε περιπτώσεις όπου ο Διευκρινιστής είναι υπερβολικά καλός σε σχέση με το Παράγωγο. Πιο συγκεκριμένα, σε τέτοιου είδους περιπτώσεις ο Παράγωγος ανατροφοδοτείται με κλίσης οι οποίες τείνουν προς το μηδέν με αποτέλεσμα η εκ τον Διευκρινιστή πληροφορία να μην είναι χρήσιμη για το Παράγωγο και ο τελευταίος να μην σημειώνει πρόοδο κατά την εκπαίδευση του. Το προαναφερθέν δεν είναι άλλο από το γνωστό πρόβλημα, το οποίο παρουσιάζουν συχνά τα Νευρωνικά Δίκτυα, της εξαφάνισης των διανυσμάτων κλίσης (Vanishing Gradients Problem).

5.3.3.1.1 Είδη GANs

Ως παραγωγικά μοντέλα, τα GANs προτάθηκαν πρόσφατα από τους ερευνητές παρόλο αυτά με τη πάροδο των χρόνων νέοι και βελτιωμένοι τρόποι εκπαίδευσης καθώς και αρχιτεκτονικές προτάθηκαν από τους ερευνητές. Σε αυτό το σημείο παραθέτουμε μια σύντομη αναφορά σε αυτά τα οποία σχετίζονται με τη παρούσα διπλωματική εργασία.

Vanilla GANs

Αρχικά ξεκινώντας από τα απλά (GANs) αυτά αποτελούνται από τα Vanilla GANs και τα Wasserstein GANs (WGANs). Η βασική αρχιτεκτονική δεν διαφοροποιείται μεταξύ των δύο ειδών GANs (Σχήμα 5.6). Η πρώτη κατηγορία των Vanilla GANs αποτελούν στη πραγματικότητα τα GANs στα οποία αναφερόμασταν παραπάνω. Η διαφορά των Wasserstein GANs έγκειται στο είδος της συνάρτησης απώλειας την οποία χρησιμοποιούν.



Σχήμα 5.6: *Βασική Αρχιτεκτονική - GANs*

Wasserstein GANs

Αντίθετα με τα Vanilla GANs τα οποία χρησιμοποιούν συναρτήσεις (Kullback-Leibler diverenge, Jensen–Shannon diverenge) της οικογένειας των $\phi-$αποκλίσεων (f-divergence), οι οποίες ποσοτικοποιούν τη διαφορά μεταξύ δύο κατανομών πιθανότητας, τα Wasserstein GANs (WGANs) χρησιμοποιούν μεθόδους μετρικής ακέραιας πιθανότητας (IPM) οι οποίες, καθώς είναι σχεδιασμένες με σκοπό την αντιμετώπιση προβλημάτων τα οποία εμφανίζουν οι συναρτήσεις της φ-οικογένειας, παρουσιάζουν μια συνεπή απόσταση μεταξύ των κατανομών. Μια εξ αυτών είναι και η αποκαλούμενη ως Wasserstein-1 ή απόσταση Earth Mover's Distance (EMD) η οποία ορίζεται ως εξής:

$$W\left(p_r, p_g\right) = \inf_{\gamma \in \Pi\left(p_r, p_g\right)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|],$$

όπου $\|x - y\|$ αποτελεί τη συνάρτηση σφάλματος, τα p_r, p_g συμβολίζουν τις κατανομές πιθανότητας και $\Pi\left(p_r, p_g\right)$ αποτελεί το σύνολο όλων των από κοινού κατανομών $\gamma(x, y)$. Για αυτόν το λόγο οι συγγραφείς του [83] εφάρμοσαν την Kantorovich-Rubinstein duality με αποτέλεσμα η απόσταση EMD να λαμβάνει την ακόλουθη μορφή:

$$\min_G \max_{w \in W} \mathbb{E}_{x\sim P_{\text{data}}}[f(x; w)] - \mathbb{E}_{z\sim p_z}\left[f\left(G\left(z; \theta_g\right); w\right)\right]$$

Επιπλέον η σιγμοειδής συνάρτηση δεν εφαρμόζεται στο τελευταίο τμήμα του Διευκρινιστή με

αποτέλεσμα η συνάρτηση EMD να εφαρμόζεται απευθείας στα logits παράγοντας ένα σκορ το οποίο ποσοτικοποιεί το πόσο κάθε παρατήρηση είναι πραγματική ή συνθετική. Η συνάρτηση Lipschitz χρησιμοποιείται με σκοπό τον περιορισμό του προβλήματος βελτιστοποίησης χρησιμοποιώντας weight clipping για τα βάρη του Διευκρινιστή ενώ συνολικά χρησιμοποιείται ο βελτιστοποιητής RMSProp.

Το μοντέλο WGAN παρουσιάζει σταθερότητα κατά την διαδικασία εκπαίδευσης. Παρόλο αυτά κάτι το οποίο το ξεχωρίζει είναι η ανθεκτικότητα στην επιλογή υπερπαραμέτρων και αρχιτεκτονικής. Επιπροσθέτως φαινόμενα Mode Collapse εμφανίζονται με χαμηλότερη πιθανότητα. Το σημαντικότερο όμως πλεονέκτημα είναι η συνεχής, κατά την εκπαίδευση, εκτίμηση της EMD εκπαιδεύοντας τον Διευκρινιστή μέχρις ότου λάβει τις βέλτιστες τιμές. Αντιθέτως η χρήση weight clipping είναι πιθανόν να προκαλέσει προβλήματα καθώς για μεγάλες τιμές της παραμέτρου clipping η εκπαίδευση του μοντέλου αυξάνεται σημαντικά.

Wasserstein GANs & Gradient Penalty

Για αυτό το λόγο το 2017 [84] εισήγαγαν την τεχνική gradient penalty για να επιτύχουν το περιορισμό Lipschitz έναντι του weight clipping. Πιο συγκεκριμένα, το μοντέλο τιμωρείται όταν η νόρμα των κλίσεων μετατοπίζεται μακριά από τη τιμή στόχο της μονάδας. Επειδή το gradient penalty εφαρμόζεται σε κάθε δείγμα, δεν είναι εφικτή η χρησιμοποίηση τεχνικών κανονικοποίησης συστάδας (batch normalization) ανταυτου τεχνικές κανονικοποίησης επιπέδου είναι δυνατόν να χρησιμοποιηθούν καθώς δεν συσχετίζουν τα επιμέρους δείγματα.

Η συνάρτηση απώλειας με εφαρμοσμένα την απόσταση Wasserstein καθώς και gradient penalty ορίζεται με βάση την ακόλουθη συνάρτηση

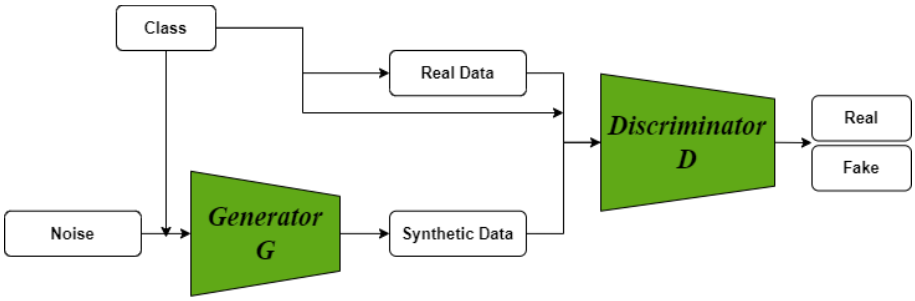
$$\max_{w \sim W} \mathbb{E}_{x \sim P_{\text{data}}} [f(x; w)] - \mathbb{E}_{z \sim p_z} \left[f\left(G\left(z; \partial_g\right); w\right)\right] + \lambda \mathbb{E}_{Z \sim p_z} \left[\left(\left\| \nabla_w f\left(G\left(z; \partial_g\right); w\right) \right\|_2 - 1 \right)^2 \right],$$

όπου λ αποτελεί το συντελεστή ποινικοποίησης.

Conditional GANs

Το 2014 οι Mehdi Mirza και Simon Osindero πρότειναν τη χρήση Παραγωγικών Ανταγωνιστικών Μοντέλων Υπό Όρους (Conditional GANs) [85] με σκοπό τη χρήση GANs για παραγωγή συνθετικών δεδομένων με συγκεκριμένα επιθυμητά χαρακτηριστικά. Αυτή η ιδιότητα των μοντέλων να χειρίζονται το λανθάνων χώρο με τρόπο τέτοιο ώστε να είναι σε θέση να παράγουν δεδομένα με συγκεκριμένα χαρακτηριστικά κάνει τα μοντέλα να ξεχωρίζουν λόγο της χρησιμότητας του. Κάτι τέτοιο επιτυγχάνεται μέσω της χρήσης βοηθητικής πληροφορίας, όπως ετικέτα κλάσης, επίπεδο κατηγορικής μεταβλητής, η οποία δίδεται τόσο στο Παράγωγο όσο και στο Διευκρινιστή (Σχήμα 5.7).

Πιο συγκεκριμένα, κατά τη διαδικασία εκπαίδευσης ο Παράγωγος μαθαίνει να παράγει δεδομένα για κάθε ετικέτα, στο σύνολο εκπαίδευσης, ενώ ο Διευκρινιστής μαθαίνει να διακρίνει ανάμεσα στα συνθετικά και αληθινά ζεύγη δεδομένων και ετικετών. Ο Διευκρινιστής δεν θεωρεί συνθετικά μόνο τα ζεύγη συνθετικών δεδομένων αλλά και τα πραγματικά δεδομένα για τα οποία δεν έχει γίνει κατάλληλη αντιστοίχηση με την ετικέτα. Συνεπώς για να



Σχήμα 5.7: *Βασική Αρχιτεκτονική - Conditional GANs*

κερδίσει ο Παράγωγος δεν αρκεί μόνο να παράξει αληθοφανή δεδομένα αλλά και να μπορεί να παράγει δεδομένα από τις σωστές ετικέτες. Η συνάρτηση απώλειας για το ανταγωνιστικό παιχνίδι στα (Conditional GANs) περιγράφεται ως εξής:

$$\min_G \max_D J(G,D) = \mathbb{E}_{x\sim p_{\delta\alpha\tau\alpha}(x)}[\log D(x \mid y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(x \mid y)))]$$

Τα GANs έχουν δείξει ότι έχουν εξαιρετικές δυνατότητες στη παραγωγή συνθετικών δεδομένων εικόνας και ήχου παρόλο αυτά η παραγωγή συνθετικών δεδομένων μορφής πίνακα εισάγει ορισμένες δυσκολίες λόγω της ύπαρξης μεικτών ειδών δεδομένων. Επιπλέον οι κατανομές των δεδομένων ανά στήλη συνήθως διαφέρουν ανά μεταβλητή και περιγράφονται από πολυτροπικές (multimodal) κατανομές οι οποίες δεν προέρχονται από Κανονικούς πληθυσμούς. Επιπροσθέτως λόγω της ύπαρξης κατηγορικών μεταβλητών καταλήγουμε σε αραιούς και μη ισορροπημένους χώρους, ως προς τα επίπεδα των κατηγορικών μεταβλητών. Ένα από τα πιο πρόσφατα μοντέλα το οποίο προσπαθεί να αντιμετωπίσει όλα τα παραπάνω προβλήματα είναι το Conditional Tabular GAN [86] το οποίο προτάθηκε το 2019 από τους Xu et al.

5.3.3.1.2 Conditional Tabular GAN

Το Conditional Tabular GAN (CTGAN) αποτελεί μια αρχιτεκτονική η οποία σχεδιάστηκε με σκοπό τη παραγωγή συνθετικών δεδομένων από δεδομένα μορφής πίνακα [86]. Μέσω αυτής της αρχιτεκτονικής οι ερευνητές προσπαθούν να αντιμετωπίσουν προβλήματα σχετιζόμενα με τη μοντελοποίηση δεδομένων μορφής πίνακα.

Αρχικά προτείνεται ένα πλαίσιο προ-επεξεργασίας, κατά το οποίο πραγματοποιείται κανονικοποίηση με βάση τις επικρατούσες τιμές της εκάστοτε δειγματικής κατανομής τέτοια ώστε κάθε συνεχή μεταβλητή ανεξαρτήτου κατανομής να μετασχηματίζεται σε ένα διάνυσμα του οποίου οι τιμές αποτελούν μέλη ενός φραγμένου συνόλου. Πιο συγκεκριμένα σε κάθε συνεχής στήλη εφαρμόζεται ανεξάρτητα το μοντέλο Variational Gaussian Mixture [87]. Επιπλέον χρησιμοποιείτε ένας υπό όρους Παράγωγος και εκπαίδευση μέσω δειγματοληψίας με σκοπό την αντιμετώπιση προβλημάτων μη ισορροπίας στα επίπεδα των κατηγορικών μεταβλητών. Η δειγματοληψία χρησιμοποιείται έτσι ώστε να εξασφαλιστεί ότι όλα τα επίπεδα των κατηγορικών μεταβλητών εμφανίζονται ισοκατανεμημένα κατά την διαδικασία της εκπαίδευσης. Ο υπό όρους Παράγωγος λαμβάνει ως δεδομένα εισόδου θόρυβο σε συνδυασμό με ένα διάνυσμα το οποίο αποτελεί προϊόν one-hot-encoding και βοηθάει στην υπό όρους

λειτουργία του μοντέλου. Η έξοδος του υπό όρους Παραγώγου αξιολογείται από το Διευκρινιστή μέσω του υπολογισμού της απόστασης μεταξύ της πραγματικής και της εκτιμώμενης υπό όρους κατανομής. Στη συνέχεια παραθέτουμε την δομή του Παραγώγου (Σχήμα 5.8).

$$\begin{cases} h_0 = z \oplus cond \\ h_1 = h_0 \oplus \texttt{ReLU}(\texttt{BN}(\texttt{FC}_{|cond|+|z|\to 256}(h_0))) \\ h_2 = h_1 \oplus \texttt{ReLU}(\texttt{BN}(\texttt{FC}_{|cond|+|z|+256\to 256}(h_1))) \\ \hat{\alpha}_i = \texttt{tanh}(\texttt{FC}_{|cond|+|z|+512\to 1}(h_2)) & 1 \le i \le N_c \\ \hat{\beta}_i = \texttt{gumbel}_{0.2}(\texttt{FC}_{|cond|+|z|+512\to m_i}(h_2)) & 1 \le i \le N_c \\ \hat{\mathbf{d}}_i = \texttt{gumbel}_{0.2}(\texttt{FC}_{|cond|+|z|+512\to |D_i|}(h_2)) & 1 \le i \le N_d \end{cases}$$

Σχήμα 5.8: *Αρχιτεκτονική Παραγώγου - CTGAN*

Καθώς και του Διευκρινιστή (Σχήμα 5.9) ο οποίος χρησιμοποιεί τη δομή ενός PacGAN [88] με 10 δείγματα σε κάθε pac με σκοπό την αποφυγή προβλημάτων Κατάρρευσης Λειτουργίας.

$$\begin{cases} h_0 = \mathbf{r}_1 \oplus \ldots \oplus \mathbf{r}_{10} \oplus cond_1 \oplus \ldots \oplus cond_{10} \\ h_1 = \texttt{drop}(\texttt{leaky}_{0.2}(\texttt{FC}_{10|\mathbf{r}|+10|cond|\to 256}(h_0))) \\ h_2 = \texttt{drop}(\texttt{leaky}_{0.2}(\texttt{FC}_{256\to 256}(h_1))) \\ \mathcal{C}(\cdot) = \texttt{FC}_{256\to 1}(h_2) \end{cases}$$

Σχήμα 5.9: *Αρχιτεκτονική Διευκρινιστή - CTGAN*

Τέλος αξίζει να σημειωθεί ότι για την εκπαίδευση του μοντέλου χρησιμοποιείται ο βελτιστοποιητής Adam με ρυθμό εκπαίδευσης ίσο με $2*10^{-4}$ και η απώλεια WGAN προσαυξημένη με το gradient penalty.

5.3.4 Μάθηση με ευαισθησία στο κόστος (Cost-Sensitive Learning)

Η μάθηση με ευαισθησία στο κόστος αποτελεί υποπεριοχή της μηχανικής μάθησης η οποία στοχεύει στην αντιμετώπισή προβλημάτων ταξινόμησης σε μη ισορροπημένα σύνολα δεδομένων [89], [90].

Στα πλαίσια του Cost-Sensitive Learning ορίζουμε ένα Πίνακα Κόστους (Πίνακας 5.2) ο οποίος στα πλαίσια της δυαδικής ταξινόμησης αποτελεί ένα $2x2$ πίνακα διπλής εισόδου όπου το κελί (i,j) συμβολίζει το κόστος ταξινόμησης μιας παρατήρησης από τη κλάση i στη κλάση j για $i,j = 0, 1$ [89].

	Y_0	Y_1
$\hat{Y}_0$	c_{00}	c_{01}
$\hat{Y}_1$	c_{10}	c_{11}

Πίνακας 5.2: *Πίνακας Κόστους*

Είναι προφανές ότι για παρατηρήσεις οι οποίες ταξινομούνται στις σωστές κλάσης το

κόστος είναι μηδενικό $c_{00} = c_{01} = 0$ ενώ για τις λάθος ταξινομήσεις το κόστος είναι μη μηδενικό. Πρακτικά η εκτίμηση αυτόν τον τιμών είναι αρκετά δύσκολή καθώς τόσο οι εσφαλμένες θετικές ταξινομήσεις όσο και οι εσφαλμένες αρνητικές ταξινομήσεις συμβάλλουν αρνητικά στα αποτελέσματα του εκάστοτε προβλήματος. Αντ' αυτού χρησιμοποιούνται ορισμένες ευριστικές με την πιο κοινή από αυτές, για μη ισορροπημένα σύνολα δεδομένων, να είναι η αναλογία μη ισορροπίας όπως αυτή ορίζεται παρακάτω.

Έστω X το σύνολο δεδομένων εκπαίδευσης και $|X_1|$, $|X_0|$ η πληθικότητα των υποσυνόλων δεδομένων από τη μειοψηφική και πλειοψηφική κλάση αντίστοιχα, τότε η αναλογία μη ισορροπίας ορίζεται ως εξής:

$$IR = \frac{|X_1|}{|X_0|},$$

Σε αυτή τη περίπτωση, ο τελικός Πίνακας Κόστους έχει μορφή όπως αυτή περιγράφεται στο Πίνακα 5.3.

	Y_0	Y_1
$\hat{Y}_0$	0	1
$\hat{Y}_1$	IR	0

Πίνακας 5.3: *Εκτιμώμενος Πίνακας Κόστους*

Χρησιμοποιώντας τα παραπάνω κόστη παραμετροποιούμε ανάλογα τις συναρτήσεις απώλειας των εκάστοτε υπό εκπαίδευση μοντέλων με αποτέλεσμα να "τιμωρούνται" σημαντικά περισσότερο οι περιπτώσεις κατά οποίες κάποια παρατήρηση από τη μειοψηφική κλάση ταξινομείται εσφαλμένα. Με αυτό το τρόπο το όριο απόφασης απομακρύνεται από τις παρατηρήσεις της μειοψηφικής κλάσης προσφέροντας στο εκάστοτε μοντέλο καλύτερες ιδιότητες γενίκευσης. Τέλος αξίζει να σημειωθεί ότι παρόλο που η παραπάνω εκτίμηση του Πίνακα Κόστους αποτελεί μια αρκετά καλή αφετηρία, στη πράξη έχει ορισμένους περιορισμούς κυρίως σε προβλήματα όπου τα σύνολα δεδομένα είναι είτε αρκετά μικρά είτε οι κλάσεις είναι επικαλυπτόμενες. Για αυτό το λόγο συνίσταται ο χειρισμός των βαρών ως υπερπαραμέτρους, λαμβάνοντας πάντα υπόψη τις προαναφερθείσες εκτιμήσεις.

5.4 Επιλογή Υπερπαραμέτρων

Η επιλογή βέλτιστων υπερπαραμέτρων καθίσταται αναγκαία στο πλαίσιο της δημιουργίας ταξινομητών βασισμένους σε μη ισορροπημένα δεδομένα. Εξαιτίας της ιδιαιτερότητας των δεδομένων πολλές σύνηθες επιλογές για τις υπερπαραμέτρους των εκάστοτε μοντέλων αποτυγχάνουν να παράξουν ικανοποιητικά αποτελέσματα. Στα πλαίσια της παρούσας διπλωματικής επιλέχθηκε ο καθορισμός των υπερπαραμέτρων μέσω της χρήσης αλγόριθμων Μπεϋζιανής Βελτιστοποίησης [91], [92]. Κύριο λόγο για την επιλογή αυτών των αλγορίθμων αποτελεί τόσο η δυνατότητα βελτιστοποίησης ακριβών αντικειμενικών συναρτήσεων με κόστος σημαντικά μικρότερο σε σχέση με άλλες επιλογές (π.χ. Διασταυρώμενη Επικύρωση, Γεννητικοί Αλγόριθμοι) όσο και η βελτιστοποίηση μη κυρτών αντικειμενικών συναρτήσεων [93].

5.4.1 Μπεϋζιανοί Αλγόριθμοι Βελτιστοποίησης

Το πρόβλημα εύρεσης των βέλτιστων υπερπαραμέτρων στα πλαίσια της μηχανικής μάθησης αποτελεί τη προσπάθεια εύρεσης των υπερπαραμέτρων για τις οποίες ελαχιστοποιούνται οι απώλειες του μοντέλου κάτι το οποίο θα μπορούσε να εκφραστεί σε μαθηματικούς όρους ως εξής:

$$x^* = \arg\min_{x \in X} f(x),$$

όπου η $f(x)$ αποτελεί μια αντικειμενική συνάρτηση ή μετρική την οποία θέλουμε να ελαχιστοποιήσουμε γύρω από ένα σύνολο πιθανών υπερπαραμέτρων X. Οι Μπεϋζιανοί Αλγόριθμοι Βελτιστοποίησης σε αντίθεση με άλλες μεθόδους συνυπολογίζουν την επίδραση προηγούμενων επιλεγέντων συνόλων υπερπαραμέτρων στη βελτιστοποίηση της αντικειμενικής συνάρτησης και τις χρησιμοποιούν με σκοπό τη δημιουργία ενός πιθανοθεωρητικού μοντέλου για την αντικειμενική συνάρτηση. Τέτοια πιθανοθεωρητικά μοντέλα αποκαλούνται στη βιβλιογραφία Μοντέλα Υποκατάστασης και συμβολίζονται ως $p(y|x)$. Σε γενικές γραμμές οι αλγόριθμοι αυτοί λειτουργούν ως εξής:

1. Δημιουργία μοντέλου Υποκατάστασης για την αντικειμενική συνάρτηση
2. Εύρεση βέλτιστων υπερπαραμέτρων με βάση την απόδοση τους στο μοντέλο Υποκατάστασης.
3. Εφαρμογή των επιλεγμένων υπερπαραμέτρων στη πραγματική αντικειμενική συνάρτηση και ανανέωση του μοντέλου Υποκατάστασης με βάση τα αποτελέσματα.
4. Επανάληψη των βημάτων 2-3 για ένα προκαθορισμένο αριθμό επαναλήψεων.

Η παραπάνω διαδικασία γίνεται εφικτή μέσω της χρήσης μοντέλων Διαδοχικής Βελτιστοποίησης, Sequential Model Based Optimization (SMBO). Η αντικειμενική συνάρτηση προσεγγίζεται από ένα μοντέλο Υποκατάστασης το οποίο βελτιστοποιείται με βάση κάποιο κριτήριο, καθώς η βελτιστοποίηση τους είναι φθηνότερη. Η αντικειμενική συνάρτηση υπολογίζεται στη συνέχεια στο βέλτιστο σημείο x^* και μοντέλο $\mathcal{M}$ ανανεώνεται δοθέντος του τρέχοντος ιστορικού $\mathcal{H}$ (Αλγόριθμος 5.2).

Αλγοριθμος 5.2: *Ψευδο-κώδικας γενικής δομής του SMBO*

Είσοδος: f, M_0, T, S
1: $\mathcal{H} \leftarrow \emptyset$
2: Για $t = 1$ έως T:
3: $\quad x^* \leftarrow \underset{x}{\text{argmin}} \quad \mathrm{S}(\mathrm{x}, M_{t-1})$
4: $\quad$ Αποτίμησε τη $f(x^*)$
5: $\quad \mathcal{H} \leftarrow \mathcal{H} \cup (x^*, f(x^*))$
6: $\quad$ Προσάρμοσε νέο μοντέλο M_t στο $\mathcal{H}$
7: Επέστρεψε το $\mathcal{H}$

Ως προς βελτιστοποίηση κριτήριο για την επιλογή του επόμενου συνόλου από υπερπαραμέτρους ορίζεται το Expected Improvement (EI) [94]. Το EI αποτελεί την προσδοκία για

ένα μοντέλο $\mathcal{M}$ της $f : \mathcal{X} \to \mathbb{R}^N$ η τιμή της $f(x)$ να υπερβεί (αρνητικά) ένα κατώφλι y:

$$\mathrm{EI}_{y^*}(x) := \int_{-\infty}^{\infty} \max(y^* - y, 0)\, p_M(y \mid x) dy$$

Ο σκοπός είναι η μεγιστοποίηση του EI ως προς x δηλαδή η επιλογή του βέλτιστου συνόλου υπερπαραμέτρων για το τρέχον μοντέλο Υποκατάστασης. Με τη χρήση του συγκεκριμένου κριτηρίου καταφέρνουμε την εξερεύνηση του χώρου αποφεύγοντας ταυτόχρονα προβλήματα υπερπροσαρμογής.

Σχετικά με την επιλογή του μοντέλου Υποκατάστασης υπάρχουν διάφορα προτεινόμενα μοντέλα όπως Gaussian Processes, Random Forests Regressors καθώς και το Tree-structured Parzen Estimator (TPE) [91] η χρήση του οποίου επιλέχθηκε στα πλαίσια της παρούσας διπλωματικής. Σε αντίθεση με την προσέγγιση των Gaussian Processes, αποφεύγεται η απευθείας μοντελοποίηση της $p(y|x)$ μέσω της χρήσης του κανόνα του Bayes.

Η προσέγγιση του TPE μοντελοποιεί την $p(x|y)$, η οποία αποτελεί την κατανομή των υπερπαραμέτρων δοθέντος του σκορ της αντικειμενικής συνάρτησης, χρησιμοποιώντας δύο συναρτήσεις πυκνότητας:

$$p(x \mid y) = \begin{cases} \ell(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases}$$

όπου $\ell(x)$ ορίζεται ως η κατανομή των παρατηρήσεων $\left\{x^{(i)}\right\}$ για τις οποίες η τιμή της συνάρτησης απώλειας $f\left(x^{(i)}\right)$ είναι χαμηλότερη από κάποιο κατώφλι y^* και $g(x)$ η κατανομή των υπόλοιπων παρατηρήσεων. Για τον καθορισμό του κατωφλιού y^* ο αλγόριθμος TPE επιλέγει ένα ποσοστημόριο γ τέτοιο ώστε $p(y < y^*) = \gamma$. Εφαρμόζοντας το κανόνα του Bayes στο EI το κριτήριο μετασχηματίζεται σε

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y)\, p(y \mid x) dy = \int_{-\infty}^{y^*} (y^* - y)\, \frac{p(x \mid y) p(y)}{p(x)} dy$$

Στα πλαίσια της χρήσης του αλγορίθμου TPE μπορούμε να θεωρήσουμε $\gamma = p(y < y^*)$ και $p(x) = \int_{\mathbb{R}} p(x \mid y) p(y) dy = \gamma \ell(x) + (1-\gamma) g(x)$ συνεπώς $\int_{-\infty}^{y^*} (y^* - y)\, p(x \mid y) p(y) dy = \ell(x) \int_{-\infty}^{y^*} (y^* - y)\, p(y) dy = \gamma y^* \ell(x) - \ell(x) \int_{-\infty}^{y^*} p(y) dy$. Αντικαθιστώντας στον τύπο της EI καταλήγουμε στο εξής κριτήριο:

$$EI_{y^*}(x) = \frac{\gamma y^* \ell(x) - \ell(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma \ell(x) + (1-\gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{\ell(x)}(1-\gamma)\right)^{-1}$$

Είναι προφανές ότι το παρόν κριτήριο εξαρτάται από το λόγο $\frac{g(x)}{\ell(x)}$ και μεγιστοποιείται για σημεία με υψηλή πιθανότητα για την κατανομή $\ell(x)$ και μικρή πιθανότητα για τη κατανομή $g(x)$.

Τέλος αξίζει να σημειωθεί ότι η $\ell(x)$ και $g(x)$ αποτελούν κατανομές συνεπώς οι υπερπαράμετροι που δειγματοληπτούνται από αυτές είναι πιθανόν να είναι κοντά αλλά όχι ακριβώς στο μέγιστο της EI. Επιπλέον, καθώς το μοντέλο υποκατάστασης αποτελεί μια εκτίμηση της αντικειμενικής συνάρτησης, επιλεγμένες υπερπαράμετροι για τις οποίες ενδέχεται να μην υπάρχει βελτίωση των αποτελεσμάτων οδηγούν πάραυτα σε ενημέρωση του μοντέλου υποκατάστασης.

Μέρος III

Περιπτωσιολογική Μελέτη

Κεφάλαιο 6

Μεθοδολογία

6.1 Σύνολο Δεδομένων & Προ-Επεξεργασία

Στα πλαίσια της παρούσας διπλωματικής εργασίας επιλέχθηκε ένα σύνολο δεδομένων από τη περιοχή της ανίχνευσης εισβολής από δεδομένα δικτύου. Το συγκεκριμένο σύνολο δεδομένων περιέχει μετρήσεις σχετικές με κινήσεις δικτύου καθώς και αν η εκάστοτε κίνηση αποτελεί προϊόν επίθεσης ή όχι. Πιο συγκεκριμένα για την εκπαίδευση των μοντέλων χρησιμοποιήθηκε το σύνολο δεδομένων CSE-CIC-IDS2018 [95] το οποίο δημιουργήθηκε από το Institute for Cybersecurity το 2018. Το σύνολο δεδομένων έχει προκύψει έπειτα από προσομοιώσεις κινήσεων δικτύου σε ένα διάστημα δέκα ημερών σε ένα ελεγχόμενο περιβάλλον στο Amazon Web Service (AWS). Το τελικό σύνολο δεδομένων αποτελείται από αρχεία pcap τα οποία περιέχουν τις ακατέργαστες προσομοιώσεις δικτύου καθώς και αρχεία μορφής csv τα οποία δημιουργήθηκαν μέσω του CICFlowMeter [96] και περιέχουν τα εξαχθέντα στατιστικά χαρακτηριστικά που θα χρησιμοποιηθούν με σκοπό την δημιουργία μοντέλων εύρεσης ανωμαλιών.

Στη τελική μορφή του, το σύνολο δεδομένων περιέχει χαρακτηριστικά όπως Timestamp, Protocol, Flow ID, Source IP, Destination IP, Source Port και Destination Port, τα οποία θα αγνοηθούν στα πλαίσια της ανάλυση καθώς οι διευθύνσεις και οι θύρες δύναται να αντικατασταθούν από τον επιτιθέμενο ενώ σε ότι αφορά τα υπόλοιπα χαρακτηριστικά υποθέτουμε ότι η μορφή των δεδομένων αποτελεί σημαντικότερο παράγοντα.

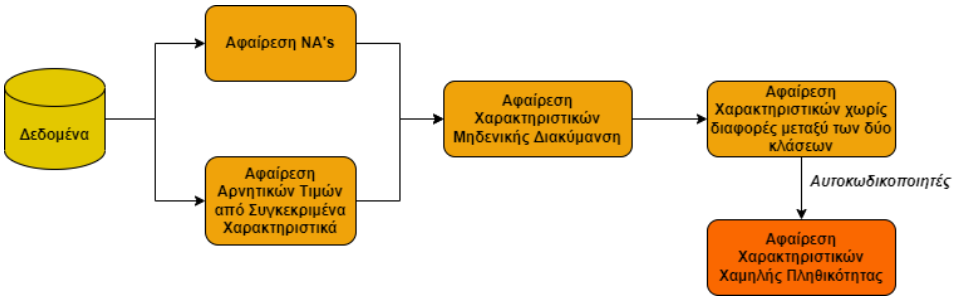
Επιπλέον διαθέτουμε περίπου 70 χαρακτηριστικά, τα οποία αφορούν στατιστικές μετρήσεις εξαχθέντες εκ των προσομοιωμένων κινήσεων δικτύου, τα οποία και περιγράφονται στον ακόλουθο Πίνακα 6.1. Ενδεικτικά αναφέρουμε ότι τα χαρακτηριστικά αφορούν το πλήθος και το μέγεθος των απεσταλμένων πακέτων, το μήκος και τη διάρκεια της ροής, το χρόνο ανάμεσα σε δύο απεσταλμένα πακέτα ή ροες, το πλήθος bytes, κ.α.

Χαρακτηριστικό	**Περιγραφή Χαρακτηριστικού**
flow_duration	Διάρκεια Ροής
tot_fwd_pkt	Συνολικά πακέτα στην εμπρόσθια κατεύθυνση
tot_bwd_pkt	Συνολικά πακέτα στην οπίσθια κατεύθυνση
totlen_fwd_pkt	Συνολικό μέγεθος πακέτων στην εμπρόσθια κατεύθυνση
fwd_pkt_len_max	Μέγιστο μέγεθος πακέτων στην εμπρόσθια κατεύθυνση
fwd_pkt_len_min	Ελάχιστο μέγεθος πακέτων στην εμπρόσθια κατεύθυνση
fwd_pkt_len_avg	Μέσο μέγεθος πακέτου στην εμπρόσθια κατεύθυνση
fwd_pkt_len_std	Τυπική απόκλιση μεγέθους του πακέτου στην εμπρόσθια κατεύθυνση
bwd_pkt_len_max	Μέγιστο μέγεθος πακέτων στην οπίσθια κατεύθυνση
bwd_pkt_len_min	Ελάχιστο μέγεθος πακέτων στην οπίσθια κατεύθυνση
bwd_pkt_len_avg	Μέσο μέγεθος πακέτου στην οπίσθια κατεύθυνση
bwd_pkt_len_std	Τυπική απόκλιση μεγέθους του πακέτου στην οπίσθια κατεύθυνση
flow_byt_s	Αριθμός bytes που μεταφέρθηκαν ανά δευτερόλεπτο
flow_pkts_s	Αριθμός πακέτων που μεταφέρθηκαν ανά δευτερόλεπτο
flow_iat_avg	Μέσος χρόνος ανάμεσα σε δύο ροές
flow_iat_std	Τυπική απόκλιση του χρόνου ανάμεσα σε δύο ροές
flow_iat_max	Μέγιστος χρόνος ανάμεσα σε δύο ροές
flow_iat_min	Ελάχιστος χρόνος ανάμεσα σε δύο ροές
fwd_iat_tot	Συνολικός χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην εμπρόσθια κατεύθυνση
fwd_iat_avg	Μέσος χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην εμπρόσθια κατεύθυνση
fwd_iat_std	Τυπική απόκλιση χρόνου ανάμεσα σε δύο πακέτα απεσταλμένα στην εμπρόσθια κατεύθυνση
fwd_iat_max	Μέγιστος χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην εμπρόσθια κατεύθυνση
fwd_iat_min	Ελάχιστος χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην εμπρόσθια κατεύθυνση
bwd_iat_tot	Συνολικός χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην οπίσθια κατεύθυνση
bwd_iat_avg	Μέσος χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην οπίσθια κατεύθυνση
bwd_iat_std	Τυπική απόκλιση χρόνου ανάμεσα σε δύο πακέτα απεσταλμένα στην οπίσθια κατεύθυνση
bwd_iat_max	Μέγιστος χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην οπίσθια κατεύθυνση
bwd_iat_min	Ελάχιστος χρόνος ανάμεσα σε δύο πακέτα απεσταλμένα στην οπίσθια κατεύθυνση
fwd_psh_flag	# φορών που το PSH flag δόθηκε σε πακέτα προς την εμπρόσθια κατεύθυνση
bwd_psh_flag	# φορών που το PSH flag δόθηκε σε πακέτα προς την οπίσθια κατεύθυνση
fwd_urg_flag	# φορών που το URG flag δόθηκε σε πακέτα προς την εμπρόσθια κατεύθυνση
bwd_urg_flag	# φορών που το URG flag δόθηκε σε πακέτα προς την οπίσθια κατεύθυνση
fwd_hdr_len	Συνολικό πλήθος bytes στις επικεφαλίδες για την εμπρόσθια κατεύθυνση
bwd_hdr_len	Συνολικό πλήθος bytes στις επικεφαλίδες για την οπίσθια κατεύθυνση
fwd_pkt_s	Αριθμός εμπρόσθιων πακέτων ανά δευτερόλεπτο
bwd_pkt_s	Αριθμός οπίσθιων πακέτων ανά δευτερόλεπτο
pkt_len_min	Ελάχιστο μήκος ροής
pkt_len_max	Μέγιστο μήκος ροής
pkt_len_avg	Μέσο μήκος ροής
pkt_len_std	Τυπική απόκλιση μήκους ροής
pkt_len_va	Ελάχιστος χρόνος ενδιάμεσης άφιξης του πακέτου
fin_flag_cnt	Αριθμός πακέτων με FIN
syn_flag_cnt	Αριθμός πακέτων με SYN
rst_flag_cnt	Αριθμός πακέτων με RST
pst_flag_cnt	Αριθμός πακέτων με PUSH
ack_flag_cnt	Αριθμός πακέτων με ACK
urg_flag_cn	Αριθμός πακέτων με URG
cwe_flag_cnt	Αριθμός πακέτων με CWE
ece_flag_cnt	Αριθμός πακέτων με ECE
down_up_ratio	Αναλογία μεταφόρτωσης και λήψης
pkt_size_avg	Μέσο μέγεθος πακέτου
fwd_seg_avg	Μέσο παρατηρηθέν μέγεθος στην εμπρόσθια κατεύθυνση
bwd_seg_avg	Μέσο παρατηρηθέν μέγεθος στην οπίσθια κατεύθυνση
fwd_byt_blk_avg	Μέση αναλογία bytes bulk στην εμπρόσθια κατεύθυνση
fwd_pkt_blk_avg	Μέσος αριθμός πακέτων bytes bulk στην εμπρόσθια κατεύθυνση
fwd_blk_rate_avg	Μέσος αριθμός αναλογίας bulk στην εμπρόσθια κατεύθυνση
bwd_byt_blk_avg	Μέση αναλογία bytes bulk στην οπίσθια κατεύθυνση
fwd_blk_rate_avg	Μέσος αριθμός αναλογίας bulk στην εμπρόσθια κατεύθυνση
bwd_byt_blk_avg	Μέση αναλογία bytes bulk στην οπίσθια κατεύθυνση
bwd_pkt_blk_avg	Μέσος αριθμός πακέτων bytes bulk στην οπίσθια κατεύθυνση
bwd_blk_rate_avg	Μέσος αριθμός αναλογίας bulk στην οπίσθια κατεύθυνση

Χαρακτηριστικό	Περιγραφή Χαρακτηριστικού
subfl_fwd_pk	Μέσος αριθμός πακέτων σε μια υπό ροή στην εμπρόσθια κατεύθυνση
subfl_fwd_byt	Μέσος αριθμός bytes σε μια υπό ροή στην εμπρόσθια κατεύθυνση
subfl_bwd_pkt	Μέσος αριθμός πακέτων σε μια υπό ροή στην οπίσθια κατεύθυνση
subfl_bwd_byt	Μέσος αριθμός bytes σε μια υπό ροή στην οπίσθια κατεύθυνση
init_fwd_win_byts	# bytes που στάλθηκαν σε ένα αρχικό παράθυρο στην εμπρόσθια κατεύθυνση
init_bwd_win_byts	# bytes που στάλθηκαν σε ένα αρχικό παράθυρο στην οπίσθια κατεύθυνση
fwd_act_pkt	# πακέτων με τουλάχιστον 1 byte από TCP δεδομένα στην εμπρόσθια κατεύθυνση
fwd_seg_size_min	Ελάχιστο παρατηρηθέν μέγεθος τμήματος στην εμπρόσθια κατεύθυνση
atv_avg	Μέσος χρόνος κατά τον οποίο μια ροή ήταν ενεργή προτού να γίνει idle
atv_std	Τυπική απόκλιση χρόνου κατά τον οποίο μια ροή ήταν ενεργή προτού να γίνει idle
atv_max	Μέγιστος χρόνος κατά τον οποίο μια ροή ήταν ενεργή προτού να γίνει idle
atv_min	Ελάχιστος χρόνος κατά τον οποίο μια ροή ήταν ενεργή προτού να γίνει idle
idl_avg	Μέσος χρόνος κατά τον οποίο μια ροή ήταν ενεργή idle προτού να γίνει ενεργή
idl_std	Τυπική απόκλιση χρόνου κατά τον οποίο μια ροή ήταν idle προτού να γίνει ενεργή
idl_max	Μέγιστος χρόνος κατά τον οποίο μια ροή ήταν idle προτού να γίνει ενεργή
idl_min	Ελάχιστος χρόνος κατά τον οποίο μια ροή ήταν idle προτού να γίνει ενεργή

Πίνακας 6.1: *Περιγραφή εξαχθέντων χαρακτηριστικών μέσω CICFlowMeter*

Το σύνολο δεδομένων περιέχει τόσο Φυσιολογικές κινήσεις δικτύου όσο και τα πιο συνήθη είδη επιθέσεων (Πίνακας 6.4). Ο σκοπός της διπλωματικής είναι η αναγνώριση μη φυσιολογικών κινήσεων δικτύου (επιθέσεις) αγνοώντας το είδος της επίθεσης σε μια προσπάθεια δημιουργίας ενός ενοποιημένου συστήματος το οποίο μέσω της χρήσης τεχνικών μηχανικής μάθησης θα είναι σε θέση να αναγνωρίζει κάθε είδους επιθέσεις ταυτοποιόντας την ως μη φυσιολογική. Όπως παρατηρούμε και στο Πίνακα 6.4 λαμβάνοντας υπόψη μας το είδος της κάθε επίθεσης καταλήγουμε μετά την προ-επεξεργασία σε κατηγορίες οι οποίες διαθέτουν ελάχιστες παρατηρήσεις με τις οποίες δεν θα ήταν εφικτή η δημιουργία κάποιο μοντέλου πολυταξικής ταξινόμησης κάτι το οποίο αντιμετωπίζεται μετατρέποντας το πρόβλημα από πολυταξική σε δυαδική ταξινόμηση.



Σχήμα 6.1: *Μεθοδολογία Προ-Επεξεργασίας Δεδομένων*

Στα πλαίσια της προ-επεξεργασίας των δεδομένων (Σχήμα 6.1) αφαιρούμε από το σύνολο δεδομένων τόσο παρατηρήσεις με ελλιπείς τιμές όσο και παρατηρήσεις στις οποίες εμφανίζονται αρνητικές τιμές εξαιρώντας τις μεταβλητές init_fwd_win_byts και init_bwd_win_byts για τις οποίες οι αρνητικές τιμές μπορούν να εξηγηθούν. Στη συνέχεια, αφαιρούμε χαρακτηριστικά τα οποία εμφανίζουν μηδενική διακύμανση στο δείγμα καθώς και χαρακτηριστικά τα οποία δεν εμφανίζουν στατιστικά σημαντικές διαφορές στη δειγματική κατανομή ανάμεσα στις δύο κλάσεις ενδιαφέροντος. Το τελευταίο γίνεται εφικτό μέσω της χρήσης του μη παραμετρικού ελέγχου Two Sample Kolmogorov-Smirnov.

Ο μη παραμετρικός έλεγχος Two Sample Kolmogorov-Smirnov χρησιμοποιεί την εμπειρική κατανομή του δείγματος με σκοπό τη διερεύνηση διαφορών στα επίπεδα κάποιας κατηγορικής μεταβλητής, ορίζοντάς τον εξής έλεγχο:

$$H_0 : F_{1,n}(x) = F_{2,m}(x), \forall x$$

$$H_a : F_{1,n}(x)! = F_{2,m}(x)$$

Το στατιστικό Kolmogorov–Smirnov υπολογίζεται ως

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

όπου $F_{1,n}$ και $F_{2,m}$ αποτελούν τη εμπειρική δειγματική κατανομή των δύο δειγμάτων μεγέθους n και m (ένα δείγμα ανά επίπεδο της κατηγορικής μεταβλητής). Για μεγάλα μεγέθη δείγματος η μηδενική υπόθεση απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας a εάν ικανοποιείται η ακόλουθη συνθήκη:

$$D_{n,m} > c(a)\sqrt{\frac{n+m}{n \cdot m}}$$

Η προαναφερθείσα διαδικασία καταλήγει στην αφαίρεση από το σύνολο δεδομένων 10 χαρακτηριστικών (Πίνακας 6.2) εκ των οποίων οκτώ χαρακτηριστικά εμφανίζουν μηδενική διακύμανση ενώ για τα υπόλοιπα δύο δεν απορρίπτεται η μηδενική υπόθεση του ελέγχου Kolmogorov–Smirnov.

Χαρακτηριστικό
bwd_blk_rate_avg
bwd_byts_b_avg
bwd_pkts_b_avg
bwd_psh_flags
bwd_urg_flags
fwd_blk_rate_avg
fwd_byts_b_avg
fwd_pkts_b_avg

Χαρακτηριστικό
cwe_flag_count
fwd_urg_flags

Πίνακας 6.2: *Χαρακτηριστικά Μηδενικής Διακύμανσης (Αριστερά) και Χωρίς Διαφορές μεταξύ των δύο Κλάσεων (Δεξιά)*

Επιπροσθέτως στα πλαίσια της προετοιμασίας των δεδομένων με σκοπό την εκπαίδευση Αυτοκωδικοποιήτων, αφαιρέθηκαν επιπλέον οκτώ χαρακτηριστικά χαμηλής πληθικότητας (Σχήμα 6.3) καθώς πειραματικά παρατηρήθηκε ότι με την απουσία αυτών κατά τη περίοδο της εκπαίδευσης, τα μοντέλα κατέληγαν σε καλύτερα αποτέλεσμά στο πλαίσιο της γενίκευσης.

Χαρακτηριστικό
fwd_psh_flags
fin_flag_cnt
syn_flag_cnt
rst_flag_cnt
psh_flag_cnt
ack_flag_cnt
urg_flag_cnt
ece_flag_cnt

Πίνακας 6.3: *Χαρακτηριστικά Χαμηλής Πληθικότητας*

Στο τέλος της προ-επεξεργασίας καταλήγουμε σε ένα σύνολο από 66 επεξηγηματικά χαρακτηριστικά (58 επεξηγηματικά χαρακτηριστικά για τους Αυτοκωδικοποιητές). Συνολικά το μέγεθος του συνόλου δεδομένων μετά την προ-επεξεργασία αποτελείται τόσο από ένα πλήθος παρατηρήσεων, επαρκή για την εκπαίδευση και των πιο απαιτητικών από άποψη όγκου παρατηρήσεων μοντέλων, καθώς και από δύο κλάσης, Φυσιολογική (Benign) και Επίθεση (Attack), τη δεύτερη να καλύπτει περίπου το 17% των συνολικών παρατηρήσεων έναντι 83% για την κλάση των Φυσιολογικών. Είναι προφανές ότι τα δεδομένα χαρακτηρίζονται από σημαντική ανισορροπία ανάμεσα στις δύο κλάσης, αλλά και ειδικότερα ως προς το είδος των επιθέσεων (Πίνακας 6.4).

Τύπος Κινήσεων Δικτύου	**Συχνότητα** (%)
Φυσιολογική Κίνηση Δικτύου	13390234 (82.978)
DDoS attack-HOIC	686012 (4.251)
DDoS attack-LOIC-HTTP	576191 (3.571)
DoS attack-Hulk	461912 (2.862)
Botnet	286191 (1.773)
FTP-Brute Force	193354 (1.198)
SSH-Brute Force	187589 (1.162)
Infiltration	160639 (0.995)
DoS attack-SlowHTTPTest	139890 (0.867)
DoS attack-GoldenEye	41508 (0.257)
DoS attack-Slowloris	10990 (0.068)
DDoS attack-LOIC-UDP	1730 (0.011)
Brute Force-Web	611 (0.004)
Brute Force-XSS	230 (0.001)
SQL Injection	87 (0.001)

Πίνακας 6.4: *Κατανομή Προσομοιωμένων Επιθέσεων (CSE-CIC-IDS2018)*

Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκαν διαφορετικές τεχνικές, συνεπώς και για το διαχωρισμό των δεδομένων σε δείγματα Εκπαίδευσης, Επικύρωσης και Ελέγχου.

Ξεκινώντας με τα μοντέλα επιβλεπόμενης μάθησης, χρησιμοποιήθηκε στρωματοποιημένος διαχωρισμός στα επίπεδα των διαφορετικών ειδών επιθέσεων με σκοπό να εξασφαλίσουμε ότι η κατανομή των δεδομένων ανάμεσα στα διαφορετικά υποσύνολα δεν θα διαφέρει ως προς την ανεξάρτητη μεταβλητή και το μοντέλο να έχει τη δυνατότητα να εκπαιδευτεί όσο το δυνατό πιο δίκαια σε όλο το φάσμα των καταγεγραμμένων χαρακτηριστικών. Μέσω αυτού του

διαχωρισμού δύναται τα μοντέλα να μάθουν γενικά χαρακτηριστικά τα οποία χαρακτηρίζουν τη γενικότερη συμπεριφορά μια ροής δικτύου. Πιο συγκεκριμένα, χρησιμοποιήθηκε διαχωρισμός της τάξης 80/10/10 σε δεδομένα Εκπαίδευσης (Train), Επικύρωσης (Validation) και Ελέγχου (Test) με τα αποτελέσματα να παρουσιάζονται στο Πίνακα 6.5.

	Ετικέτα Κινήσης Δικτύου	
	Benign (%)	Attacks (%)
Train	10712187(83)	2197547(17)
Validation	1339023(83)	274694(17)
Test	1339024(83)	274693(17)

Πίνακας 6.5: *Κατανομή Παρατηρήσεων ανά Υποσύνολο Δεδομένων*

Επιπλέον, εξαιτίας της μη ισορροπίας μεταξύ των δύο κλάσεων οφείλουμε να διαφυλάξουμε ότι τα χαρακτηριστικά όπως αυτά κατανέμονται στα τρία προαναφερθέν σύνολα, προέρχονται από τους ίδιους πληθυσμούς. Για αυτό το λόγο χρησιμοποιήθηκε εκ νέου ο έλεγχος Kolmogorov-Smirnov (KS) χρησιμοποιώντας αυτή τη φορά τη διόρθωση Holm–Bonferroni [97] για τα υπολογισθέντα p-values, λόγο της πραγματοποίησης πολλαπλών ελέγχων. Τα τελικά p-values είναι όλα μεγαλύτερα από 0.05 συνεπώς δεν υπάρχουν επαρκή στοιχεία για να θεωρήσουμε ότι τα χαρακτηριστικά μας συμπεριφέρονται με τρόπο διαφορετικό ανάμεσα στα σύνολα Εκπαίδευσης, Επικύρωσης και Ελέγχου και συνεπώς ο διαχωρισμός μπορεί να θεωρηθεί αντιπροσωπευτικός.

Συνεχίζοντας με τα μοντέλα μιας κλάσης, τα οποία χρησιμοποιήθηκαν για την εύρεση ανωμαλιών, χρησιμοποιήθηκε διαφορετική προσέγγιση καθώς τα προαναφερθείσα μοντέλα εκπαιδεύονται χρησιμοποιώντας αποκλειστικά δεδομένα από τη πλειοψηφική κλάση. Πιο συγκεκριμένα για την εκπαίδευση των μοντέλων χρησιμοποιήθηκε το 80% από το σύνολο των παρατηρήσεων από τη φυσιολογική κλάση. Τα υπολειπόμενα δεδομένα από τη κλάση των φυσιολογικών δεδομένων δικτύου συνδυάστηκαν με τα δεδομένα από τη κλάση των επιθέσεων και το σύνολο τους διαχωρίστηκε χρησιμοποιώντας στρωματοποίηση στα επίπεδα του είδους των επιθέσεων με αναλογία 0/50/50 σε Train, Validation και Test αντίστοιχα. Στο Πίνακα 6.6 παρατίθενται τα αποτελέσματα του προαναφερθέν διαχωρισμού.

	Ετικέτα Κίνησης Δικτύου	
	Benign (%)	Attacks (%)
Train	10712187(100)	0(0.0)
Validation	1339023(49.4)	1373467(50.6)
Test	1339024(49.4)	1373467(50.6)

Πίνακας 6.6: *Κατανομή Παρατηρήσεων ανά Υποσύνολο Δεδομένων (Μοντέλα μιας Κλάσης)*

Επιπροσθέτως, το σύνολο δεδομένων CIC-IDS2017 [95] χρησιμοποιήθηκε για λόγους επικύρωσης των τελικών αποτελεσμάτων για τα μοντέλα τα οποία εκπαιδεύτηκαν στο σύνολο δεδομένων CSE-CIC-IDS2018. Το συγκεκριμένο σύνολο δεδομένων αποτελεί μια προγενέστερη έκδοση του συνόλου δεδομένων CSE-CIC-IDS2018 με χαρακτηριστικά αντίστοιχα των εξαχθέντων στατιστικών όπως αυτά περιγράφηκαν για το CSE-CIC-IDS2018. Χαρακτηρίζεται από μικρότερο πλήθος και εύρος από προσομοιωμένα είδη επιθέσεων. Για λόγους συ-

νάφειας επαναλαμβάνουμε τα αντίστοιχα βήματα προ-επεξεργασίας όπως αυτά περιγράφτηκαν για το αρχικό σύνολο δεδομένων. Έπειτα από την προ-επεξεργασία η τελική κατανομή των διαφορών τύπων κινήσεων δικτύου παρουσιάζεται στον ακόλουθο Πίνακα 6.7.

Τύπος Κινήσεων Δικτύου	**Συχνότητα** (%)
Φυσιολογική Κίνηση Δικτύου	2268624 (80.306)
DoS attack-Hulk	229965 (8.140)
PortScan	158804 (5.621)
DDoS	128006 (4.531)
DoS attack-GoldenEye	10288 (0.364)
FTP-Brute Force	7931 (0.281)
SSH-Brute Force	5895 (0.209)
DoS attack-Slowloris	5796 (0.205)
DoS attack-SlowHTTPTest	5499 (0.195)
Botnet	1956 (0.069)
Brute Force-Web	1507 (0.053)
Brute Force-XSS	652 (0.023)
Infiltration	35 (0.0012)
SQL Injection	21 (0.0007)
Heartbleed	7 (0.0002)

Πίνακας 6.7: *Κατανομή Προσομοιωμένων Επιθέσεων (CIC-IDS2017)*

Όπως παρατηρούμε το σύνολο δεδομένων παρουσιάζει ανισορροπία μεταξύ της κλάσης των επιθέσεων και των φυσιολογικών κινήσεων δικτύου σε αντίστοιχη κλίμακα με εκείνη του συνόλου δεδομένων που θα χρησιμοποιηθεί με σκοπό την εκπαίδευση των μοντέλων. Παρόλο αυτά, όπως παρατηρούμε, το είδος των επιμέρους επιθέσεων παρουσιάζει σημαντικά διαφορετική κατανομή καθώς είδη επιθέσεων, όπως για παράδειγμα DoS attack-Hulk και DoS attack-GoldenEye, καλύπτουν σημαντικά μεγαλύτερο ποσοστό επί του συνόλου των κινήσεων ενώ αντίστοιχα άλλα είδη κινήσεων τοποθετούνται χαμηλότερα ως προς το ποσοστό των επιθέσεων. Επιπλέον, η δεύτερη μεγαλύτερη κλάση επιθέσεων ταυτοποιείται ως DDoS χωρίς να συγκεκριμενοποιείται το είδος της ενώ στη συνέχεια παρατηρούμε ότι εισάγονται διαφορετικού τύπου επιθέσεις στο σύνολο (PortScan, Heartbleed) οι οποίες και απουσίαζαν από το πρωταρχικό σύνολο δεδομένων. Για αυτούς τους λόγους η αξιολόγηση σύμφωνα με αυτό το σύνολο δεδομένων παρουσιάζει εξαιρετικό ενδιαφέρον, καθώς έρχεται να δοκιμάσει τα μοντέλα σε ένα πλαίσιο το οποίο θα μπορούσαμε να θεωρήσουμε ότι προσομοιάζει συνθήκες κατά τις οποίες τα μοντέλα θα έρθουν αντιμέτωπα στη περίπτωση όπου επιλεχθούν για χρήση στα πλαίσια της παραγωγής.

6.2 Μετρικές Αξιολόγησης

6.2.1 Ακρίβεια (Accuracy)

Η Ακρίβεια ορίζεται ως το ποσοστό σωστά ταξινομημένων παρατηρήσεων σε σχέση με το σύνολο των παρατηρήσεων.

$$\text{Ακρίβεια} = \frac{TP + TN}{\text{Συνολικό Μέγεθος Δείγματος}}$$

Το παράδοξο της Ακρίβειας [98] είναι ότι σε περιπτώσεις όπου τα δεδομένα είναι μη ισορροπημένα η παραπάνω μετρική δίνει αποτελέσματα τα οποία φαίνονται καλύτερα σε σχέση με τη πραγματικότητα. Για παράδειγμα σε ένα πρόβλημα δυαδικής ταξινόμησης όπου η πλειοψηφική κλάση αποτελείται από 95 παρατηρήσεις έναντι της μειοψηφικής κλάσης η οποία αποτελείται από πέντε παρατηρήσεις τότε εάν το μοντέλο προβλέπει όλες τις παρατηρήσεις στη πλειοψηφική κλάση τότε η Ακρίβεια του μοντέλου υπολογίζεται ίση με 95% ποσοστό ιδιαίτερα υψηλό. Στη πραγματικότητα όμως το μοντέλο έχει αγνοήσει πλήρως την μειοψηφική κλάση. Για αυτό το λόγο εισάγουμε στην ανάλυση μας μετρικές οι οποίες λαμβάνουν υπόψη τους την ανισορροπία των κλάσεων και δύναται να ποσοτικοποιήσουν αντιπροσωπευτικά ως προς τα δεδομένα την απόδοση των μοντέλων.

6.2.2 Ευστοχία (Precision)

Το Precision ορίζεται ως η αναλογία των σωστά ταξινομημένων παρατηρήσεων σε σχέση με το σύνολο των σωστά και λάθος ταξινομημένων παρατηρήσεών από τη θετική κλάση.

$$\text{Precision} = \frac{TP}{TP + FP}$$

6.2.3 Ανάκληση/Ευαισθησία (Recall/Sensitivity)

Το Recall ορίζεται ως η αναλογία των σωστά ταξινομημένων παρατηρήσεων σε σχέση με το σύνολο των σωστά ταξινομημένων παρατηρήσεων από τη θετική κλάση καθώς και των ψευδώς αρνητικών περιπτώσεων.

$$\text{Recall} = \frac{TP}{TP + FN}$$

6.2.4 F1 - Score

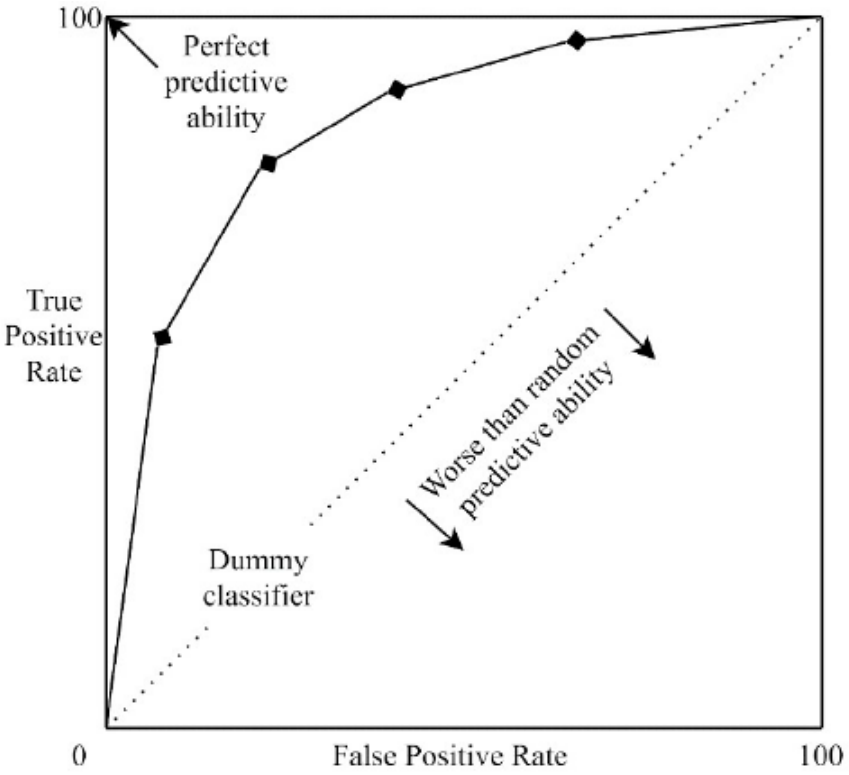
Ως F_1 - Score ορίζουμε τον αρμονικό μέσο της Ανάκλησης και της Ευστοχίας.

$$F_1 \text{ - Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.2.5 Χαρακτηριστική Καμπύλη ROC (ROC Curve)

Μέσω της Χαρακτηριστικής Καμπύλης ROC [99] καταφέρνουμε να απεικονίσουμε τους συνδυασμούς της αναλογίας των ψευδών θετικών περιπτώσεων, γνωστό και ως (1−Ειδικότητα),

και της Ευαισθησίας, γνωστή και ως η αναλογία των αληθώς θετικών περιπτώσεων, (στους άξονες X και Y αντίστοιχα) για όλες τις τιμές που παρατηρούμε στο δείγμα. Με αυτό το τρόπο καταφέρνουμε να απεικονίσουμε τη προβλεπτική ικανότητα του μοντέλου μας σε διαφορετικά κατώφλια ταξινόμησης. Οι συγκεκριμένες καμπύλες παρουσιάζουν μονοτονική συμπεριφορά, με αποτέλεσμα όσο αυξάνεται η αναλογία των αληθώς θετικών περιπτώσεων τόσο να αυξάνεται η αναλογία των ψευδών θετικών περιπτώσεων. Όπως φαίνεται και στο Σχήμα 6.2 ένα μοντέλο χειρότερο από ένα τυχαίο ταξινομητή δίνει τιμές κάτω από τη διαγώνιο ενώ αντίθετα για ένα τέλειο μοντέλο περιμένουμε τιμές κοντά στην πάνω αριστερή γωνία του σχήματος.



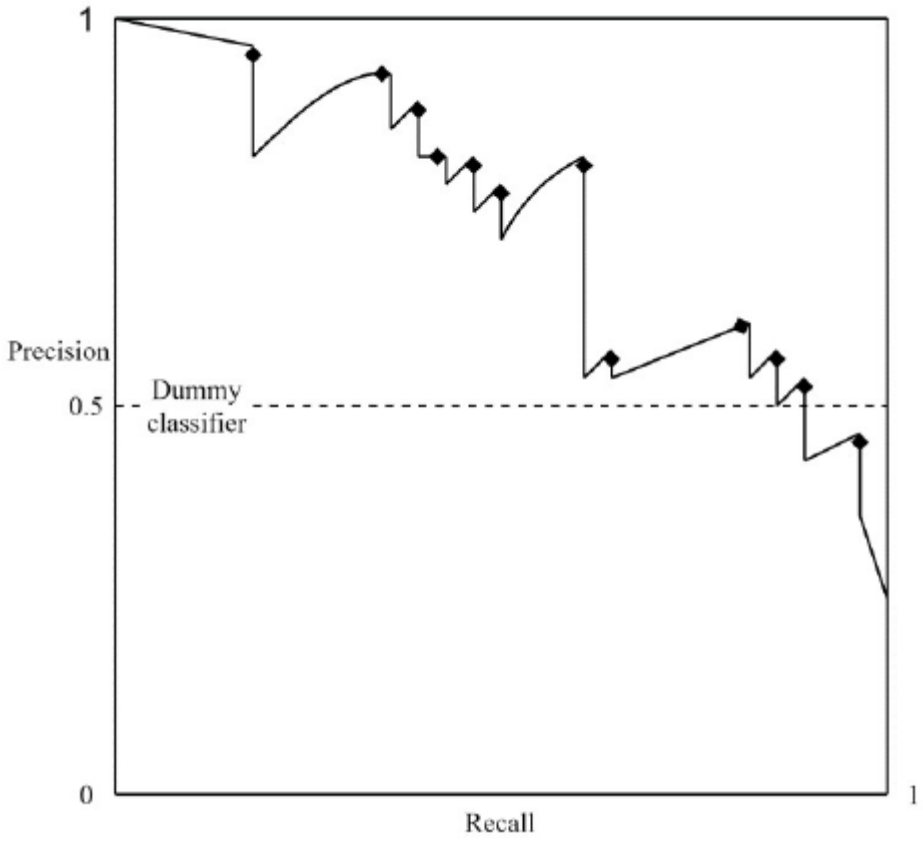
Σχήμα 6.2: *Χαρακτηριστική Καμπύλη ROC*

Μια άλλη σημαντική ποσότητα η οποία σχετίζεται με τις καμπύλες ROC είναι η Περιοχή κάτω από τη Καμπύλη ROC (ROC-AUC). Η τιμή αυτή δεν είναι τίποτε άλλο από το εμβαδόν που περικλείεται μεταξύ της καμπύλης ROC και του άξονα των X. Η ποσότητα αυτή έχει πιθανοθεωρητική ερμηνεία και μπορεί να ερμηνευθεί ως η πιθανότητα τα σκόρς ενός ταξινομητή να κατατάσσουν ένα τυχαίο σημείο από τη θετική κλάση υψηλότερα από ένα τυχαίο σημείο από την αρνητική κλάση. Η ROC-AUC παίρνει τιμές μεταξύ 0 και 1 και χρησιμοποιείται για να πάρουμε μια γενική εικόνα της συνολικής απόδοσης του μοντέλου μας. Στη πράξη για ένα ισορροπημένο σύνολο δεδομένων ένα μοντέλο το οποίο μαντεύει τυχαία ανάμεσα δύο κλάσεων έχει ROC-AUC ίση με 0.5 συνεπώς οτιδήποτε κάτω από αυτή τη τιμή θεωρείται χειρότερο από τυχαίο. Παρόλο αυτά σε μη ισορροπημένα σύνολα δεδομένων, η απόδοση των καμπύλων ROC και της μετρικής ROC-AUC έχει αμφισβητηθεί καθώς τείνει να είναι παραπλανητική [100], [101].

6.2.6 Καμπύλη Ευστοχίας - Ανάκλησης (PR Curve)

Οι καμπύλες Ευστοχίας - Ανάκλησης δείχνουν τις μεταβολές ανάμεσα στην Ευστοχία και την Ανάκληση για διάφορα κατώφλια ταξινόμησης. Σε αντίθεση με τις Καμπύλες ROC σε περίπτωση όπου το σύνολο των δεδομένων είναι εξαιρετικά μη ισορροπημένο οι PR Curves δύναται να αποτυπώσουν με μεγαλύτερη ακρίβεια την προβλεπτική ικανότητα του μοντέλου [102], [103]. Παρόλο αυτά υπάρχει ένα προς ένα σχέση μεταξύ των δύο καμπύλων με αποτέλεσμα ένας ταξινομητής με τη βέλτιστη PR Curve να επιτυγχάνεται μαζί την αντίστοιχη βέλτιστη ROC Curve [104].

Στο Σχήμα 6.3 κάθε σημείο διακοπής μεταξύ των γραμμών αποτελεί την Ευστοχία και την Ανάκληση για ένα δοθέν προβλεπόμενο κατώφλι πιθανότητας. Αντίθετα με τις καμπύλες ROC οι καμπύλες Ευστοχίας - Ανάκλησης δεν χαρακτηρίζονται από μονοτονία (όπως φαίνεται και στο Σχήμα 6.3 η Ευστοχία μπορεί να αυξάνεται ή να μειώνεται καθώς αυξάνεται η Ανάκληση).



Σχήμα 6.3: *PR Curve για ισορροπημένο σύνολο δεδομένων*

Αντίστοιχα με τις καμπύλες ROC η Περιοχή κάτω από τη Καμπύλη Ευστοχίας - Ανάκλησης (AUPRC) ποσοτικοποιεί την προβλεπτική ικανότητα του μοντέλου αντικατοπτρίζοντας παρόλο αυτά καλύτερα την επίδοση του μοντέλου ειδικά σε προβλήματα τα οποία χαρακτηρίζονται από ανισορροπίας στα δεδομένων. Για δυαδικά προβλήματα ταξινόμησης, σε ένα ισορροπημένο σύνολο δεδομένων ένας τυχαίος ταξινομητής θα έχει AUPRC ίση με 0.5 συνεπώς όπως βλέπουμε και στο Σχήμα 6.3 μοντέλα με καλύτερη προβλεπτική ικανότητα

έχουν υψηλότερες τιμές AUPRC δηλαδή το σχήμα τους βρίσκεται κοντά στην άνω δεξιά γωνία του παραπάνω σχήματος. Τέτοιες τιμές αντιπροσωπεύουν τόσο υψηλές τιμές Precision όσο και Recall, όπου υψηλές τιμές Ανάκλησης σχετίζονται με χαμηλή αναλογία ψευδώς αρνητικών παρατηρήσεων ενώ αντίστοιχα υψηλή Ευστοχία σχετίζεται με χαμηλή αναλογία ψευδώς αρνητικών παρατηρήσεων. Αντίθετα στη περίπτωση του παραδείγματος από την υποενότητα 6.2.1, μοντέλα με ανάλογη απόδοση καταλήγουν σε τιμές AUPRC ίσες με $5/(95+5) = 0.05$ το οποίο μπορεί να γραφτεί στη γενική περίπτωση και ως $|\mathcal{X}_1| / (|\mathcal{X}_0| + |\mathcal{X}_1|)$ όπου $|\mathcal{X}_1|$ και $|\mathcal{X}_0|$ η πληθικότητα των υποσυνόλων δεδομένων από τη μειοψηφική και πλειοψηφική κλάση αντίστοιχα.

Για την προσέγγιση της τιμής της περιοχής κάτω από τη Καμπύλη Ευστοχίας - Ανάκλησης (AUPRC) έχει προταθεί μια πληθώρα μεθόδων [104]. Για την παρούσα διπλωματική εργασία επιλέχθηκε η χρήση της μέσης τιμής της Ευστοχίας, για τον υπολογισμό αυτής.

Ως μέση τιμή της Ευστοχίας (AP) [105] ορίζεται η σταθμισμένη μέση τιμή της Ευστοχίας στα διάφορα κατώφλια ταξινόμησης, χρησιμοποιώντας ως βάρη την αύξηση της τιμής της Ανάκλησης σε σχέση με το προηγούμενο κατώφλι ταξινόμησης,

$$\mathrm{AP} = \sum_n (R_n - R_{n-1})\, P_n,$$

όπου R_n και P_n είναι η Ανάκληση και η Ευστοχία αντίστοιχα στο ν-ιοστό κατώφλι ταξινόμησης.

6.3 Υλοποίηση

Για την πραγματοποίηση των πειραμάτων επιλέχθηκε η πλατφόρμα Google Colaboratory [106] και συγκεκριμένα ένα session του Google Colab Pro+ καθώς λόγου του όγκου των δεδομένων ήταν απαραίτητη η χρήση υψηλής RAM. Πιο συγκεκριμένα, στο επιλεχθέν session χρησιμοποιήθηκε Python 3.7.13 σε τετραπύρηνο επεξεργαστή με 52 GB RAM και μια Tesla P100 GPU. Οι βιβλιοθήκες που χρησιμοποιήθηκαν για την προ-επεξεργασία, οπτικοποίηση των δεδομένων καθώς και την εκπαίδευση των μοντέλων αποτελούνται από τα Pandas [107], NumPy [108], Matplotlib [109], Seaborn [110], Scikit-learn [111], Imbalanced-learn [112], XGBoost [71], LightGBM [72], CatBoost [73], Tensorflow [113], Hyperopt [114], Optuna [115].

Η ύπαρξη μεγάλης ανισορροπίας ανάμεσα στις δύο κλάσεις ενδιαφέροντος μας οδήγησε στην εστίαση σε μετρικές αξιολόγησης ανθεκτικές στην ανισορροπία των κλάσεων όπως τα Recall, (Precision), F_1 - Score και κυρίως το AP το οποιο όπως προαναφέρθηκε συνοψίζει τη Καμπύλη Ευστοχίας - Ανάκλησης.

Οι έξοδοι των προτεινόμενων μοντέλων είναι είτε ετικέτες κλάσης (0 και 1) είτε αριθμοί ανάμεσα στο μηδέν και ένα, οι οποίοι αντιπροσωπεύουν τη πιθανότητα μια κίνηση δικτύου να είναι στη πραγματικότητα κακόβουλη. Χρησιμοποιώντας ένα επιλεγέν κατώφλι πιθανότητας, η κάθε κίνηση δικτύου ταξινομείται ως επίθεση ή όχι, ανάλογα με το αν η πιθανότητα είναι μεγαλύτερη από το επιλεγμένο κατώφλι. Το κατώφλι επιλέγεται με τρόπο τέτοιο ώστε να μεγιστοποιούνται τόσο οι τιμές του Precision όσο και του Recall. Αξίζει να σημειωθεί ότι ως πρόβλημα η ανίχνευση εισβολής σε δεδομένα δικτύου η μη ανίχνευση μιας πραγματικής

επίθεσης, δηλαδή η εσφαλμένα ταξινόμηση μιας επιθετικής κίνησης ως φυσιολογική, αποτελεί σημαντικότερο λάθος έναντι της εσφαλμένης ταξινόμησης μιας φυσιολογικής κίνησης ως επίθεση καθώς η δεύτερη μπορεί να έχει ως αποτέλεσμα μια μικρή και παροδική δυσκολία για το τελικό χρήστη ενώ αντιθέτως η μη ανίχνευση μιας επίθεσης δύναται να οδηγήσει σε καταστροφικά αποτελέσματα για τα θύματα. Ως αποτέλεσμα περαιτέρω προσοχή δόθηκε και στη μετρική της Ανάκλησης.

Αρχικά στο πλαίσιο των μοντέλων μιας κλάσης χρησιμοποιήθηκαν ως baseline δύο μοντέλα απλά αλλά ταυτόχρονα πολύ ισχυρά, ένα Δέντρο Απομόνωσης και μια Μηχανή Διανυσμάτων Υποστήριξης μιας Κλάσης. Για τον καθορισμό των σημαντικότερων υπερπαραμέτρων των μοντέλων χρησιμοποιήθηκε Τυχαία Αναζήτηση σε ένα εύρος από λογικές τιμές. Ακολούθως επιλέχθηκε η χρήση διαφόρων ειδών Αυτοκωδικοποιήτων με σκοπό τη μοντελοποίηση της φυσιολογικής κλάσης. Οι Αυτοκωδικοποιητές εκπαιδεύτηκαν χωρίς επίβλεψη χρησιμοποιώντας αποκλειστικά παρατηρήσεις από την πλειοψηφική κλάση σε μια προσπάθεια εκμάθησης κάλων απεικονίσεων στο λανθάνων χώρο για αυτή.

Στη συνέχεια χρησιμοποιήθηκαν Μοντέλα Ενισχυτικής Κλίσης με σκοπό την δημιουργία μοντέλων εύρεσης ανωμαλιών μέσω χρήσης επιβλεπόμενης μάθησης. Για το βέλτιστο εξ αυτών μελετήθηκε η επίδραση της χρήση τεχνικών δειγματοληψίας στην τελική απόδοση του μοντέλου.Οι τεχνικές δειγματοληψίας εφαρμόστηκαν τόσο για τη μειοψηφική όσο και τη πλειοψηφική κλάση. Επιπροσθέτως το σύνολο δεδομένων εκπαίδευσης εμπλουτίστηκε με το παραγόμενο διάνυσμα σφάλματος ανακατασκευής όπως αυτό προέκυψε έπειτα από την εκπαίδευση και τροφοδότηση ενός Αυτοκωδικοποιητή με τα δεδομένα εκπαίδευσης. Αποτέλεσμα αυτού ήταν η εκπαίδευση Μοντέλων Ενισχυτικής Κλίσης με ημί-επιβλεπόμενο τρόπο. Τέλος για όλα τα Μοντέλα Ενισχυτικής Κλίσης επιλέχθηκε ο καθορισμός των υπερπαραμέτρων τους μέσω της χρήσης Μπεϋζιανής Βελτιστοποίησης και πιο συγκεκριμένα του αλγορίθμου TPE [91].

6.3.1 Δομή Προτεινόμενων Μοντέλων

6.3.1.1 Μοντέλα μιας Κλάσης

- **Μοντέλα Βάσης**

Δύο μοντέλα μικρότερης πολυπλοκότητας, εκπαιδεύτηκαν χρησιμοποιώντας αποκλειστικά παρατηρήσεις προερχόμενες από την κλάση των φυσιολογικών κινήσεων δικτύου με σκοπό τη χρήση τους ως μοντέλα βάσης τέτοια ώστε να χρησιμοποιηθούν για τη σύγκριση των αποτελεσμάτων τους με μοντέλα μεγαλύτερης πολυπλοκότητας. Τα δεδομένα εισόδου κανονικοποιήθηκαν χρησιμοποιώντας Standard Scaler. Τα μοντέλα τα οποία επιλέχθηκαν για αυτό το σκοπό είναι:

(α') **Isolation Forests**
Για το μοντέλο επιλέχθηκε, έπειτα από Τυχαία Αναζήτηση, η χρήση 100 εκτιμητών βάσης για τον ταξινομητή συνόλου ενώ το ποσοστό ακραίων τιμών στο σύνολο δεδομένων καθορίστηκε ίσο με το λόγο μειοψηφικής κλάσης έναντι του μεγέθους του συνόλου δεδομένων δηλαδή 17%.

(β') **One-Class SVM**

Για το μοντέλο επιλέχθηκε, έπειτα από Τυχαία Αναζήτηση, η χρήση RBF kernel ενώ το ανώτατο ποσοστό από σφάλματα κατά την εκπαίδευση και ταυτόχρονα κατώτερο ποσοστό support vectors καθορίστηκε ίσο με 17%.

Έχοντας εκπαιδεύσει τα μοντέλα στη κλάση των φυσιολογικών κινήσεων δικτύου στη συνέχεια κατά τη διαδικασία της επικύρωσης τα μοντέλα αξιολογήθηκαν στην ικανότητα διαχωρισμού των δύο κλάσεων.

- **Αυτοκωδικοποιητές**

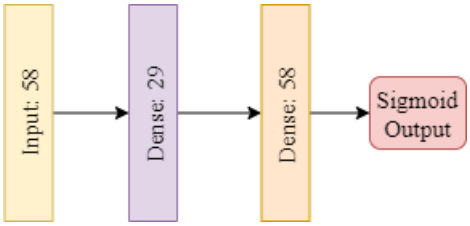
Οι Αυτοκωδικοποιητές εκπαιδεύτηκαν εξίσου χωρίς επίβλεψη χρησιμοποιώντας αποκλειστικά φυσιολογικές κινήσεις δικτύου. Πιο συγκεκριμένα αφου τα δεδομένα εκπαίδευσης κανονικοποιήθηκαν μέσω της χρήσης Min-Max Scaler, κάθε Αυτοκωδικοποιητής εκπαιδεύτηκε με τρόπο τέτοιο ώστε να μάθει να ανακατασκευάζει τα δεδομένα εισόδου από το σύνολο εκπαίδευσης. Έχοντας εκπαιδευτεί μόνο σε φυσιολογικά δεδομένα περιμένουμε ο Αυτοκωδικοποιητής να επιστρέφει χαμηλές τιμές για το σφάλμα ανακατασκευής όταν αυτός τροφοδοτείται με φυσιολογικές κινήσεις δικτύου ενώ αντίθετα για επιθέσεις των οποίων την κατανομή δεν έχει μάθει υψηλές τιμές. Με αυτό το τρόπο, μετρώντας το σφάλμα ανακατασκευής και συγκρίνοντας το με κάποιο προκαθορισμένο κατώφλι μπορούμε να ταξινομήσουμε τις παρατηρήσεις ως φυσιολογικές ή όχι.

Τα βάρη των νευρωνικών δικτύων αρχικοποιήθηκαν μεσώ της χρήσης He Normal, δηλαδή τυχαία δειγματοληψία από μια Tuncated Normal κατανομή με μέση τιμή μηδέν και τυπική απόκλιση $std. = \sqrt{2/\#\text{μονάδων εισόδου στον τανυστή βάρους}}$. Ως συνάρτηση ενεργοποίησης επιλέχθηκε η χρήση της ELU [116] για όλα τα επίπεδα με εξαίρεση το τελευταίο επίπεδο των decoder στο οποίο χρησιμοποιήθηκε η σιγμοειδής συνάρτηση ώστε να ωθήσουμε τις εξόδους των δικτύων στο $[0, 1]$. Το πλήθος των εποχών εκπαίδευσης καθορίστηκε ίσο με 50 χρησιμοποιώντας παρόλο αυτά πρόωρο τερματισμό σε περίπτωση όπου η μετρική υπό παρακολούθηση δεν βελτιωθεί για περισσότερες από 15 εποχές. Τα φυσιολογικά δεδομένα εκπαίδευσης καθώς και το υποσύνολο με φυσιολογικά δεδομένα από το σύνολο επικύρωσης χρησιμοποιήθηκαν σε κάθε εποχή για τον υπολογισμό της απώλειας (Δυαδική Διασταυρούμενη Εντροπία) και την αποφυγή προβλημάτων υπερπροσαρμογής. Επιπροσθέτως μετά το πέρας κάθε εποχής, χρησιμοποιώντας εξολοκλήρου το σύνολο επικύρωσης (φυσιολογικές κινήσεις δικτύου καθώς και επιθέσεις) υπολογίζουμε το AP, μετρική η οποία χρησιμοποιήθηκε και στα πλαίσια του πρόωρου τερματισμού με σκοπό την εύρεση του βέλτιστου μοντέλου. Για τον υπολογισμό του AP χρησιμοποιήθηκε το μέσο τετραγωνικό σφάλμα (MSE) ανάμεσα στα πραγματικά δεδομένα επικύρωσης και τα ανακατασκευασμένα δεδομένα από το σύνολο επικύρωσης ως μέτρο απόφασης σε συνδυασμό με τις ετικέτες. Για υπολογιστικούς λόγους το μέγεθος παρτίδας καθορίστηκε ίσο με 1096. Τέλος για την ελαχιστοποίηση της Δυαδικής Διασταυρούμενης Εντροπίας επιλέχθηκε ο βελτιστοποιητής Adam με learning rate 10^{-3} για το οποίο ορίστηκε πρόγραμμα μείωσης κατά 0.2 σε περίπτωση όπου η μετρική αξιολόγησης (AP) δεν βελτιωθεί για περισσότερες από τρεις εποχές.

Για τον καθορισμό του κατωφλίου για το σφάλμα ανακατασκευής, επιλέγεται μια τιμή για την οποία παρατηρήσεις με μεγαλύτερο σφάλμα ταξινομούνται ως επιθέσεις. Λόγο της

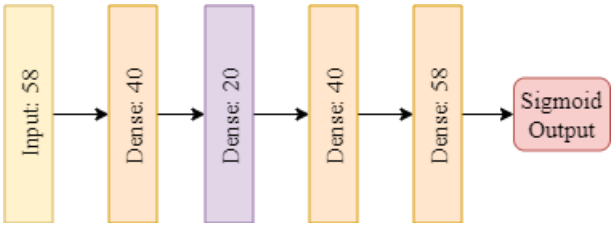
φύσης του προβλήματος κρίθηκε ως μέγιστης σημασίας η δημιουργία μοντέλων τα οποία θα παρουσίαζαν όσο το δυνατό λιγότερα ψευδώς αρνητικά αποτελέσματα για την κλάση των επιθέσεων. Για αυτό το λόγο σε κάθε μοντέλο η τιμή του κατωφλίου καθορίστηκε ώστε να ορίζει πεδίο απόφασης ικανό να επιτύχει, στο σύνολο επικύρωσης, τιμές Ανάκλησης ίσες με 0.9 για την κλάση των Επιθέσεων. Συνολικά εκπαιδεύτηκαν τέσσερα διαφορετικά είδη Αυτοκωδικοποιητών όπως αυτοί παρουσιάζονται λεπτομερώς στη συνέχεια. Η επιλογή τον αρχιτεκτονικών έγινε εμπειρικά δοκιμάζοντας διάφορες υπερπαραμέτρους και παρατηρώντας τη συμπεριφορά των μοντέλων στο σύνολο επικύρωσης.

1. **Απλός Υποπλήρης Αυτοκωδικοποιητής**
   Ένας απλός Υποπλήρης Αυτοκωδικοποιητής (Σχήμα 6.4) με ένα κρυφό επίπεδο με 29 νευρώνες, μειώνοντας με αυτό το τρόπο το πλήθος των χαρακτηριστικών κατά το μισό. Στη πραγματικότητα αποτελεί την απλούστερη μορφή Αυτοκωδικοποιητών με τις λανθάνουσες αναπαραστάσεις να αποτελούνται από το κρυφό επίπεδο.



Σχήμα 6.4: *Αρχιτεκτονική Υποπλήρη Αυτοκωδικοποιητή*

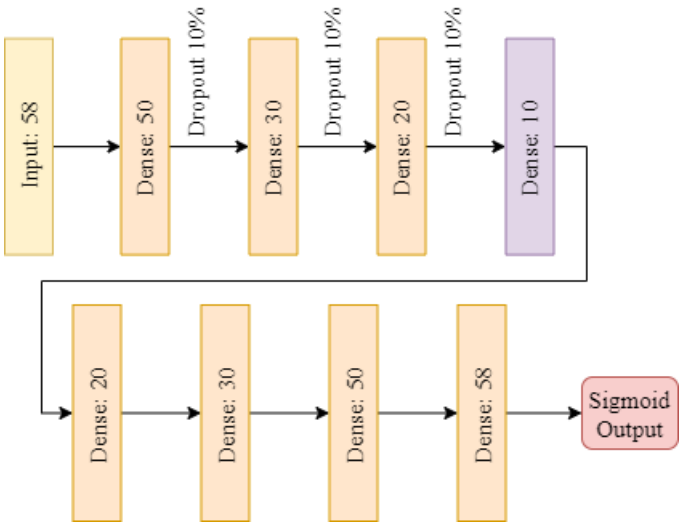
2. **Στοιβαγμένοι Υποπλήρης Αυτοκωδικοποιητές**
   Δύο Αυτοκωδικοποιητές με περισσότερα κρυφά επίπεδα χρησιμοποιήθηκαν με σκοπό την εκμάθηση πιο περίπλοκων απεικονίσεων, ικανές να οδηγήσουν ενδεχομένως και σε καλύτερα αποτελέσματα. Ο πρώτος εκ των δύο χαρακτηρίζεται από ένα κωδικοποιητή με δύο κρυφά επίπεδα 40 και 20 νευρώνων αντίστοιχα (συμπεριλαμβανομένου του bottleneck) και ένα συμμετρικό αποκωδικοποιητή (Σχήμα 6.5). Για τον δεύτερο Στοιβαγμένο Αυτοκωδικοποιητή (Σχήμα 6.6) επι-



Σχήμα 6.5: *Αρχιτεκτονική Στοιβαγμένου Υποπλήρη Αυτοκωδικοποιητή*

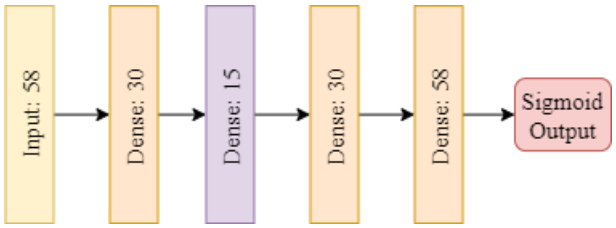
λέχθηκε μια βαθύτερη αρχιτεκτονική με τέσσερα κρυφά επίπεδα διάστασης 50, 30, 20, 10 αντίστοιχα (συμπεριλαμβανομένου του bottleneck) για τον κωδικοποιητή και ένα συμμετρικό αποκωδικοποιητή. Για λόγους αποφυγής προβλημάτων υπερπροσαρμογής και εκμάθησης της ταυτοτικής συνάρτησης χρησιμοποιήθηκε dropout [117], [118] της τάξης τους 10% μετά από κάθε κρυφό επίπεδο του

κωδικοποιητή σε συνδυασμό με περιορισμό μοναδιαίας νόρμας για τα βάρη των αντίστοιχων κρυφών επιπέδων.



Σχήμα 6.6: *Αρχιτεκτονική Βαθύ Στοιβαγμένου Υποπλήρη Αυτοκωδικοποιητή*

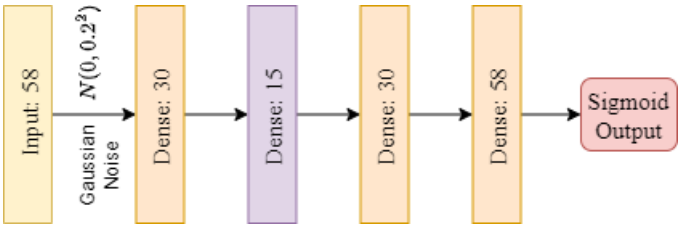
3. **Αραιός Αυτοκωδικοποιητής**
   Χρησιμοποιώντας L1 κανονικοποίηση για τη συνάρτηση απώλειας δημιουργήσαμε ένα Αραιό Αυτοκωδικοποιητή δύο κρυφών επιπέδων διάστασης 30 και 15 για τον κωδικοποιητή (συμπεριλαμβανομένου του bottleneck) και ένα συμμετρικό αποκωδικοποιητή.



Σχήμα 6.7: *Αρχιτεκτονική Αραιού Αυτοκωδικοποιητή*

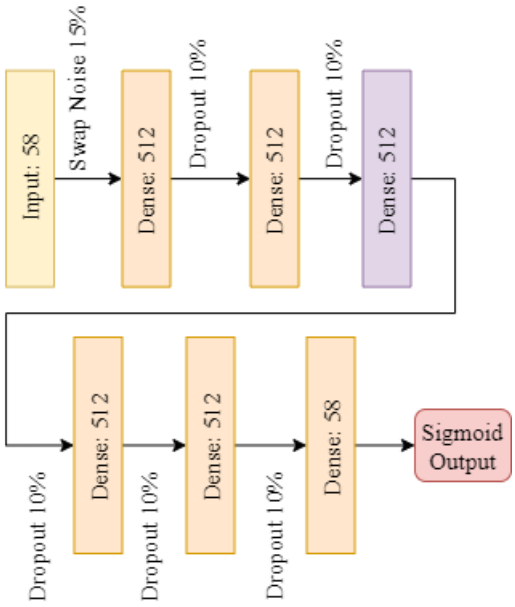
4. **Αυτοκωδικοποιητές Αφαίρεσης Θορύβου**
   Στα πλαίσια των Αυτοκωδικοποιήτων Αφαίρεσης Θορύβου επιλέχθηκαν δύο διαφορετικές τεχνικές διαφθοράς των δεδομένων εισόδου. Αρχικά μέσω της χρήσης θορύβου από μια Γκαουσιανή κατανομή, κεντραρισμένη γύρω από το μηδέν και με τυπική απόκλιση 0.2, διαφθείραμε τα δεδομένα εισόδου προτού αυτά χρησιμοποιηθούν για την εκπαίδευση ενός Αυτοκωδικοποιητή του οποίου η αρχιτεκτονική παρατίθεται στο Σχήμα 6.8.

   Στην πραγματικότητα όμως η χρήση θορύβου από Κανονικές Κατανομές με σκοπό τη διαφθορά των δεδομένων εισόδου δεν αποτελεί βέλτιστη πρακτική καθώς τα δεδομένα μορφής πίνακα αποτελούνται από χαρακτηριστικά τα οποία χαρακτηρίζονται τόσο από διαφορετικές κλίμακες αλλά και χαρακτηριστικά διακριτών



Σχήμα 6.8: *Αρχιτεκτονική Αυτοκωδικοποιητή Αφαίρεσης Θορύβου (Gaussian Noise)*

τιμών για τα οποία η πρόσθεση θορύβου δεν έχει νόημα. Συνεπώς αναζητήθηκαν διαφορετικές τεχνικές οι οποίες θα μπορούσαν να χρησιμοποιηθούν για να μολυνθούν τα δεδομένα εισόδου, γεγονός το οποίο μας οδήγησε στον Michael Jahrer και τη νικηφόρα στρατηγική η οποία εφαρμόστηκε από τον ίδιο στα πλαίσια ενός διαγωνισμού στο Kaggle [119]. Πιο συγκεκριμένα ο Michael Jahrer προτείνει μια τεχνική μόλυνσης των δεδομένων εισόδων, Swap Noise, η οποία βασίζεται στην εναλλαγή των τιμών των χαρακτηριστικών κατά κάποιο προκαθορισμένο ποσοστό, δειγματοληπτώντας με επανάθεση αποκλειστικά από το σύνολο των παρατηρηθέντων τιμών για το εκάστοτε χαρακτηριστικό. Με αυτό το τρόπο επιτυγχάνεται η δειγματοληψία από την εμπειρική κατανομή κάθε χαρακτηριστικού, αποφεύγοντας τη μοντελοποίηση της πραγματικής κατανομής και διατηρώντας ταυτόχρονα χαμηλό υπολογιστικό κόστος. Έχοντας λοιπόν διαφθείρει τα δεδομένα κατά 15% επιλέχθηκε η μοντελοποίηση τους μέσω ενός Υπερπλήρη Αυτοκωδικοποιητή Αφαίρεσης Θορύβου (Σχήμα 6.9) ο οποίος αποτελείτο από ένα κωδικοποιητή τριών κρυφών επιπέδων, συμπεριλαμβανομένου του bottleneck, διάστασης 512 τα οποία ακολουθούνται από τον αντίστοιχο συμμετρικό αποκωδικοποιητή.



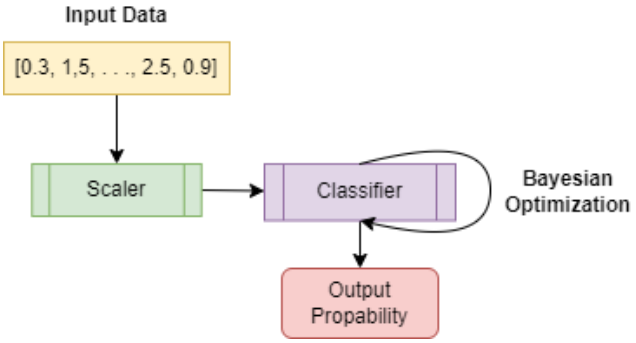
Σχήμα 6.9: *Αρχιτεκτονική Αυτοκωδικοποιητή Αφαίρεσης Θορύβου (Swap Noise)*

Τόσο το ποσοστό επιμόλυνσης των δεδομένων εισόδου όσο και η αρχιτεκτονική του δικτύου προέκυψαν εμπειρικά λαμβάνοντας υπόψη τις προτάσεις του Michael

Jahrer. Φυσικά τα δεδομένα επιμολύνθηκαν αποκλειστικά κατά την διαδικασία εκπαίδευσης των μοντέλων ενώ κατά τη διαδικασία ελέγχου οι Αυτοκωδικοποιητές λαμβάνουν αυτούσιες τις παρατηρήσεις των εκάστοτε συνόλων.

6.3.1.2 Μοντέλα Επιβλεπόμενης Μάθησης

Στα πλαίσια της επιβλεπόμενης ανίχνευσης ακραίων σημείων επιλέχθηκε η χρήση και σύγκριση τριών μοντέλων Ενισχυτικής Κλίσης. Συγκεκριμένα χρησιμοποιήθηκαν τα XGBoost, LightGBM και CatBoost. Πριν από την εκπαίδευση των μοντέλων τα δεδομένα κανονικοποιήθηκαν χρησιμοποιώντας Standard Scaler και στη συνέχεια οι σημαντικότερες υπερπαράμετροι των τριών μοντέλων καθορίστηκαν μέσω Μπεϋζιανής Βελτιστοποίησης.



Σχήμα 6.10: *Γενική Μορφή Επιβλεπόμενης Αρχιτεκτονικής*

Για την επιλογή των υπερπαραμέτρων επιλέχθηκε η βελτιστοποίηση της μετρικής AP έναντι του συνόλου επικύρωσης χρησιμοποιώντας 150 γύρους βελτιστοποίησης. Το πλήθος επαναλήψεων για τον αλγόριθμο TPE επιλέχθηκε με σκοπό την εξασφάλιση σύγκλισης. Στους Πίνακες 6.8, 6.9, 6.10 παραθέτουμε τις προς βελτιστοποίηση υπερπαραμέτρους όπως αυτές καθορίστηκαν για τα τρία μοντέλα. Τέλος χρησιμοποιώντας κατώφλι πιθανότητας 0.5 οι κινήσεις δικτύου ταξινομήθηκαν ως φυσιολογικές ή όχι.

Υπερπαράμετρος	**Περιγραφή**
n_estimators	Πλήθος gradient boosted δέντρων
max_depth	Μέγιστο βάθος δέντρου
gamma	Ελάχιστη μείωσης απώλειας για διαχωρισμό κόμβου
learning_rate	Step size shrinkage
min_child_weight	Ελάχιστο άθροισμα βαρών για ένα νέο κόμβο
subsample	Αναλογία υποδειγματοληψίας δεδομένων εκπαίδευσης
colsample_bytree	Αναλογία υποδειγματοληψίας χαρακτηριστικών
scale_pos_weight	Αναλογία μεταξύ των δύο κλάσεων (Cost Sensitive)

Πίνακας 6.8: *Υπερπαράμετροι - XGBoost*

Επιπροσθέτως εξετάστηκε η επίδραση τεχνικών δειγματοληψίας στην απόδοση του βέλτιστου εκ των τριών προαναφερθέντων μοντέλων.Όπως θα αναλύσουμε και στη συνέχεια, στα πλαίσια της παρούσας διπλωματικής ως βέλτιστος αλγόριθμος καθορίστηκε ο LightGBM ο οποίος και συνδυάστηκε με τις ακόλουθες παραλλαγές:

Υπερπαράμετρος	Περιγραφή
n_estimators	Πλήθος gradient boosted δέντρων
max_depth	Μέγιστο βάθος δέντρου
random_strength	Τυχαιότητα για το σκορ των διαχωρισμών
border_count	Πλήθος διαχωρισμών
lambda_l2	L2 Κανονικοποίηση
scale_pos_weight	Αναλογία μεταξύ των δύο κλάσεων (Cost Sensitive

Πίνακας 6.9: *Υπερπαράμετροι - CatBoost*

Υπερπαράμετρος	Περιγραφή
n_estimators	Πλήθος gradient boosted δέντρων
max_depth	Μέγιστο βάθος δέντρου
num_leaves	Μέγιστο πλήθος φύλλων ανά δέντρου
subsample	Αναλογία υποδειγματοληψίας δεδομένων εκπαίδευσης
colsample_bytree	Αναλογία υποδειγματοληψίας χαρακτηριστικών
bagging_freq	Συχνότητα για χρήση Bagging
min_child_samples	Ελάχιστο πλήθος δεδομένων ανά φύλλο
lambda_l1	L1 Κανονικοποίηση
lambda_l2	L2 Κανονικοποίηση
scale_pos_weight	Αναλογία μεταξύ των δύο κλάσεων (Cost Sensitive)

Πίνακας 6.10: *Υπερπαράμετροι - LightGBM*

- **LightGBM με Υποδειγματοληψία Πλειοψηφικής Κλάσης**
  Τυχαία υποδειγματοληψία (Random Undersampling - RUS) πραγματοποιήθηκε στη πλειοψηφική κλάση έτσι ώστε να επιτευχθεί λόγος 0.5 μεταξύ των δύο κλάσεων.

- **LightGBM με SMOTE**
  Χρησιμοποιώντας τη μέθοδο SMOTE με πέντε κοντινότερους γείτονες ενισχύθηκαν όλα τα είδη επιθέσεων των οποίον το πλήθος στο σύνολο εκπαίδευσης ήταν χαμηλότερο από 10000 έτσι ώστε να καθορίσουμε αυτό ως ελάχιστο ανά είδος επίθεσης.

- **LightGBM με Υποδειγματοληψία Πλειοψηφικής Κλάσης και SMOTE**
  Χρησιμοποιώντας αντίστοιχες τεχνικές όπως αυτές περιγράφησαν παραπάνω συνδυάσαμε τις δύο μεθόδους.

- **LightGBM με PCA και Υποδειγματοληψία Πλειοψηφικής Κλάσης και SMOTE**
  Τα δεδομένα μετασχηματίστηκαν σε κύριες συνιστώσες εκ των οποίων επιλέχθηκε η χρήση των πρώτων 29, καθώς αυτές περιείχαν το 99% της συνολικής διακύμανσης του συνόλου εκπαίδευσης. Στη συνέχεια τυχαία υποδειγματοληψία πραγματοποιήθηκε στη πλειοψηφική κλάση έτσι ώστε να επιτευχθεί λόγος 0.5 μεταξύ των δύο κλάσεων ενώ έπειτα μέσω χρήσης SMOTE με πέντε κοντινότερους γείτονες η συνθήκη ισορροπίας αποκαταστάθηκε μεταξύ των δύο κλάσεων.

- **LightGBM με CTGAN**
  Η μειοψηφική κλάση των δεδομένων εκπαίδευσης εμπλουτίστηκε κατά 50% με συνθετικά δεδομένα παραχθέντα από ένα CTGAN μοντέλο το οποίο για υπολογιστικούς

λόγους εκπαιδεύτηκε χρησιμοποιώντας αποκλειστικά δεδομένα από τη μειοψηφική κλάση. Για την αξιολόγηση των παραχθέντων δεδομένων από το CTGAN μοντέλο χρησιμοποιήθηκαν τόσο στατιστικά μέτρα όσο και τεχνικές μηχανικής μάθησης. Πιο συγκεκριμένα χρησιμοποιήθηκε ο έλεγχος KS με σκοπό τον έλεγχο της υπόθεσης ότι τόσο τα συνθετικά όσο και τα πραγματικά δεδομένα προέρχονται από την ίδια κατανομή. Επιπροσθέτως χρησιμοποιήθηκαν δύο μοντέλα μηχανικής μάθησης (Logistic Regression, SVC) σε μια προσπάθεια διάκρισης εντός συνόλου των δεδομένων, τις πραγματικές από τις συνθετικές παρατηρήσεις. Στο Πίνακα 6.11 παραθέτουμε τα τελικά αποτελέσματα της αξιολόγησης.

	Statistical Metrics	**Detection Metrics**	
	KS-Test	**Logistic Regr.**	**SVC**
CTGAN	0.84	0.77	0.79

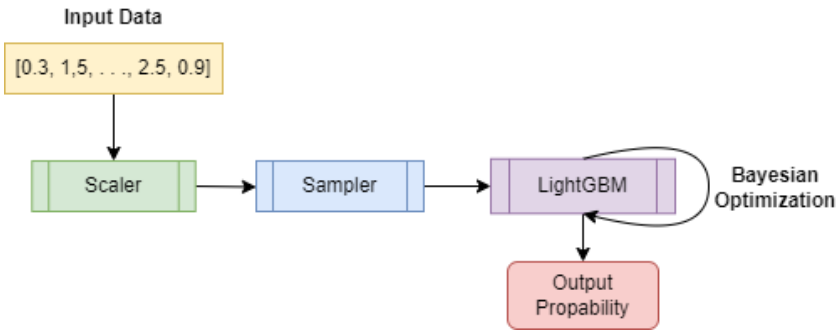
Πίνακας 6.11: *Αξιολόγηση συνθετικών δεδομένων - CTGAN*

Για τον έλεγχο KS παρατίθεται ο μέσος όρος των επιμέρους p-values για κάθε μεταβλητή ενώ για τις μετρικές μηχανικής μάθησης παραθέτουμε τη τιμή του 1 − ROC AUC όπως αυτή έχει υπολογιστεί έπειτα από 3-folds διασταυρούμενη επικύρωση. Είναι προφανές ότι για όλες τις μεταβλητές οι ιδανικές τιμές, στα πλαίσια του παρόντος προβλήματος, καθορίζονται ως αυτές οι οποίες τοποθετούνται κοντά στη μονάδα. Κάτι τέτοιο υποδεικνύει αρχικά ότι οι επιμέρους KS έλεγχοι έχουν κατά μέσο όρο υψηλά p-values με αποτέλεσμα να μην απορρίπτεται η μηδενική υπόθεση του ελέγχου για όμοιες πληθυσμιακές κατανομές στα χαρακτηριστικά των πραγματικών και συνθετικών δεδομένων. Επιπροσθέτως σε ότι αφορά τα μοντέλα μηχανικής μάθησης, τιμές κοντά στη μονάδα αποτυπώνουν την αδυναμία τον μοντέλων να διακρίνουν ανάμεσα στις δύο κλάσης (πραγματικά & συνθετικά δεδομένα) με επιτυχία. Αυτή η αδυναμία ταξινόμησης των δεδομένων στις σωστές κλάσης αποτελεί την εύλογη απόδειξη ότι τα παραχθέντα δεδομένα είναι πράγματι αξιόπιστα.

Κατά την εκπαίδευση των μοντέλων στα νέα σύνολα δεδομένων οι υπερπαράμετροι του εκάστοτε LightGBM επιλέχθηκαν μέσω της χρήσης Μπεϋζιανής Βελτιστοποίησης όπως αυτή σκιαγραφήθηκε παραπάνω. Τέλος στο Σχήμα 6.11 παραθέτουμε τη γενική μορφή της διαδικασίας εκπαίδευσης μοντέλων επιβλεπόμενης μάθησης λαμβάνοντας υπόψη και τις τεχνικές δειγματοληψίας.

6.3.1.3 Μοντέλα Ημί-Επιβλεπόμενης Μάθησης

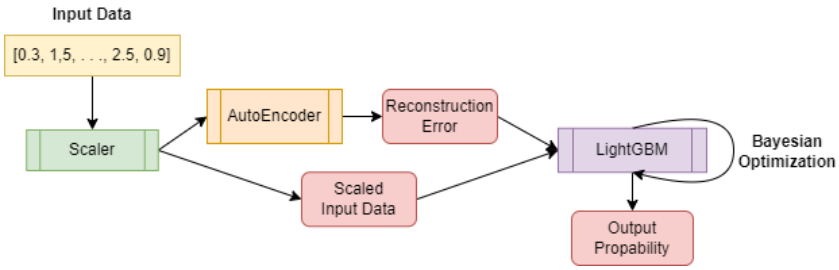
Στα πλαίσια της ημι-επιβλεπόμενης μάθησης επιλέχθηκε ο εμπλουτισμός των δεδομένων εκπαίδευσης με έναν επιπλέον χαρακτηριστικό. Πιο συγκεκριμένα εκπαιδεύοντας ένα Αυτοκωδικοποιητή στο σύνολο εκπαίδευσης, εξάγουμε το σφάλμα ανακατασκευής για το σύνολο των δεδομένων και το χρησιμοποιούμε ως ξεχωριστό χαρακτηριστικό με σκοπό την επαύξηση του συνόλου δεδομένων. Στα πλαίσια των πειραμάτων εφαρμόστηκε μια πληθώρα αρχιτεκτονικών για τον Αυτοκωδικοποιητή παρόλο αυτά επιλέχθηκε η χρήση ενός απλού Υποπλήρη Αυτοκωδικοποιητή, με ένα κρυφό επίπεδο διάστασης 33 μειώνοντας τα αρχικά



Σχήμα 6.11: *Γενική Μορφή Επιβλεπόμενης Αρχιτεκτονικής με Τεχνικές Δειγματοληψίας*

χαρακτηριστικά στα μισά. Αυτή η επιλογή έγινε καθώς τα αποτελέσματα πιο σύνθετων αρχιτεκτονικών δεν πρόσφεραν σημαντικές βελτιώσεις λαμβάνοντας υπόψη και το υπολογιστικό κόστος. Για την εκπαίδευση του Αυτοκωδικοποιητή χρησιμοποιήθηκε το ίδιο πλαίσιο όπως αυτό περιγράφηκε παραπάνω στην ενότητα για τα Μοντέλα μιας Κλάσης ενώ τα δεδομένα κανονικοποιήθηκαν χρησιμοποιώντας τον Min-Max Scaler.

Στη συνέχεια εκπαιδεύσαμε και αξιολογήσαμε ένα μοντέλο LightGBM στο εμπλουτισμένο σύνολο δεδομένων. Αφορμή για την επιλογή του συγκεκριμένου μοντέλου υπήρξε το γεγονός ότι στα πλαίσια της εκπαίδευσης μοντέλων επιβλεπόμενης μάθησης, το παρών μοντέλο κατέληξε σε βέλτιστα αποτελέσματα έναντι των άλλων δύο μοντέλων Ενισχυτικής Κλίσης. Τέλος κατά την εκπαίδευση του μοντέλου πραγματοποιήθηκε επιλογή υπερπαραμέτρων με βάση το σύνολο επικύρωσης χρησιμοποιώντας Μπεϋζιανή Βελτιστοποίηση όπως αυτή περιγράφηκε και στα πλαίσια της επιβλεπόμενης μάθησης. Στο Σχήμα 6.12 παραθέτουμε τη διαδικασία ημι-επιβλεπόμενης εκπαίδευσης του μοντέλου.



Σχήμα 6.12: *Γενική Μορφή Ημι-Επιβλεπόμενης Αρχιτεκτονικής*

Κεφάλαιο 7

Αποτελέσματα

7.1 Μοντέλα Βάσης

Όπως παρατηρούμε στους Πίνακες 7.1 - 7.2, τόσο το One Class SVM όσο και το Isolation Forest αποτυγχάνουν να μοντελοποιήσουν επιτυχώς την κλάση των φυσιολογικών παρατηρήσεων. Τα αποτελέσματα στο Validation Set, αν και καλύτερα από τη χρήση ενός τυχαίος ταξινομητή, δεν μπορούμε σε καμία περίπτωση να θεωρήσουμε ότι είναι ικανοποιητικά. Το παρών αποτελεί μια ένδειξη ότι τα δεδομένα είναι στη πραγματικότητα αρκετά περίπλοκα για να μοντελοποιηθούν σύμφωνα με τις απλοϊκές υποθέσεις των δύο μοντέλων. Επιπλέον η αντίστοιχη υπέρπαράμετρος και των δύο μοντέλων που σχετίζεται με το ποσοστό ακραίων τιμών στο δείγμα, φαίνεται να λειτουργεί με τρόπο περιοριστικό για τα μοντέλα καθώς οφείλει να καθοριστεί πριν από την εκπαίδευση τους.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.4947	0.8281	0.6194	1339023
Attack	0.5115	0.1755	0.2613	1373467
Macro Avg.	0.5031	0.5018	0.4404	2712490
Weighted Avg.	0.5032	0.4976	0.4381	2712490
Accuracy			0.4976	2712490
ROC-AUC			0.5172	2712490
Average Precision			0.544	2712490

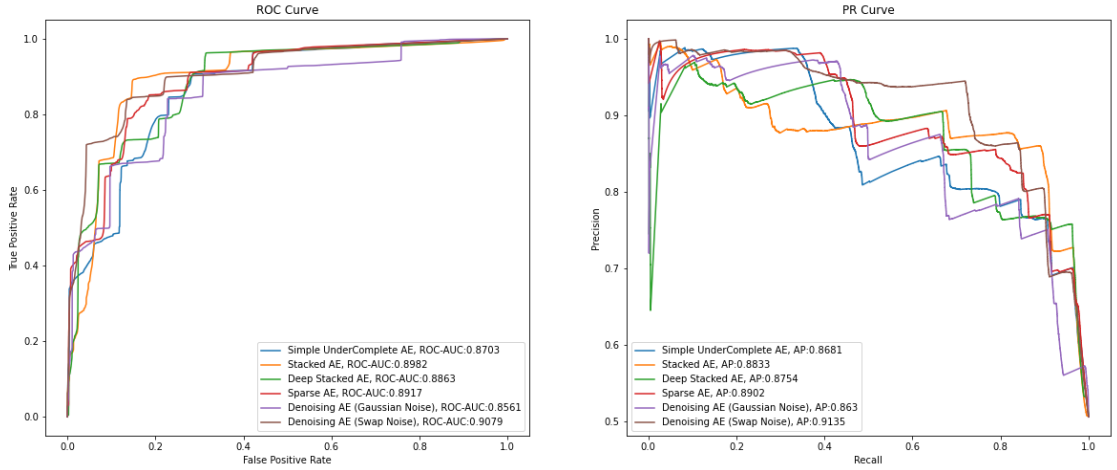
Πίνακας 7.1: *Αξιολόγηση Validation Set - One Class SVM*

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.4631	0.8302	0.5946	1339023
Attack	0.2715	0.0617	0.1005	1373467
Macro Avg.	0.3673	0.4459	0.3475	2712490
Weighted Avg.	0.3661	0.4411	0.3444	2712490
Accuracy			0.4411	2712490
ROC-AUC			0.5268	2712490
Average Precision			0.504	2712490

Πίνακας 7.2: *Αξιολόγηση Validation Set - Isolation Forest*

7.2 Αυτοκωδικοποιητές

Σε αυτό το σημείο παρουσιάζουμε την επίδοση των Αυτοκωδικοποιητών, σύμφωνα με τις προτεινόμενες αρχιτεκτονικές όπως αυτές παρουσιάστηκαν σε προηγούμενα κεφάλαια. Χρησιμοποιώντας τις εξόδους των εκάστοτε μοντέλων είμαστε στη θέση να δημιουργήσουμε ROC Curves καθώς και PR Curves (Σχήμα 7.1) για κάθε μοντέλο έτσι ώστε να αξιολογήσουμε τη γενική απόδοση των μοντέλων χωρίς να λαμβάνουμε υπόψη κάποιο συγκεκριμένο κατώφλι.



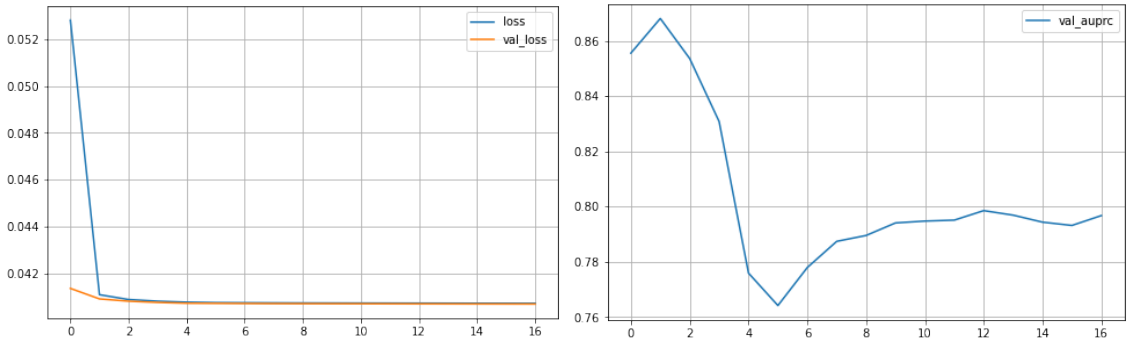
Σχήμα 7.1: *Αυτοκωδικοποιητές - ROC & PR Curves*

Όπως παρατηρούμε η εκπαίδευση πολυπλοκότερων μοντέλων μιας κλάσης φαίνεται να έχει επιτύχει αξιοπρεπή αποτελέσματα στο σύνολο επικύρωσης. Σε γενικές γραμμές είμαστε στη θέση να θεωρήσουμε ότι η χρήση Αυτοκωδικοποιήτων με σκοπό την εκμάθηση λανθανουσών αναπαραστάσεων για τα δεδομένα δικτύου από τη φυσιολογική κλάση έχει στεφθεί με επιτυχία. Με άλλα λόγια η βασική υπόθεση της μεθόδου, η οποία εφαρμόστηκε με σκοπό την εύρεση ανωμαλιών, φαίνεται να ικανοποιείται κάτι το οποίο στη πράξη μεταφράζεται στο γεγονός ότι έχοντας εκπαιδευτεί στην ανακατασκευή φυσιολογικών κινήσεων δικτύου, οι Αυτοκωδικωδικοποιητές αποτυγχάνουν στην ανακατασκευή παρατηρήσεων οι οποίες προέρχονται από άλλες κατανομές κάτι το οποίο οδηγεί και στον εντοπισμό τους.

Πιο συγκεκριμένα, ο Αυτοκωδικοποιητής Αφαίρεσης Θορύβου κατά την εκπαίδευση του οποίου τα δεδομένα εισόδου διαφθείρθηκαν μέσω χρήσης της τεχνικής Swap Noise φαίνεται να καταλήγει στα καλύτερα αποτελέσματα στο σύνολο επικύρωσης παρουσιάζοντας ROC AUC και AP ίσα με 0.9079 και 0.9135 αντίστοιχα. Αντίθετα η χρήση θορύβου από Γκαουσιανές κατανομές οδηγεί στα χειρότερα αποτελέσματα κάτι το οποίο αναδεικνύει τις προβληματικές της συγκεκριμένης μεθόδου διαφθοράς των δεδομένων εισόδου. Επιπροσθέτως παρατηρούμε ότι η χρήση βαθύτερων αρχιτεκτονικών δεν οδήγησε σε καλύτερα αποτελέσματα, ενώ αντίθετα ρηχά δίκτυα καθώς και η χρήση μεθόδων κανονικοποίησης αποτελούν σύμφωνα και με τα αποτελέσματα βέλτιστες πρακτικές. Στη συνέχεια παρατίθενται λεπτομέρειες σχετικά με την εκπαίδευση, το καθορισμό του κατωφλιού για το σφάλμα ανακατασκευής αλλά και την επίδοση των προτεινόμενων αρχιτεκτονικών.

- **Απλός Υποπλήρης Αυτοκωδικοποιητής**

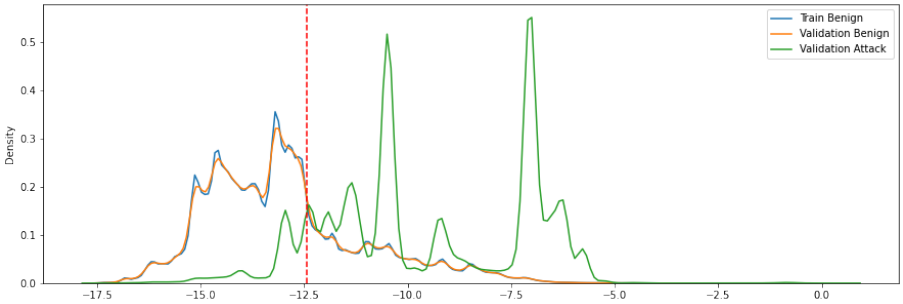
Ξεκινώντας από τον Απλό Υποπλήρη Αυτοκωδικοποιητή, στο Σχήμα 7.2 παρατίθενται τόσο η εξέλιξη της συνάρτησης απώλειας όσο και της μετρικής AP καθώς η τελευταία χρησιμοποιήθηκε με σκοπό τον πρόωρο τερματισμό της εκπαίδευσης αλλά και το προγραμματισμό της μείωσης του learning rate.



(α') *Συνάρτηση Απώλειας ανά Εποχή* (β') *AP ανά Εποχή*

Σχήμα 7.2: *Εκπαίδευση Απλού Υποπλήρη Αυτοκωδικοποιητή*

Παρατηρούμε ότι η απώλεια στο σύνολο επικύρωσης είναι κοντά στην απώλεια του σύνολο εκπαίδευσης συνεπώς έχουμε αποφύγει προβλήματα υπερπροσαρμογής ενώ από τη δεύτερη κιόλας εποχή το μοντέλο επιτύχει τη βέλτιστη απόδοση του στο σύνολο επικύρωσης. Όσο αφορά το σφάλμα ανακατασκευής των δεδομένων (Σχήμα 7.3), παρατηρούμε ότι υπάρχει επικάλυψη μεταξύ των κατανομών. Παρόλο αυτά η επιλογή του κατωφλίου φαίνεται επιτυγχάνει το διαχωρισμό μεταξύ των περιοχών υψηλής πυκνότητας για την εκάστοτε κλάση. Επιπλέον το σφάλμα ανακατασκευής για τα φυσιολογικά δεδομένα από το σύνολο εκπαίδευσης παρουσιάζει αντίστοιχη συμπεριφορά με τα αντίστοιχα δεδομένα από το σύνολο επικύρωσης.



Σχήμα 7.3: *Σφάλμα Ανακατασκευής - Απλός Υποπλήρης Αυτοκωδικοποιητής*

Όπως προαναφέρθηκε στην υπο-υποενότητα 6.3.1.1 το κατώφλι για το σφάλμα ανακατασκευής καθορίστηκε με απώτερο σκοπό την επίτευξη τιμών Ανάκλησης ίσης με 0.9. Κάτι τέτοιο επιτεύχθηκε για το παρόν μοντέλο λαμβάνοντας ως κατώφλι τη τιμή $e^{-12.42912}$. Τέλος στο Πίνακα 7.3 παρουσιάζεται η συνολική απόδοση του μοντέλου, μέσω της χρήσης του παραπάνω κατωφλίου ανακατασκευής, στο σύνολο επικύρωσης. Το μοντέλο καταφέρνει να διαχωρίσει τα δεδομένα στις δύο κλάσεις παρόλο αυτά παρουσιάζει αρκετές παρατηρήσεις

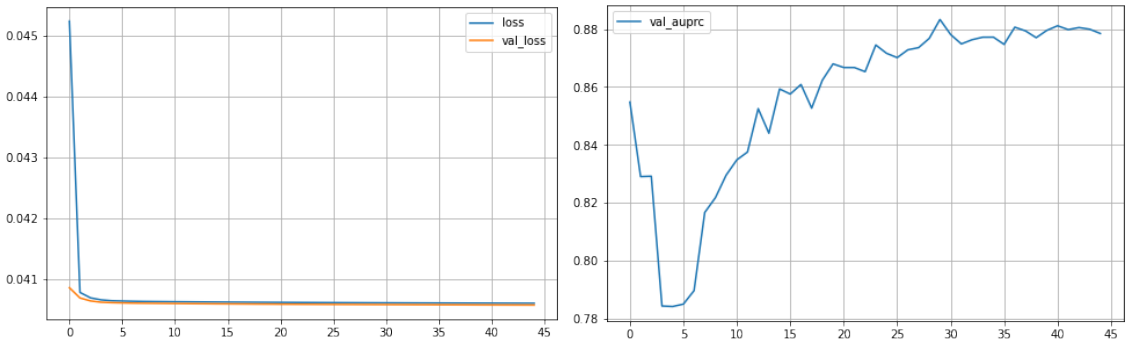
ταξινομημένες ψευδώς ως επιθέσεις, ένα τίμημα άμεσα συσχετισμένο με το καθορισμό υψηλής Ανάκλησης για την κατανομή των επιθέσεων.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8748	0.7170	0.7881	1339023
Attack	0.7654	0.9000	0.8272	1373467
Macro Avg.	0.8201	0.8085	0.8077	2712490
Weighted Avg.	0.8194	0.8097	0.8079	2712490
Accuracy			0.8097	2712490
ROC-AUC			0.8703	2712490
Average Precision			0.8681	2712490

Πίνακας 7.3: *Αξιολόγηση Validation Set - Απλός Υποπλήρης Αυτοκωδικοποιητής*

- **Στοιβαγμένος Υποπλήρης Αυτοκωδικοποιητής**

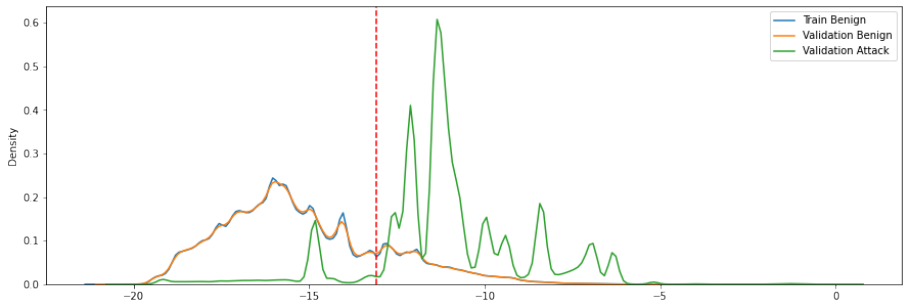
Στο Σχήμα 7.4 παραθέτουμε την εξέλιξη της συνάρτησης απώλειας και της μετρικής AP ανά εποχή εκπαίδευσης. Αντίστοιχα με το προηγούμενο μοντέλο παρατηρούμε ότι η συνάρτηση απώλειας συγκλίνει γρήγορα στο τοπικό ελάχιστο, παρόλο αυτά για την επίτευξη της βέλτιστης τιμής ως προς τη μετρική αξιολόγησης απαιτήθηκαν περισσότερες εποχές.



(α') *Συνάρτηση Απώλειας ανά Εποχή* (β') *AP ανά Εποχή*

Σχήμα 7.4: *Εκπαίδευση Στοιβαγμένου Αυτοκωδικοποιητή*

Όπως φαίνεται στο Σχήμα 7.5 το σφάλμα ανακατασκευής παρουσιάζει και σε αυτό το μοντέλο επικαλύψεις ανάμεσα στις δύο κλάσης, παρόλο αυτά η επιλογή κατωφλιού ίσο με $e^{-13.08758}$ καταφέρνει να διαχωρίσει ικανοποιητικά τις δύο κλάσης.



Σχήμα 7.5: *Σφάλμα Ανακατασκευής - Στοιβαγμένος Αυτοκωδικοποιητής*

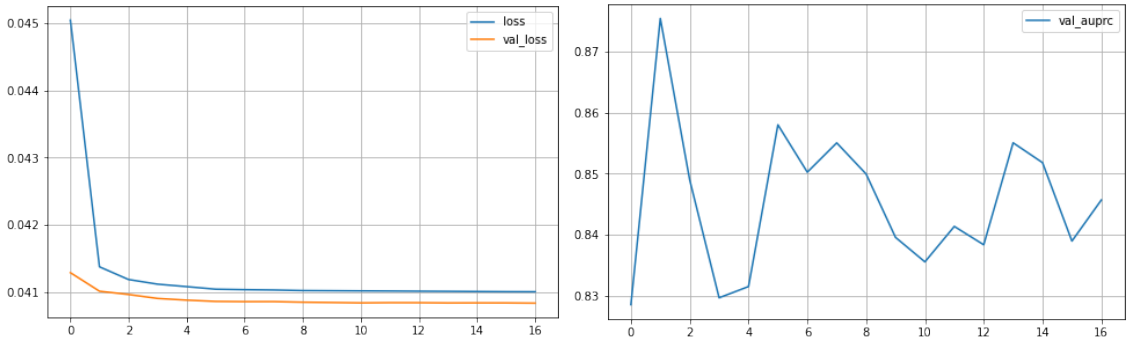
Το τελευταίο αντικατοπτρίζεται και στις υπόλοιπες μετρικές του Πίνακα 7.4. Η μέση τιμή ανάμεσα στις κλάσης των επιμέρους μετρικών είναι κοντά στο 0.86% κάτι το οποίο επιδεικνύει μια ισορροπία ανάμεσα στις ψευδώς αρνητικές και θετικές ταξινομήσεις.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8879	0.8127	0.8486	1339023
Attack	0.8313	0.9000	0.8643	1373467
Macro Avg.	0.8596	0.8563	0.8565	2712490
Weighted Avg.	0.8593	0.8569	0.8566	2712490
Accuracy			0.8569	2712490
ROC-AUC			0.8982	2712490
Average Precision			0.8833	2712490

Πίνακας 7.4: *Αξιολόγηση Validation Set - Στοιβαγμένος Αυτοκωδικοποιητής*

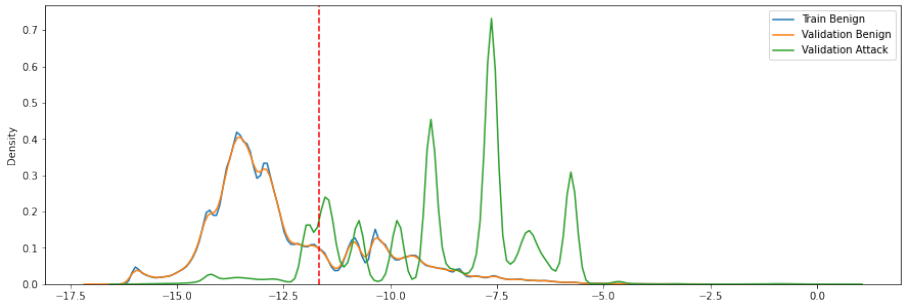
- **Βαθύς Στοιβαγμένος Υποπλήρης Αυτοκωδικοποιητής**

Ως αποτέλεσμα της χρήσης dropout η αποτίμηση της συνάρτησης απώλειας στο σύνολο επικύρωσης είναι τιμές σταθερά χαμηλότερες από τις αντίστοιχες στο σύνολο εκπαίδευσης (Σχήμα 7.6α'). Επιπλέον αντίθετα με τις προηγούμενες δύο περιπτώσεις η τιμή της AP στο σύνολο επικύρωσης φαίνεται να παρουσιάζει ομοιόμορφη κατανομή (Σχήμα 7.6β').



(α') *Συνάρτηση Απώλειας ανά Εποχή* (β') *AP ανά Εποχή*

Σχήμα 7.6: *Εκπαίδευση Βαθύ Στοιβαγμένου Αυτοκωδικοποιητή*



Σχήμα 7.7: *Σφάλμα Ανακατασκευής - Βαθύς Στοιβαγμένος Αυτοκωδικοποιητής*

Όπως παρατηρούμε στο Σχήμα 7.7 το σφάλμα ανακατασκευής των δεδομένων παρουσιάζει σημαντικές επικαλύψεις ανάμεσα στις δύο κλάσης. Ως αποτέλεσμα η επιλογή ενός

ικανοποιητικού κατωφλίου για αυτό φαντάζει δύσκολή.

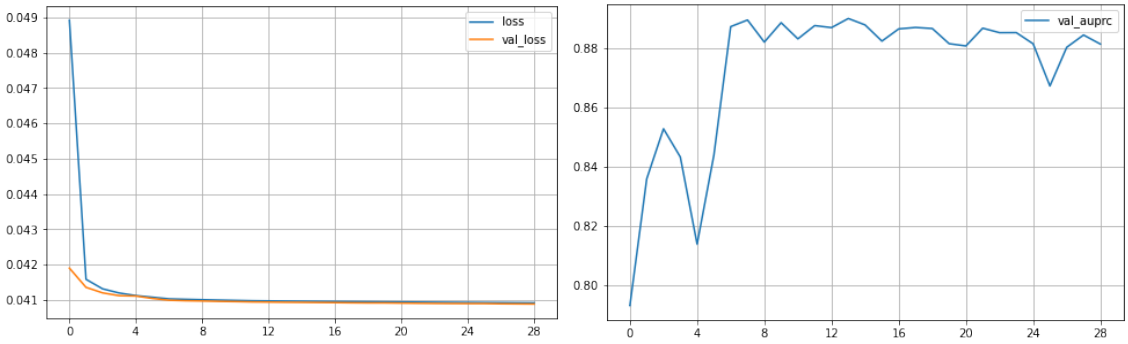
Σύμφωνα με τον επιλεγέντα κανόνα, επιλέχθηκε για κατώφλι η τιμή $e^{-11.67036}$ για την οποία, όπως φαίνεται και στο Σχήμα 7.7, θεωρούμε μεγάλο μέρος από τη κατανομή των φυσιολογικών παρατηρήσεων ως επιθέσεις. Αντίστοιχη είναι και η εικόνα των μετρικών αξιολόγησης (Πίνακας 7.5), οι οποίες και μαρτυρούν την ύπαρξη πολλών ψευδών θετικών ταξινομήσεων.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8750	0.7183	0.7890	1339023
Attack	0.7662	0.9000	0.8277	1373467
Macro Avg.	0.8206	0.8091	0.8083	2712490
Weighted Avg.	0.8199	0.8103	0.8086	2712490
Accuracy			0.8103	2712490
ROC-AUC			0.8863	2712490
Average Precision			0.8754	2712490

Πίνακας 7.5: *Αξιολόγηση Validation Set - Βαθύς Στοιβαγμένος Αυτοκωδικοποιητής*

- **Αραιός Αυτοκωδικοποιητής**

Κατά την εκπαίδευση του Αραιού Αυτοκωδικοποιητή η αποτίμηση της συνάρτησης απώλειας παρουσιάζει παρόμοια συμπεριφορά με άλλα μοντέλα ενώ αντίθετα κατά την αποτίμηση της AP παρατηρούμε ότι χρειάζονται περισσότερες εποχές εκπαίδευσης μέχρις ότου οι τιμές της μετρικής να συγκλίνουν στη βέλτιστη τιμή.

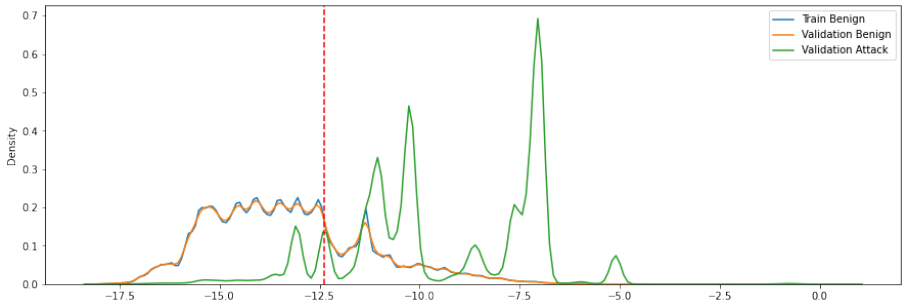


(α') *Συνάρτηση Απώλειας ανά Εποχή* (β') *AP ανά Εποχή*

Σχήμα 7.8: *Εκπαίδευση Αραιού Αυτοκωδικοποιητή*

Η κατανομή του σφάλματος ανακατασκευής για το συγκεκριμένο μοντέλο παρουσιάζει σημαντική διακύμανση μεταξύ των δύο κατανομών, με τη κατανομή των επιθέσεων να παρουσιάζει σημαντικά υψηλότερη πυκνότητα στην ουρά της δεύτερης κατανομής.

Το επιλεγέν κατώφλι αποτιμάται στη τιμή $e^{-12.38057}$ η οποία όπως φαίνεται και στο Σχήμα 7.9 καταφέρνει να διαχωρίσει τις δύο κατανομές σχεδόν στη μέση ξεχωρίζοντας με αυτό το τρόπο δύο αντίστοιχες περιοχές υψηλής πυκνότητας. Παρόλο αυτά είναι προφανές ότι δεν παύει να υπάρχει επικάλυψη ανάμεσα στις δύο κατανομές, γεγονός το οποίο επηρεάζει και τις τιμές των εκάστοτε μετρικών (Πίνακας 7.6).



Σχήμα 7.9: *Σφάλμα Ανακατασκευής - Αραιός Αυτοκωδικοποιητής*

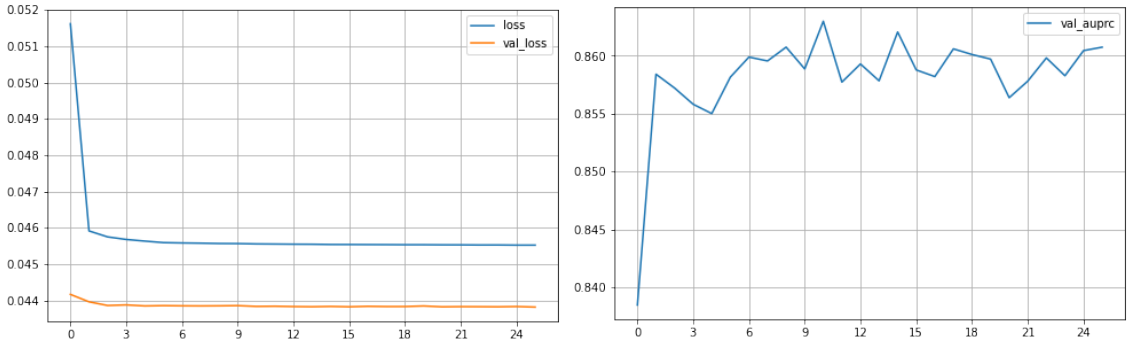
	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8761	0.7251	0.7935	1339023
Attack	0.7706	0.9000	0.8303	1373467
Macro Avg.	0.8233	0.8126	0.8119	2712490
Weighted Avg.	0.8226	0.8137	0.8121	2712490
Accuracy			0.8137	2712490
ROC-AUC			0.8917	2712490
Average Precision			0.8902	2712490

Πίνακας 7.6: *Αξιολόγηση Validation Set - Αραιός Αυτοκωδικοποιητής*

- **Αυτοκωδικοποιητές Αφαίρεσης Θορύβου**

  (Α΄) Διαφθορά μέσω χρήσης Gaussian Noise

Κατά τη διάρκεια της εκπαίδευσης παρατηρούμε ότι το σφάλμα στο σύνολο επικύρωσης είναι διαρκώς χαμηλότερο από το αντίστοιχο σφάλμα στο σύνολο εκπαίδευσης (Σχήμα 7.10α΄), ενώ η τιμή τιμή της μετρικής AP (Σχήμα 7.10β΄) παρουσιάζει αντίστοιχη συμπεριφορά όπως αυτή περιγράφηκε και στα πλαίσια του Αραιού Αυτοκωδικοποιητή.



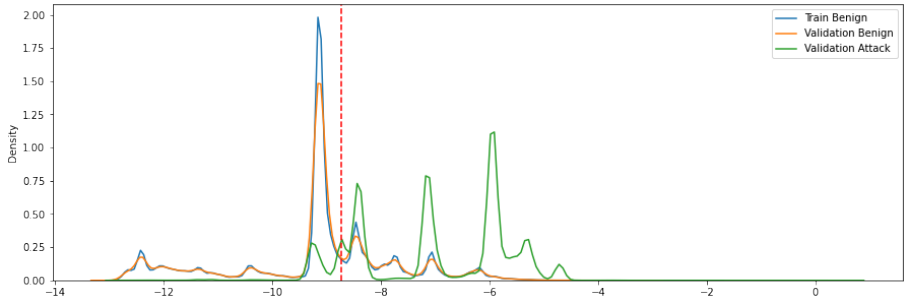
(α΄) *Συνάρτηση Απώλειας ανά Εποχή*

(β΄) *AP ανά Εποχή*

Σχήμα 7.10: *Εκπαίδευση Αυτοκωδικοποιητή Αφαίρεσης Θορύβου (Gaussian Noise)*

Λαμβάνοντας υπόψη την εικόνα του Σχήματος 7.11, είναι προφανές ότι το μοντέλο έχει αποτύχει στο διαχωρισμό ανάμεσα στις δύο κατανομές. Συγκεκριμένα, παρατηρούμε ότι οι κατανομές παρουσιάζουν υψηλές πυκνότητες σε αντίστοιχα σημεία ενώ, με εξαίρεση τις ουρές των κατανομών, η μορφή των κατανομών είναι αρκετά όμοια χωρίς να διακρίνουμε σημαντικές διαφορές ανάμεσα σε αυτές. Ως αποτέλεσμα η επιλογή του κατωφλιού ($e^{-8.720863}$),

οδηγεί στα χειρότερα αποτελέσματα (Πίνακας 7.7) συγκριτικά με τις υπόλοιπες αρχιτεκτονικές.



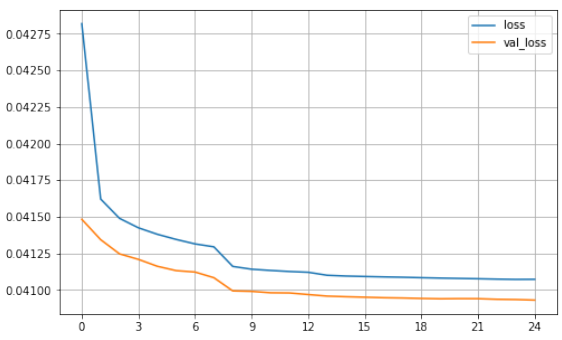
Σχήμα 7.11: *Σφάλμα Ανακατασκευής - Αυτοκωδικοποιητής Αφαίρεσης Θορύβου (Gaussian Noise)*

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8708	0.6914	0.7708	1339023
Attack	0.7495	0.9000	0.8179	1373467
Macro Avg.	0.8101	0.7957	0.7943	2712490
Weighted Avg.	0.8094	0.7970	0.7946	2712490
Accuracy			0.7970	2712490
ROC-AUC			0.8561	2712490
Average Precision			0.8630	2712490

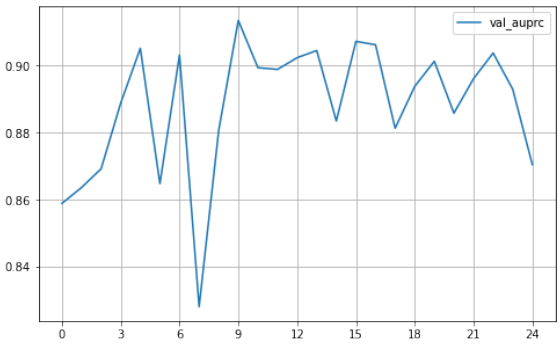
Πίνακας 7.7: *Αξιολόγηση Validation Set - Αυτοκωδικοποιητής Αφαίρεσης Θορύβου (Gaussian Noise)*

(Β΄) Διαφθορά μέσω χρήσης Swap Noise

Κατά την εκπαίδευση του συγκεκριμένου Αυτοκωδικοποιητή παρατηρούμε ότι η συνάρτηση απώλειας (Σχήμα 7.12α΄) προσεγγίζει τη ελάχιστη τιμή της ομαλότερα σε σχέση με τα προηγούμενα μοντέλα ενώ αντίστοιχα παρατηρούμε ότι οι τιμές του συνόλου επικύρωσης είναι σταθερά χαμηλότερες από τις αντίστοιχες τιμές για το σύνολο εκπαίδευσης, κάτι το αναμενόμενο λόγο της χρήσης dropout.



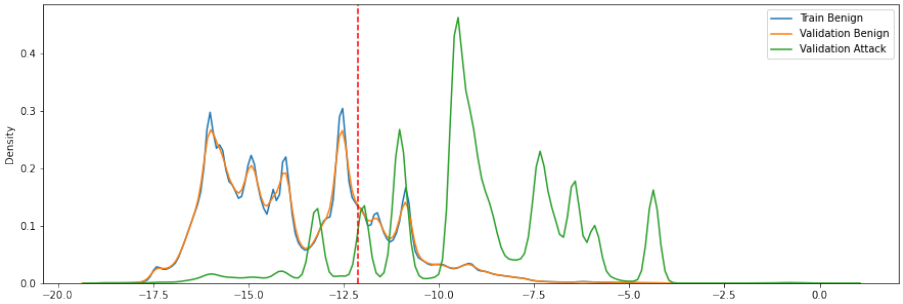
(α΄) *Συνάρτηση Απώλειας ανά Εποχή*



(β΄) *AP ανά Εποχή*

Σχήμα 7.12: *Εκπαίδευση Αυτοκωδικοποιητή Αφαίρεσης Θορύβου (Swap Noise)*

Όσο αφορά τη μετρική AP (Σχήμα 7.12β'), παρατηρούμε ότι ήδη από τη πέμπτη εποχή εκπαίδευσης η τιμή της ξεπερνάει το 0.9 καθορίζοντας με αυτό το τρόπο το μοντέλο ως το βέλτιστο εκ των προτεινόμενων αρχιτεκτονικών Αυτοκωδικοποιήτων.



Σχήμα 7.13: *Σφάλμα Ανακατασκευής - Αυτοκωδικοποιητής Αφαίρεσης Θορύβου (Swap Noise)*

Παρατηρώντας τη κατανομή του σφάλματος αναπροσαρμογής για τα διάφορα σύνολα δεδομένων (Σχήμα 7.13) είναι εμφανές ότι η κλάση των επιθέσεων παρουσιάζει υψηλότερη πυκνότητα σε σημεία όπου η κατανομή των φυσιολογικών κινήσεων παρουσιάζει τοπικά ελάχιστα. Επιπλέον αξίζει να σημειωθεί ότι ο κύριος όγκος της κατανομής καλύπτει τη δεξιά ουρά της κατανομής των φυσιολογικών κινήσεων. Θέτοντας ως κατώφλι τη τιμή $e^{-12.11822}$ επιτυγχάνουμε τη τιμή στόχο για την Ανάκληση στη κλάση ενδιαφέροντος, πληρώνοντας το ανάλογο τίμημα σχετικά με τις ψευδώς θετικές ταξινομήσεις (Πίνακας 7.8). Επιλογές κατωφλίου λιγότερο συντηρητικές (για παράδειγμα e^{-10}) είναι δυνατόν να βελτιώσουν την απόδοση του μοντέλου ως προς τις ψευδώς θετικές ταξινομήσεις μειώνοντας όμως την απόδοση του μοντέλου ως προς τη κλάση ενδιαφέροντος.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8813	0.7617	0.8172	1339023
Attack	0.7948	0.9000	0.8442	1373467
Macro Avg.	0.8381	0.8309	0.8307	2712490
Weighted Avg.	0.8375	0.8317	0.8308	2712490
Accuracy			0.8317	2712490
ROC-AUC			0.9079	2712490
Average Precision			0.9135	2712490

Πίνακας 7.8: *Αξιολόγηση Validation Set - Αυτοκωδικοποιητής Αφαίρεσης Θορύβου (Swap Noise)*

Συνολικά εκ των προαναφερθέντων μοντέλων παρατηρήσαμε ότι ο Αυτοκωδικοποιητής Αφαίρεσης Θορύβου με χρήση Swap Noise κατάφερε να μοντελοποίησει βέλτιστα τις φυσιολογικές κινήσεις δικτύου. Η επιλογή του κατωφλίου για το σφάλμα ανακατασκευής αποδεικνύεται ως μια αρκετά περίπλοκη διαδικασία η οποία ευθύνεται άμεσα για την τελική απόδοση του μοντέλου. Για αυτό το λόγο και συνίσταται η προσεκτική επιλογή αυτού με γνώμονα πάντα την εκάστοτε εφαρμογή των μοντέλων. Εναλλακτικά του κανόνα καθορισμού κατωφλιού όπως αυτός ορίστηκε στα πλαίσια της παρούσας διπλωματικής, ο ερευνητής καλείται να αναλογιστεί διαφορετικούς τρόπους καθορισμού βέλτιστου κατωφλιού όπως για παράδειγμα μέσω χρήσης ποσοστημορίων.

Σύμφωνα με τον επιλεγέντα κανόνα για τον καθορισμό του κατωφλιού ως προς το σφάλμα αναπροσαρμογής, ο Στοιβαγμένος Υποπλήρης Αυτοκωδικοποιητής ευνοείται καθώς παρόλο που δεν έχει συνολικά βέλτιστη απόδοση ως μοντέλο, καταφέρνει να ελαχιστοποιήσει τόσο τις ψευδώς αρνητικές όσο και τις ψευδώς θετικές ταξινομήσεις. Πιο συγκεκριμένα υπερτερεί του αντίστοιχου Αυτοκωδικοποιητή Αφαίρεσης Θορύβου με χρήση Swap Noise με τις σταθμισμένες μετρικές (Recall, Precision, F1-Score) να διαφέρουν κατά περίπου 0.03 υπέρ του Στοιβαγμένου Υποπλήρη Αυτοκωδικοποιητή. Παρόλο αυτά, στη περίπτωση όπου κληθούμε να επιλέξουμε κάποιο εκ των προαναφερθέντων μοντέλων, ο Αυτοκωδικοποιητής Αφαίρεσης Θορύβου με χρήση Swap Noise οφείλει να επιλεγεί καθώς μοντέλα με συνολικά καλή συμπεριφορά οφείλουν να επιλέγονται έναντι μοντέλων τα οποία υπό συγκεκριμένες συνθήκες παρουσιάζουν τα θεμιτά αποτελέσματα μιας και τα τελευταία ενδέχεται να οδηγήσουν σε προβλήματα υπερπροσαρμογής.

7.3 Μοντέλα Ενισχυτικής Κλίσης

Λαμβάνοντας υπόψη και τις δύο κλάσης και εκπαιδεύοντας αντίστοιχα τα Μοντέλα Ενισχυτικής Κλίσης καταλήγουμε σε ικανοποιητικότερα αποτελέσματα. Αυτό είναι εν μέρη αναμενόμενο καθώς ως γνωστόν σε περιπτώσεις όπου έχουμε τη δυνατότητα να εκπαιδεύσουμε μοντέλα σε ποιοτικά και επαρκή δεδομένα τα οποία καλύπτουν πλήρως το προς λύση πρόβλημα τότε τα μοντέλα συμπεριφέρονται βέλτιστα. Στους Πίνακες 7.9, 7.10, 7.11 παραθέτουμε τα αποτελέσματα τριών μοντέλων ενισχυτικής κλίσης, XGBoost, LightGBM και CatBoost. Οι παράμετροι των τριών μοντέλων έχουν καθοριστεί μέσω Μπεϋζιανής Βελτιστοποίησης.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9918	0.9913	0.9915	1339023
Attack	0.9576	0.9599	0.9587	274694
Macro Avg.	0.9747	0.9756	0.9751	1613717
Weighted Avg.	0.9859	0.9859	0.9859	1613717
Accuracy			0.9859	1613717
ROC-AUC			0.9924	1613717
Average Precision			0.9837	1613717

Πίνακας 7.9: *Αξιολόγηση Validation Set - XGBoost*

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9918	0.9918	0.9918	1339023
Attack	0.9600	0.9599	0.9600	274694
Macro Avg.	0.9759	0.9758	0.9759	1613717
Weighted Avg.	0.9864	0.9864	0.9864	1613717
Accuracy			0.9864	1613717
ROC-AUC			0.9928	1613717
Average Precision			0.9843	1613717

Πίνακας 7.10: *Αξιολόγηση Validation Set - LightGBM*

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9916	0.9928	0.9922	1339023
Attack	0.9649	0.9588	0.9618	274694
Macro Avg.	0.9782	0.9758	0.9770	1613717
Weighted Avg.	0.9870	0.9870	0.9870	1613717
Accuracy			0.9870	1613717
ROC-AUC			0.992	1613717
Average Precision			0.9834	1613717

Πίνακας 7.11: *Αξιολόγηση Validation Set - CatBoost*

Παρατηρούμε ότι και τα τρία μοντέλα παρουσιάζουν αρκετά ικανοποιητικά αποτελέσματα αγγίζοντας το 98% ως προς το Average Precision κάτι το οποίο αντικατοπτρίζεται και στις αντίστοιχες μετρικές τόσο για τη κλάση των φυσιολογικών κινήσεων δικτύου αλλά και των επιθέσεων. Όπως έχει προαναφερθεί στόχος είναι η επίτευξη όσο το δυνατόν χαμηλότερων Ψευδών Αρνητικών περιπτώσεων. Συνεπώς εκ των τριών μοντέλων επιλέχθηκε ο LightGBM καθώς, πέραν του χαμηλότερου χρόνου εκπαίδευσης που απαιτεί το μοντέλο, παρουσιάζει και τον ελάχιστο αριθμό από Ψευδών Αρνητικών ταξινομήσεων στο σύνολο Επικύρωσης (Πίνακας 7.12). Αξίζει να σημειωθεί ότι ο CatBoost παρουσιάζει σημαντικά λιγότερες Ψευδής Θετικές ταξινομήσεις παρόλο αυτά λόγο της ιεράρχησις των Ψευδών Αρνητικών ταξινομήσεων υψηλότερα δεν επιλέχθηκε η χρήση του.

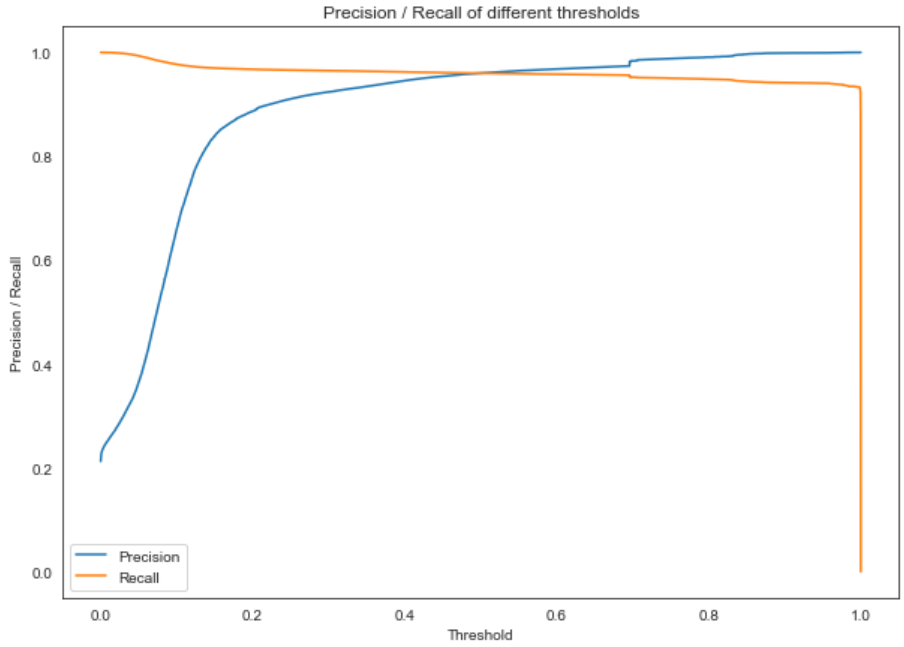
	Ψευδώς Θετικά (%)	**Ψευδώς Αρνητικά (%)**	**Χρόνος Εκπαίδευσης (min.)**
XGBoost	11682 (0.72)	11017 (0.68)	9.06
LightGBM	11024 (0.72)	10975 (0.68)	2.9
CatBoost	11315 (0.70)	9586 (0.59)	19.18

Πίνακας 7.12: *Σύγκριση Μοντέλων Ενισχυτικής Κλίσης*

Τα παραπάνω αποτελέσματα επιτεύχθηκαν έπειτα από καθορισμό των εκάστοτε υπερπαραμέτρων μέσω χρήσης 150 γύρων Μπεϋζιανής Βελτιστοποίησης. Για το επιλεγέν LightGBM μοντέλο οι εν χρήση υπερπαράμετροι παρουσιάζονται στον ακόλουθο Πίνακα 7.13.

Υπερπαράμετρος	**Τιμή**
n_estimators	422
max_depth	15
num_leaves	245
subsample	0.855
colsample_bytree	0.715
bagging_freq	6
min_child_samples	49
lambda_l1	6.528
lambda_l2	0.003
scale_pos_weight	6.004

Πίνακας 7.13: *Επιλεχθέντες Υπερπαράμετροι - LightGBM*



Σχήμα 7.14: *Κατώφλια Απόφασης - LightGBM*

Τέλος στο Σχήμα 7.14 παραθέτουμε το διάγραμμα με τη μεταβολή των Precision και Recall για διαφορετικά κατώφλια απόφασης. Όπως παρατηρούμε η επιλογή κατωφλιού πιθανότητας 0.5 δικαιολογείται πλήρως από τα αποτελέσματα στο σύνολο επικύρωσης καθώς για τιμές κοντά σε αυτό επιτυγχάνεται η από κοινού βελτιστοποίηση των δύο μετρικών.

Έχοντας καθορίσει το LightGBM ως βέλτιστο μοντέλο για το πρόβλημα μας, εφαρμόσαμε τις προαναφερθείσες τεχνικές δειγματοληψίας στο σύνολο εκπαίδευσης με σκοπό την επανεκπαίδευση και καθορισμό νέων σύνολων υπερπαραμέτρων για το LightGBM. Με εξαίρεση το πείραμα κατά το οποίο επιλέχθηκε η χρήση PCA, καταλήγουμε στο συμπέρασμα ότι η χρήση των υπολοίπων μεθόδων επιφέρει αποδόσεις οι οποίες δεν δικαιολογούν τη χρήση τους καθώς τα αποτελέσματα των μοντέλων προσεγγίζουν τα αποτελέσματα της εφαρμογής του LightGBM στα αρχικά δεδομένα χωρίς όμως να τον ξεπερνάνε (Πίνακες 7.14, 7.15, 7.16). Ειδικότερα σημειώνουμε ότι η χρήση PCA οδήγησε σε σημαντικά χειρότερα αποτελέσματα (Πίνακας 7.17), παρόλο που το σύνολο εκπαίδευσης διαθέτει χαρακτηριστικά με ισχυρές ανά δύο συσχετίσεις.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9922	0.9892	0.9907	1339023
Attack	0.9480	0.9621	0.9550	274694
Macro Avg.	0.9701	0.9756	0.9728	1613717
Weighted Avg.	0.9847	0.9846	0.9846	1613717
Accuracy			0.9846	1613717
ROC-AUC			0.9927	1613717
Average Precision			0.9842	1613717

Πίνακας 7.14: *Αξιολόγηση Validation Set - RUS & LightGBM*

	Precision	Recall	F1-Score	Support
Benign	0.9921	0.9894	0.9908	1339023
Attack	0.9492	0.9617	0.9554	274694
Macro Avg.	0.9706	0.9755	0.9731	1613717
Weighted Avg.	0.9848	0.9847	0.9847	1613717
Accuracy			0.9847	1613717
ROC-AUC			0.9927	1613717
Average Precision			0.9842	1613717

Πίνακας 7.15: *Αξιολόγηση Validation Set - SMOTE & LightGBM*

	Precision	Recall	F1-Score	Support
Benign	0.9922	0.9889	0.9905	1339023
Attack	0.9468	0.9619	0.9543	274694
Macro Avg.	0.9695	0.9754	0.9724	1613717
Weighted Avg.	0.9844	0.9843	0.9844	1613717
Accuracy			0.9843	1613717
ROC-AUC			0.9927	1613717
Average Precision			0.9842	1613717

Πίνακας 7.16: *Αξιολόγηση Validation Set - RUS & SMOTE & LightGBM*

	Precision	Recall	F1-Score	Support
Benign	0.9909	0.9940	0.9924	1339023
Attack	0.9702	0.9555	0.9628	274694
Macro Avg.	0.9805	0.9747	0.9776	1613717
Weighted Avg.	0.9874	0.9874	0.9874	1613717
Accuracy			0.9874	1613717
ROC-AUC			0.9906	1613717
Average Precision			0.9812	1613717

Πίνακας 7.17: *Αξιολόγηση Validation Set - PCA & RUS & SMOTE & LightGBM*

	Precision	Recall	F1-Score	Support
Benign	0.9914	0.9933	0.9923	1339023
Attack	0.9668	0.9580	0.9624	274694
Macro Avg.	0.9791	0.9756	0.9774	1613717
Weighted Avg.	0.9872	0.9873	0.9872	1613717
Accuracy			0.9873	1613717
ROC-AUC			0.9928	1613717
Average Precision			0.9842	1613717

Πίνακας 7.18: *Αξιολόγηση Validation Set - CTGAN & LightGBM*

Συνεπώς, καταλήγουμε στο συμπέρασμα ότι η χρήση μεθόδων δειγματοληψίας που στοχεύουν στον εμπλουτισμό της μειοψηφικής κλάσης δεν φαίνεται να καταφέρνουν να προσφέρουν στα εκάστοτε μοντέλα την αναγκαία πληροφορία έτσι ώστε να βελτιώσουν την απόδοση τους. Αντιθέτως, έχουμε ενδείξεις ότι μεγάλο μέρος της πλειοψηφικής κλάσης αποτελείται

από επικαλυπτόμενη πληροφορία κάτι το οποίο αποτυπώνεται στο γεγονός ότι αφαιρώντας το 50% αυτών καταλήγουμε σε επιδόσεις ανάλογες με εκείνες που επιτεύχθηκαν με τη χρήση όλων των παρατηρήσεων. Το παραπάνω μας οδηγεί στο συμπέρασμα ότι κατά την εκπαίδευση παρόμοιων μοντέλων θα μπορούσε να εφαρμοστεί υποδειγματοληψία (τυχαία ή βασισμένη σε clusters) στη πλειοψηφική κλάση με σκοπό τη μείωση του υπολογιστικού κόστους.

Τέλος αξιολογώντας τον εμπλουτισμό του Συνόλου Εκπαίδευσης με το διάνυσμα του Σφάλματος Ανακατασκευής όπως αυτό προέκυψε από τον αντίστοιχο Αυτοκωδικοποιητή καταλήγουμε στο συμπέρασμα ότι η χρήση της μεθόδου δεν είναι σε θέση να βελτιώσει τα τελικά αποτελέσματα σε σημείο όπου να δικαιολογείται η χρήση της (Πίνακας 7.19).

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9922	0.9887	0.9905	1339023
Attack	0.9460	0.9623	0.9541	274694
Macro Avg.	0.9691	0.9755	0.9723	1613717
Weighted Avg.	0.9844	0.9842	0.9843	1613717
Accuracy			0.9842	1613717
ROC-AUC			0.9927	1613717
Average Precision			0.9843	1613717

Πίνακας 7.19: *Αξιολόγηση Validation Set - AE & LightGBM*

7.4 Αξιολόγηση Επιλεγέντος Μοντέλου (Σύνολο Ελέγχου)

Εκ των προαναφερθέντων μοντέλων μιας κλάσης αλλά και τα μοντέλα ενισχυτικής κλίσης, επιλέχθηκε το μοντέλο LightGBM όπως αυτό καθορίστηκε έπειτα από εκπαίδευση χωρίς χρήση μεθόδων δειγματοληψίας και καθορισμό υπερπαραμέτρων μέσω Μπεϋζιανής Βελτιστοποίησης. Το μοντέλο αυτό επιλέχθηκε καθώς, πέραν των καλυτέρων αποτελεσμάτων που παρουσίασε σε σύγκριση με τα μοντέλα μιας κλάσης, επιτυγχάνει τη βέλτιστη τιμή για τη μετρική στόχο (Average Precision) συνδυάζοντας χαμηλό υπολογιστικό κόστος (χρόνος εκπαίδευσης και μέγεθος μοντέλου).

Με σκοπό την περαιτέρω αξιολόγηση του επιλεχθέντος μοντέλου, σε δεδομένα τα οποία δεν χρησιμοποιήθηκαν κατά την διαδικασία εκπαίδευσης και καθορισμού υπερπαραμέτρων, χρησιμοποιήθηκε το κομμάτι σύνολο ελέγχου από το σύνολο δεδομένων CSE-CIC-IDS2018 καθώς και το εξολοκλήρου το σύνολο δεδομένων CIC-IDS2017. Στις ακόλουθες ενότητες αναλύουμε λεπτομερώς τα αποτελέσματα στα δύο προαναφερθέν σύνολα.

7.4.1 CSE-CIC-IDS2018

Το επιλεγέν μοντέλο ενισχυτικής κλίσης παρουσιάζει, στο σύνολο ελέγχου, παρόμοια συμπεριφορά όπως αυτή αναδείχθηκε στα πλαίσια του συνόλου επικύρωσης (Πίνακας 7.20). Πιο συγκεκριμένα, οι μετρικές Precision, Recall και F1-Score επιτυγχάνουν αποδόσεις της τάξης του 0.986 λαμβάνοντας υπόψη τον σταθμισμένο μέσο όρο ανάμεσα στις δύο κλάσης. Για την κλάση των φυσιολογικών κινήσεων δικτύου παρατηρούμε τιμές γύρω από το 0.99 ενώ αντίστοιχα για τη μειοψηφική κλάση των επιθέσεων ο ταξινομητής επιτυγχάνει αποδόσεις

της τάξεως του 0.96 για τις μετρικές ενδιαφέροντος. Συνεπώς, είναι ξεκάθαρο ότι, κατά την εκπαίδευση του μοντέλου αντιμετωπίστηκαν επαρκώς κίνδυνοι σχετικά με την υπερπροσαρμογή του μοντέλου καθώς και προβλήματα σχετικά με την ανισοκατανομή ανάμεσα στις δύο κλάσης.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9918	0.9919	0.9919	1339024
Attack	0.9606	0.9599	0.9603	274693
Macro Avg.	0.9762	0.9759	0.9761	1613717
Weighted Avg.	0.9865	0.9865	0.9865	1613717
Accuracy			0.9865	1613717
ROC-AUC			0.9928	1613717
Average Precision			0.9844	1613717

Πίνακας 7.20: *Αξιολόγηση Test Set - LightGBM*

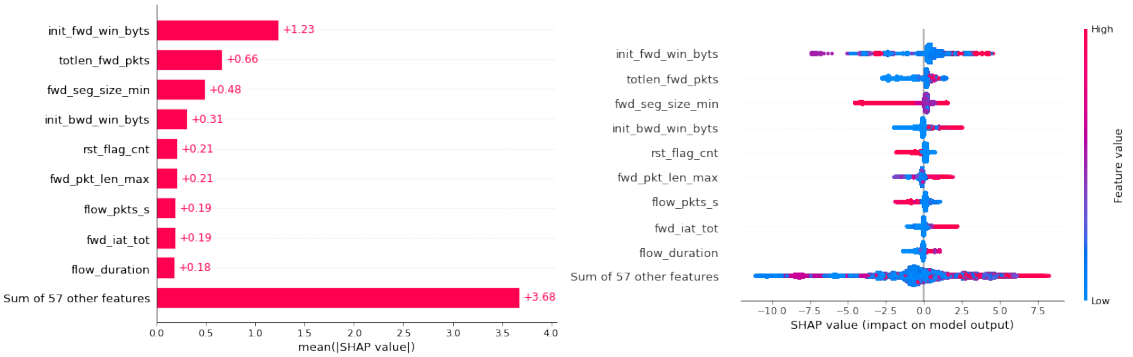
Στη συνέχεια, παραθέτουμε, στο Πίνακα 7.21, την κατανομή των σφαλμάτων στο σύνολο ελέγχου. Όπως παρατηρούμε ένα πολύ μικρό ποσοστό (0.8%) από τις φυσιολογικές κινήσεις δικτύου έχει ταξινομηθεί ψευδώς ως επιθέσεις. Επιπροσθέτως, παρατηρούμε ότι για το σύνολο των επιθέσεων ο κύριος όγκος των εσφαλμένων ταξινομήσεων παρουσιάζεται για επιθέσεις της κατηγορίας Infiltration. Η συγκεκριμένη κατηγορία επιθέσεων φαίνεται να ξεφεύγει από το μοντέλο καθώς σε ποσοστό 68.2% οι επιθέσεις αυτές δεν αναγνωρίζονται από το μοντέλο. Αντίστοιχα, παρατηρούμε ότι, για τις κατηγορίες επιθέσεων Brute Force-Web, Brute Force-XSS, SQL Injection, δεδομένου του μικρού μεγέθους των κατηγοριών, το μοντέλο συμπεριφέρεται ικανοποιητικά, ειδικά αν αναλογιστούμε το γεγονός ότι λόγο του μικρού μεγέθους των κατηγοριών ενδέχεται οι παρατηρήσεις να μην είναι δυνατόν να περιγράψουν πλήρως τα χαρακτηριστικά των εκάστοτε κατανομών. Τέλος αξίζει να σημειωθεί ότι για το 50% από τα είδη των διαθέσιμων επιθέσεων το μοντέλο καταφέρνει να επιτύχει μηδενικές εσφαλμένες ταξινομήσεις.

Τέλος θα χρησιμοποιήσουμε τη μέθοδο SHAP (SHapley Additive exPlanations) με σκοπό την εξερεύνηση των αιτίων πίσω από τις τελικές αποφάσεις του μοντέλου. Η μέθοδος SHAP προτάθηκε το 2017 από τους Lundberg και Lee [120] και βασίζεται στην εύρεση της βέλτιστης θεωρητικής τιμής Shapley. Ο στόχος της μεθόδου είναι να εξηγήσει την πρόβλεψη ενός στιγμιότυπου, υπολογίζοντας τη συμβολή κάθε χαρακτηριστικού στην πρόβλεψη. Η μέθοδος επεξήγησης SHAP υπολογίζει τις τιμές Shapley από τη θεωρία συνασπισμών παιγνίων. Οι τιμές χαρακτηριστικών μιας παρατήρησης λειτουργούν ως παίκτες σε έναν συνασπισμό. Οι τιμές Shapley μας λένε πώς να κατανείμουμε δίκαια τη ¨πληρωμή¨ (τη πρόβλεψη) μεταξύ των χαρακτηριστικών. Αξίζει να σημειωθεί ότι η μέθοδος υπολογίζει τη σημαντικότητα των μεταβλητών λαμβάνοντας υπόψη την αλληλεπίδραση με άλλες μεταβλητές. Για δεδομένα πίνακα ένας παίκτης μπορεί να είναι μια μεμονωμένη τιμή χαρακτηριστικού. Ένας παίκτης μπορεί επίσης να είναι μια ομάδα τιμών χαρακτηριστικών. Για παράδειγμα, για να εξηγηθεί μια εικόνα, τα pixel μπορούν να ομαδοποιηθούν σε superpixel και η πρόβλεψη να κατανεμηθεί μεταξύ τους. Στη συνέχεια παραθέτουμε την επεξήγηση για το επιλεγέν μοντέλο επιλέγοντας να εστιάσουμε στα δέκα σημαντικότερα χαρακτηριστικά. Στο Σχήμα 7.15α′ παρουσιάζονται οι σημαντικότερες μεταβλητές καθώς και οι αντίστοιχες μέσες απόλυτες τιμές Shapley, ε-

Τύπος Κινήσεων Δικτύου	Εσφαλμένες Ταξινομήσεις (%)	Σύνολο
Φυσιολογική Κίνηση Δικτύου	10809 (0.8)	1339024
Infiltration	10958 (68.2)	16064
Brute Force-Web	22 (36.1)	61
DoS attack-Slowloris	14 (1.27)	1099
Botnet	7 (0.02)	28619
Brute Force-XSS	3 (13)	23
DDoS attack-LOIC-HTTP	2 (0.003)	57619
SQL Injection	1 (12.5)	8
FTP-Brute Force	0	19336
SSH-Brute Force	0	18759
DDoS attack-HOIC	0	68601
DDoS attack-LOIC-UDP	0	173
DoS attack-GoldenEye	0	4151
DoS attack-SlowHTTPTest	0	13989
DoS attack-Hulk	0	46191

Πίνακας 7.21: *Κατανομή Εσφαλμένων Ταξινομήσεων (Test - CIC-IDS2018)*

νώ στο Σχήμα 7.15β′ παρατίθενται η κατανομή των τιμών Shapley για τις σημαντικότερες μεταβλητές.



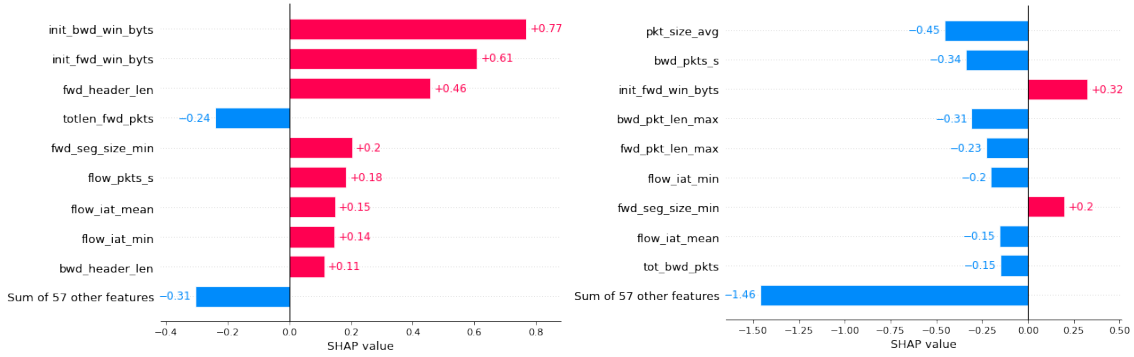
(α′) *Μέση τιμή απολύτων Shapley Values*

(β′) *Shapley Values*

Σχήμα 7.15: *Επεξήγηση Μοντέλου - SHAP*

Παρατηρούμε ότι δύο μεταβλητές οι οποίες σχετίζονται με το πλήθος των bytes τα οποία απεστάλησαν στο αρχικό παράθυρο, τόσο στην εμπρόσθια (init_fwd_win_byts) όσο και στην οπίσθια κατεύθυνση (init_bwd_win_byts), επιδρούν σημαντικά στις τελικές αποφάσεις του μοντέλου. Και για τις δύο μεταβλητές παρατηρούμε ότι υψηλές τιμές αυτών ωθούν προς τη πρόβλεψης μιας επίθεσης. Αντίστοιχη είναι η εικόνα για μεταβλητές που σχετίζονται με τη διάρκεια της ροής (flow_duration), το συνολικό χρόνο ανάμεσα σε δύο πακέτα απεσταλμένα στην εμπρόσθια κατεύθυνση (fwd_iat_tot) και το μέγιστο μέγεθος πακέτων στην εμπρόσθια κατεύθυνση fwd_pkt_len_max. Αντίθετα χαμηλές τιμές ως προς τον αριθμό πακέτων που μεταφέρθηκαν ανά δευτερόλεπτο (flow_pkts_s) και τον αριθμό πακέτων με RST (rst_flag_cnt) οδηγούν σε φυσιολογική ταξινόμηση της κίνησης.

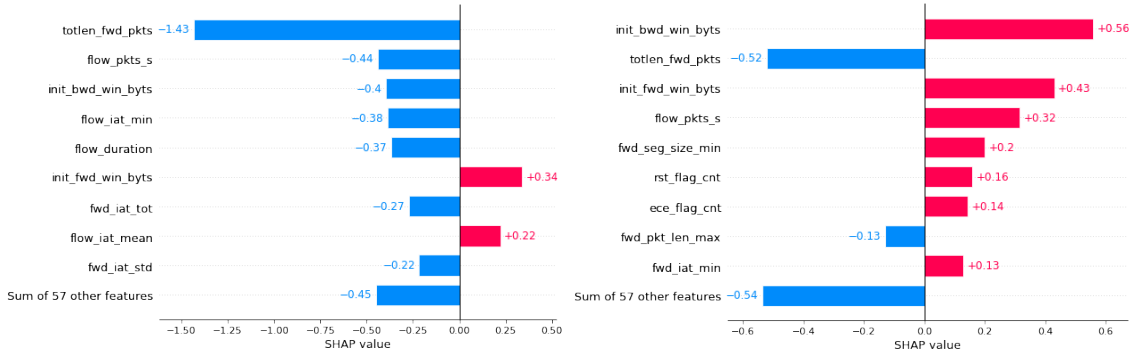
Τέλος, ερευνούμε τους λόγους πίσω από τις εσφαλμένες ταξινομήσεις για τις Φυσιολογικές Κινήσεις Δικτύου (Σχήμα 7.16) και τις Επιθέσεις τύπου Infiltration (Σχήμα 7.17). Για



(α') *Σωστή Ταξινόμηση* (β') *Εσφαλμένη Ταξινόμηση*

Σχήμα 7.16: *Μέθοδος SHAP - Φυσιολογική Κίνηση Δικτύου*

αυτό το λόγο επιλέχθηκαν δύο παρατηρήσεις από τη κάθε ομάδα, εκ των οποίων για τη μια έχει επιτευχθεί σωστή ταξινόμηση ενώ η άλλη έχει ταξινομηθεί εσφαλμένα. Όπως παρατηρούμε, τόσο οι Φυσιολογικές παρατηρήσεις όσο και οι Επιθέσεις παρουσιάζουν σημαντικά διαφορετική συμπεριφορά. Πιο συγκεκριμένα, για τις εσφαλμένες ταξινομήσεις παρατηρούμε ότι στις σημαντικότερες μεταβλητές συγκαταλέγονται χαρακτηριστικά για τα οποία δεν είχαμε ενδείξεις ως προς τη σημαντικότητα τους ενώ για τις σημαντικότερες κοινές μεταβλητές αντίθετες επιδράσεις λαμβάνουν χώρα. Το γεγονός αυτό εξηγεί τους λόγους για τους οποίους το επιλεγέν μοντέλο αδυνατεί να ταξινομήσει ορθώς τις εν λόγω παρατηρήσεις καθώς η επίδραση αυτών διαφέρει σημαντικά.



(α') *Σωστή Ταξινόμηση* (β') *Εσφαλμένη Ταξινόμηση*

Σχήμα 7.17: *Μέθοδος SHAP - Επίθεση (Infiltration)*

7.4.2 CIC-IDS2017

Σε μια προσπάθεια περαιτέρω αξιολόγησης του επιλεγέντος μοντέλου χρησιμοποιούμε το σύνολο CIC-IDS2017. Όπως προαναφέρθηκε το σύνολο αυτό αποτελείται από τα ίδια χαρακτηριστικά παρόλο αυτά το είδος των επιθέσεων καθώς και ο μηχανισμός παραγωγής τους διαφέρει έναντι του σύνολου εκπαίδευσης. Όπως παρατηρούμε στο Πίνακα 7.22 το μοντέλο παρουσιάζει ικανοποιητικές αποδόσεις για τη κλάση των Φυσιολογικών Κινήσεων Δικτύου (F1-Score = 0.8986). Αντιθέτως, τα αποτελέσματα για τη κλάση των επιθέσεων δεν είναι τα αναμενόμενα με το F1-Score να υπολογίζεται ίσο με 0.4578 και την Ανάκληση ίση με

0.3662. Συνολικά η απόδοση του μοντέλου στο σύνολο δεδομένων δεν είναι ικανοποιητική κάτι το οποίο αντικατοπτρίζεται και στις τιμές των ROC-AUC και AP.

	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.8585	0.9427	0.8986	2268624
Attack	0.6106	0.3662	0.4578	556362
Macro Avg.	0.7345	0.6544	0.6782	2824986
Weighted Avg.	0.8096	0.8292	0.8118	2824986
Accuracy			0.8292	2824986
ROC-AUC			0.6901	2824986
Average Precision			0.4607	2824986

Πίνακας 7.22: *Αξιολόγηση (CIC-IDS2017) - LightGBM*

Σημαντικό πρόβλημα παρατηρείται ως προς την αναγνώριση επιθέσεων άρνησης υπηρεσίες τόσο κατανεμημένες όσο και όχι. Αξίζει να σημειωθεί ότι ακόμη και για είδη επιθέσεων για τις οποίες επαρκές πλήθος δεδομένων ήταν διαθέσιμο κατά την εκπαίδευση του μοντέλου αλλά και το ίδιο κατάφερε να επιδείξει αρκετά υποσχόμενα αποτελέσματα στα πλαίσια του CSE-CIC-IDS2018, τώρα φαίνεται να αποτυγχάνει στη σωστή ταξινόμηση τους ενώ για προβληματικά είδη κινήσεων (για παράδειγμα Infiltration) τα αποτελέσματα είναι αντίστοιχης ποιότητας.

Τύπος Κινήσεων Δικτύου	**Εσφαλμένες Ταξινομήσεις (%)**	**Σύνολο**
Φυσιολογική Κίνηση Δικτύου	129922 (5.73)	2268624
DoS attack-Hulk	182139 (79.2)	229965
DDoS	127965 (99.97)	128006
PortScan	11670 (7.35)	158804
DoS attack-GoldenEye	7650 (74.36)	10288
FTP-Brute Force	6568 (82.81)	7931
SSH-Brute Force	5857 (99.36)	5895
DoS attack-SlowHTTPTest	4112 (74.78)	5499
DoS attack-Slowloris	2587 (44.63)	5796
Botnet	1955 (99.95)	1956
Brute Force-Web	1433 (95.09)	1507
Brute Force-XSS	652 (100)	652
Infiltration	33 (94.29)	35
SQL Injection	17 (81)	21
Heartbleed	7 (100)	7

Πίνακας 7.23: *Κατανομή Εσφαλμένων Ταξινομήσεων (CIC-IDS2017)*

Ερευνώντας τους λόγους για τους οποίους το μοντέλο αποτυγχάνει στον εντοπισμό επιθέσεων στο παρόν σύνολο δεδομένων, καταλήγουμε στο συμπέρασμα ότι κινήσεις δικτύου οι οποίες αποτελούν προϊόν διαφορετικών δικτύων και συνεπώς διαφορετικών μηχανισμών παραγωγής αυτών, χαρακτηρίζονται από στοιχεία σημαντικά διαφορετικά μεταξύ τους σε σχέση με άλλα αντίστοιχα σύνολα δεδομένων. Ως αποτέλεσμα αυτού, παρόλο που οι διαθέσιμες παρατηρήσεις ενδέχεται να αποτελούνται από κλάσης επιθέσεων για τις οποίες έχουμε ήδη παρατηρήσει δεδομένα (κατά την εκπαίδευση των μοντέλων), στη πραγματικότητα οι καινο-

ύργιες επιθέσεις είναι πιθανόν να προέρχονται από διαφορετικούς μηχανισμούς παραγωγής γεγονός το οποίο οδηγεί σε σημαντικά διαφορετικές κατανομές και εν τέλη ευθύνεται για την αδυναμία του μοντέλου να αναγνωρίσει τις εκάστοτε επιθέσεις. Για αυτό το λόγο αποφασίστηκε η διερεύνηση της απόδοσης του μοντέλου λαμβάνοντας υπόψη κατά την εκπαίδευση του μοντέλου και τμήματα του νέου συνόλου δεδομένων. Ως σύνολο αξιολόγησης θεωρούμε το σύνολο δεδομένων το οποίο αποτελείται από το συνδυασμό του σύνολο ελέγχου από το CSE-CIC-IDS2018 καθώς και τμήματα του (CIC-IDS2017). Ως βάση χρησιμοποιείται το συνδυαστικό σύνολο δεδομένων το οποίο αποτελείται από το σύνολο ελέγχου από το CSE-CIC-IDS2018 και εξολοκλήρου το σύνολο δεδομένων (CIC-IDS2017). Για το προαναφερθέν σύνολο βάσης παρουσιάζουμε τα αποτελέσματα στο Πίνακα 7.24 με τις τιμές των ROC-AUC και AP να υπολογίζονται ίσες με 0.8165 και 0.71 αντίστοιχα.

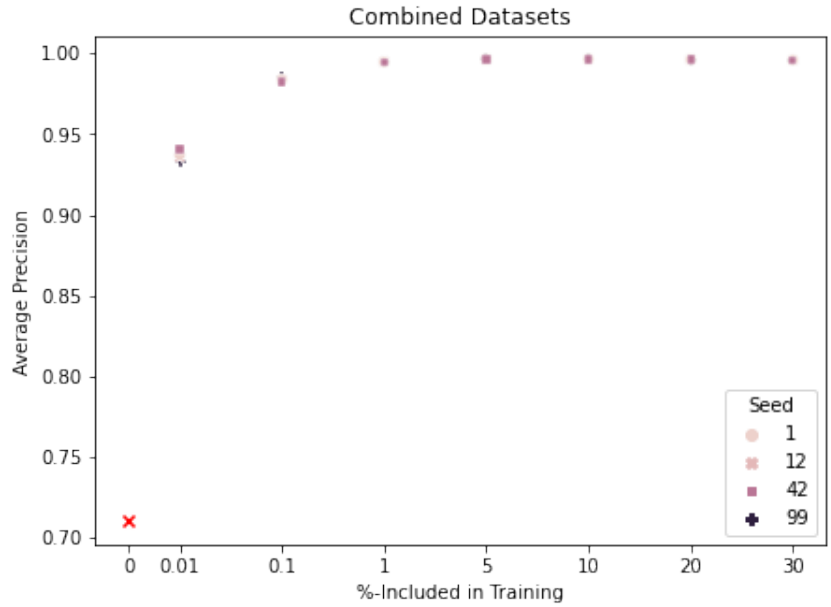
	Precision	**Recall**	**F1-Score**	**Support**
Benign	0.9051	0.9610	0.9322	3607648
Attack	0.7686	0.5624	0.6495	831055
Macro Avg.	0.8368	0.7617	0.7909	4438703
Weighted Avg.	0.8795	0.8864	0.8793	4438703
Accuracy			0.8864	4438703
ROC-AUC			0.8165	4438703
Average Precision			0.71	4438703

Πίνακας 7.24: *Αξιολόγηση Συνδυαστικά Σύνολα Δεδομένων - LightGBM*

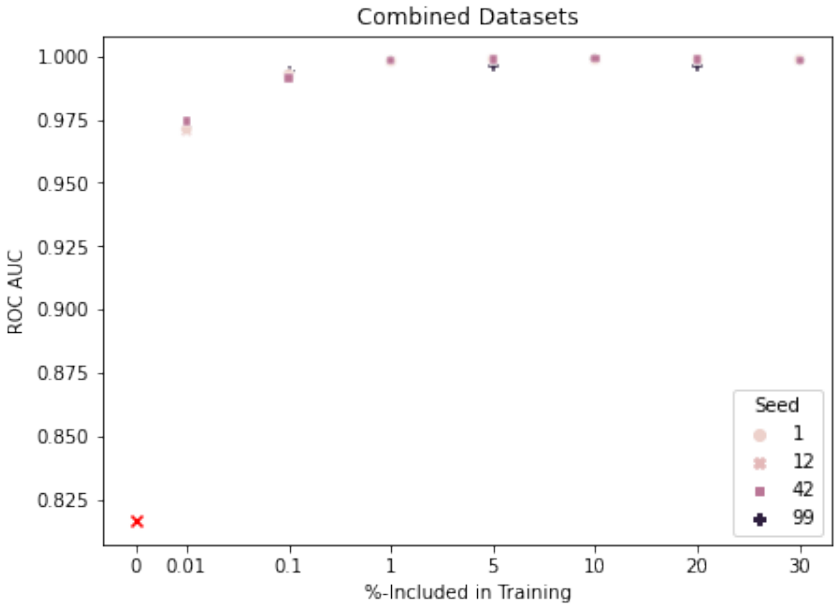
Στη συνέχεια εμπλουτίζουμε το σύνολο εκπαίδευσης με τμήματα από το CIC-IDS2017 και εκπαιδεύουμε την επιλεγέντα αρχιτεκτονική όπως αυτή περιγράφηκε παραπάνω. Όσο αφορά το τμήμα του CIC-IDS2017, το οποίο χρησιμοποιείτε κατά την εκπαίδευση, αυτό καθορίζεται μέσω στρωματοποιημένης δειγματοληψία στα επίπεδα των τύπων κινήσεων δικτύου. Τέλος ερευνούμε τη σχέση ανάμεσα στην επίδραση του μεγέθους του δείγματος από το CIC-IDS2017 και τα τελικά αποτελέσματα. Για αυτό το λόγο καθορίζουμε τα εξής ποσοστά του συνόλου CIC-IDS2017, $0.01\%, 0.1\%, 1\%, 5\%, 10\%, 20\%, 30\%$, με τα οποία εμπλουτίζεται το σύνολο εκπαίδευσης και επαναλαμβάνουμε τα πειράματα για τέσσερα διαφορετικά seed $(1, 12, 42, 99)$ έτσι ώστε να εξασφαλίσουμε την αντικειμενικότητα των αποτελεσμάτων μας.

Στα Σχήματα 7.18 - 7.19 παρουσιάζουμε τη μεταβολή στις μετρικές AP και ROC-AUC αντίστοιχα. Παρατηρούμε ότι ακόμα και μέσω της χρήσης της ελάχιστης ποσότητας του 0.01% (δηλαδή 282 παρατηρήσεις) κατά την εκπαίδευση του μοντέλου η απόδοση αυτού στο συνδυαστικό σύνολο δεδομένων εκτινάσσεστε από 0.71 σε 0.9366 (μέση τιμή ως προς όλα τα seed) για το AP ενώ αντίστοιχα η τιμή της ROC-AUC μετατοπίζεται από 0.8165 σε 0.9719 (μέση τιμή ως προς όλα τα seed). Αυξάνοντάς το μέγεθος του δείγματος που αξιοποιείτε για την εκπαίδευση του μοντέλου οι προαναφερθείσες μετρικές επιτυγχάνουν τιμές της τάξης του 0.99. Συνεπώς καταδεικνύεται με τον πιο εκκωφαντικό τρόπο ότι το μοντέλο είναι σε θέση να χρησιμοποιηθεί με σκοπό την επίλυση προβλημάτων ανίχνευσης εισβολής. Παρόλο αυτά, καθίσταται μέγιστης σημασίας η χρήση, ακόμη και ελάχιστου μέρους, δεδομένων από την εκάστοτε ροή στην οποία θα εφαρμοστεί ή η αναπροσαρμογή του εκπαιδευομένου μοντέλου σε διάφορα χρονικά παράθυρα, έτσι ώστε να είναι σε θέση

να ενσωματώνει καινούργια χρήσιμη πληροφορία σχετικά με τη κατανομή των διαφόρων επιθέσεων.



Σχήμα 7.18: *Μεταβολή AP χρησιμοποιώντας μέρος του CIC-IDS2017 στην Εκπαίδευση*



Σχήμα 7.19: *Μεταβολή ROC AUC χρησιμοποιώντας μέρος του CIC-IDS2017 στην Εκπαίδευση*

Μέρος IV

Επίλογος

Κεφάλαιο 8

Σύνοψη και Προτάσεις

Στα πλαίσια της παρούσας διπλωματικής εργασίας ερευνήθηκαν διάφορες μέθοδοι με σκοπό την ανίχνευση ανωμαλιών. Για να καταστεί το παραπάνω εφικτό χρησιμοποιήθηκαν τόσο κλασσικά μοντέλα μηχανικής μάθησης όσο και τεχνικές βαθιάς μηχανικής μάθησης. Επιπλέον για την εκπαίδευση των εκάστοτε μοντέλων εφαρμόστηκε μια πληθώρα διαφορετικών τεχνικών όπως επιβλεπόμενη μάθηση, ημι-επιβλεπόμενη μάθηση καθώς και υβριδικές τεχνικές. Η παρούσα διπλωματική εργασία δύναται να αξιοποιηθεί ως μια αναλυτική μεθοδολογία μελέτης και αντιμετώπισης προβλημάτων τα οποία σχετίζονται με την ανίχνευση ανωμαλιών σε δεδομένα μορφής πίνακα, χωρίς το τελευταίο να αποτελεί περιοριστικό παράγοντα καθώς όλες οι μεθοδολογίες που παρουσιάστηκαν μπορούν εύκολα να γενικευθούν σε κάθε τύπο δεδομένων. Εξαιτίας της φύσης των προβλημάτων εύρεσης ανωμαλίας η εύρεση ποικιλόμορφων δεδομένων, τα οποία θα μπορούσαν να αξιοποιηθούν με σκοπό τη δημιουργία προβλεπτικών μοντέλων, αποτέλεσε πρόβλημα καθώς τα συγκεκριμένα σύνολα δεδομένων συνήθως είτε δεν είναι διαθέσιμα (ευαισθησία δεδομένων και θέματα ιδιωτικότητας) είτε τα διαθέσιμα σύνολα δεν χαρακτηρίζονται από επαρκή ποιότητα.

Συνοπτικά, η ροή της εργασίας ξεκίνησε με τη μελέτη της βιβλιογραφίας εστιάζοντας σε μεθόδους σχετικές με την δημιουργία μοντέλων μηχανικής μάθησης με σκοπό την ανίχνευση ανωμαλιών. Έχοντας ερευνήσει τη βιβλιογραφία αναζητήθηκαν σύνολα δεδομένων τα οποία να είναι διαθέσιμα δημόσια και να είναι σε θέση να αξιοποιηθούν στα πλαίσια της ανίχνευσης ανωμαλιών. Το τελευταίο μας οδήγησε στη χρήση δεδομένων προερχόμενα από την ερευνητική περιοχή της Ανίχνευσης Εισβολής σε Δεδομένα Δικτύου. Έχοντας στη διάθεση μας τα δεδομένα το πρώτο πρόβλημα το οποίο έπρεπε να επιλυθεί ήταν αυτό της υπολογιστικής διαχείρισης λόγο του μεγάλου όγκου αυτών. Έχοντας προ-επεξεργαστεί τα δεδομένα, ήρθαμε αντιμέτωποι με το ίδιο το πρόβλημα της ανίχνευσης ανωμαλιών καθώς έπρεπε να δημιουργήσουμε μοντέλα ικανά να εντοπίσουν ακραίες τιμές σε ένα σύνολο δεδομένων το οποίο αποτελούνταν κατά κύριο λόγο από φυσιολογικές παρατηρήσεις.

Αρχικά, για την επίλυση του προβλήματος επιλέχθηκε η χρήση μοντέλων μιας κλάσης. Σε αυτό το σημείο επιλέξαμε να συγκρίνουμε την επίδοση ορισμένων απλών μοντέλων (OCSVM, Isolation Forest) έναντι πιο σύνθετων Αυτοκωδικοποιητών (Undercomplete AE, Stacked AE, Sparse AE, Denoising AE). Για τους Αυτοκωδικοποιητές ερευνήθηκε η χρήση μιας πληθώρας αρχιτεκτονικών με τους Αυτοκωδικοποιητές Αφαίρεσης Θορύβου με χρήση Swap Noise να παρουσιάζουν τα καλύτερα αποτελέσματα. Στα πλαίσια της εκπαίδευσης Αυτοκωδικοποιητών μιας κλάσης, παρατηρήθηκε ότι η χρήση ρηχών δικτύων σε συνδυασμό με επιβολή περιορι-

σμών κατά την εκπαίδευση αυτών οδήγησε σε καλύτερες δυνατότητες γενίκευσης έναντι της χρήσης βαθύτερων αρχιτεκτονικών.

Στη συνέχεια, επιλέχθηκε η δημιουργία μοντέλων επιβλεπόμενης μάθησης με σκοπό την ανίχνευση ανωμαλιών, εστιάζοντας σε μοντέλα ενισχυτικής κλίσης (XGBoost, LightGBM, CatBoost). Κάτι το οποίο μας απασχόλησε σε αυτό το σημείο ήταν η μεγάλη ανισορροπία ανάμεσα στις δύο κλάσης ενδιαφέροντος. Για αυτό το λόγο αναζητήθηκαν τεχνικές αντιμετώπισης του προαναφερθέντος προβλήματος. Ως αποτέλεσμα αυτού οδηγηθήκαμε στη χρήση τεχνικών δειγματοληψίας με σκοπό την υποδειγματοληψία της πλειοψηφικής κλάσης αλλά και την επαύξηση της μειοψηφικής κλάσης. Για την επαύξηση των δεδομένων επιλέχθηκε η χρήση της μεθόδου SMOTE καθώς και η χρήση ενός Παραγωγικού Ανταγωνιστικού Δικτύου (CTGAN), το οποίο έχει δημιουργήθει με σκοπό τη παραγωγή συνθετικών δεδομένων μορφής πίνακα. Επιπροσθέτως, ερευνήθηκε η χρήση Cost-Sensitive μάθησης η οποία και οδήγησε και στα μοντέλα με τις καλύτερες επιδόσεις. Συγκεκριμένα καταλήξαμε στο συμπέρασμα ότι η χρήση Cost-Sensitive μάθησης κατάφερε να παράξει τα βέλτιστα αποτελέσματα ενώ αντιθέτως χρήση μεθόδων επαύξησης των δεδομένων αυξάνει την υπολογιστική πολυπλοκότητα χωρίς να βελτιώνει απαραίτητα τα τελικά αποτελέσματα. Επιπλέον, χρησιμοποιήθηκε ένας Αυτοκωδικοποιητής με σκοπό τον εμπλουτισμού του σύνολο εκπαίδευσης με το σφάλμα ανακατασκευής και την ημι-επιβλεπόμενη εκπαίδευση των εκάστοτε μοντέλων, χωρίς όμως να οδηγήσει σε αποτελέσματα ικανά να δικαιολογήσουν τη χρήση της μεθόδου. Τέλος, εφαρμόστηκε Μπεϋζιανή Βελτιστοποίηση με σκοπό το καθορισμό των βέλτιστων υπερπαραμέτρων για τα μοντέλα ενισχυτικής κλίσης. Εκ των μοντέλων επιβλεπόμενης μάθησης τα βέλτιστα αποτελέσματα, στα σύνολα επικύρωσης/ελέγχου, επιτεύχθηκαν μέσω της χρήσης του μοντέλου LightGBM χρησιμοποιώντας Cost Sensitive μάθηση και καθορισμό των υπερπαραμέτρων μέσω Μπεϋζιανής Βελτιστοποίησης. Αν και αναμενόμενο, τα μοντέλα επιβλεπόμενης μάθησης καταφέρνουν να εντοπίσουν τις ανωμαλίες με μεγαλύτερη ακρίβεια έναντι των μοντέλων μιας κλάσης. Παρόλο αυτά, αξίζει να σημειωθεί ότι η επιβλεπόμενη εκπαίδευση ενδεχομένως να μην είναι εφικτή, σε αντίστοιχα προβλήματα ανίχνευσης εισβολής, λόγο έλλειψης δεδομένων. Σε αυτό το πλαίσιο αποδείχθηκε ότι η χρήση Αυτοκωδικοποιήτων εκπαιδευμένων στη κλάση των φυσιολογικών παρατηρήσεων δύναται να οδηγήσει σε αξιοπρεπή αποτελέσματα, ειδικά επιλέγοντας προσεκτικά το εκάστοτε κατώφλι ταξινόμησης, καθώς και να αποτελέσει αφετηρία για δημιουργία μοντέλων βασισμένα στα αποτελέσματα των Αυτοκωδικοποιήτων.

Στο σύνολο ελέγχου το LightGBM κατάφερε να επιτύχει ελάχιστο αριθμό εσφαλμένων ταξινομήσεων για τη κλάση των επιθέσεων, με τη πλειοψηφία αυτών να αποτελούνται από επιθέσεις τύπου Infiltration. Παρόλο αυτά ένα από τα σημαντικότερα προβλήματα το οποίο κληθήκαμε να αντιμετωπίσουμε στα πλαίσια της παρούσας εργασίας ήταν αυτό του concept drift. Πιο συγκεκριμένα, αναφερόμαστε στην αλλαγή των κατανομών των εκάστοτε χαρακτηριστικών στο βάθος του χρόνου. Το παρών έγινε εμφανές όταν επιλέχθηκε η χρήση ενός δεύτερου (προγενέστερου) συνόλου δεδομένων με σκοπό την αξιολόγηση του επιλεγέντος μοντέλου. Σε αυτό το σημείο παρατηρήθηκε ότι, παρόλο που το σύνολο δεδομένων αποτελούνταν από τα ίδια χαρακτηριστικά και παρόμοια είδη επιθέσεων, το μοντέλο δεν κατάφερε να ταυτοποίηση επαρκώς τις ανωμαλίες. Αντιθέτως, παρατηρήσαμε ότι χρησιμοποιώντας ένα ελάχιστο τμήμα από το δεύτερο σύνολο δεδομένων κατά την εκπαίδευση του μοντέλου τότε

το μοντέλο επιδείξει αντίστοιχη συμπεριφορά όπως αυτή παρατηρήθηκε στο αρχικό σύνολο ελέγχου. Το τελευταίο επιδεικνύει με τον πιο εκκωφαντικό τρόπο την ανάγκη για ενημέρωση των εκπαιδευμένων μοντέλων ανά τακτά χρονικά διαστήματα καθώς τα μοτίβα των κινήσεων δικτύου είναι στη πράξη εύκολα μεταβαλλόμενα κάτι το οποίο βοηθάει μελλοντικούς επιτιθέμενους στο να ξεγελάσουν ένα σύστημα μηχανικής μάθησης.

Καθώς αυξάνεται ο όγκος των διαθέσιμων δεδομένων, σε συνδυασμό με τη βελτίωση των διαθέσιμων υπολογιστικών πόρων, τόσο μεγαλύτερη είναι η δυνατότητα μας για αποδοτικότερη εκπαίδευση των εκάστοτε μοντέλων καθώς και χρήση εναλλακτικών και πιο σύνθετων μεθόδων. Ορισμένες από αυτές θα μπορούσαν να είναι η χρήση της πιθανότητας ανακατασκευής όπως αυτή ορίζεται μέσω της χρήσης Μεταβλητών Αυτοκωδικοποιητών μιας κλάσης καθώς και η εκπαίδευση νευρωνικών δικτύων μιας κλάσης με σκοπό την περαιτέρω αξιοποίηση της πλειοψηφικής κλάσης. Επιπλέον, η χρήση Παραγωγικών Ανταγωνιστικών Δικτύων με σκοπό την εύρεση ανωμαλιών GAN-AD είναι πιθανόν να οδηγήσει στην επιλύσει προβλημάτων σχετιζόμενα με το concept drift. Τέλος οφείλουμε να αναφέρουμε την ανάγκη για ερευνά ως προς τη χρήση τεχνικών μεταφοράς γνώσης στα πλαίσια τόσο της εύρεσης ανωμαλιών αλλά και ειδικότερα στα πλαίσια της ανίχνευσης εισβολής. Λόγω της έλλειψης δεδομένων αλλά και της ιδιαιτερότητας των εκάστοτε τομέων εφαρμογής, η δημιουργία και χρήση μοντέλων με σκοπό τη μεταφορά γνώσης αποτελεί ένα αρκετά υποσχόμενο πεδίο έρευνας. Το παρών σύνολο δεδομένων, όπως αυτό χρησιμοποιήθηκε στα πλαίσια της εκπαίδευσης των μοντέλων, δύναται να χρησιμοποιηθεί με σκοπό την εκπαίδευση μοντέλων τα οποία στη συνέχεια θα μπορούσαν να χρησιμοποιηθούν με σκοπό τη μεταφορά γνώσης σε αντίστοιχα προβλήματα.

Το σύνολο του κώδικα, ο οποίος χρησιμοποιήθηκε για την υλοποίηση της παρούσας διπλωματικής εργασίας, μπορεί να βρεθεί εδώ [121].

Βιβλιογραφία

[1] Yanpei Hua. *An Efficient Traffic Classification Scheme Using Embedded Feature Selection and LightGBM*. *2020 Information Communication Technologies Conference (ICTC)*, σελίδες 125–130, 2020.

[2] Qusyairi Ridho Saeful Fitni και Kalamullah Ramli. *Implementation of Ensemble Learning and Feature Selection for Performance Improvements in Anomaly-Based Intrusion Detection Systems*. *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, σελίδες 118–124, 2020.

[3] Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis και Helge Janicke. *Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study*. *Journal of Information Security and Applications*, 50:102419, 2020.

[4] Ram B. Basnet, Riad Shash, Clayton Johnson, Lucas Walgren και Tenzin Doleck. *Towards Detecting and Classifying Network Intrusion Traffic Using Deep Learning Frameworks*. *J. Internet Serv. Inf. Secur.*, 9:1–17, 2019.

[5] V. Kanimozhi και T. Prem Jacob. *Artificial Intelligence based Network Intrusion Detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing*. *ICT Express*, 5(3):211–214, 2019.

[6] Gozde Karatas, Onder Demir και Ozgur Koray Sahingoz. *Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset*. *IEEE Access*, 8:32150–32162, 2020.

[7] Xianwei Gao, Chun Shan, Changzhen Hu, Zequn Niu και Zhen Liu. *An Adaptive Ensemble Machine Learning Model for Intrusion Detection*. *IEEE Access*, 7:82512–82521, 2019.

[8] Aklil Zenebe Kiflay, Athanasios Tsokanos και Raimund Kirner. *A Network Intrusion Detection System Using Ensemble Machine Learning*. *2021 International Carnahan Conference on Security Technology (ICCST)*, σελίδες 1–6, 2021.

[9] Bayu Adhi Tama, Marco Comuzzi και Kyung Hyune Rhee. *TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System*. *IEEE Access*, 7:94497–94507, 2019.

[10] J. Ren, X. Liu, Q. Wang, H. He και X. Zhao. *An Multi-Level Intrusion Detection Method Based on KNN Outlier Detection and Random Forests*. *Jisuanji Yanjiu yu Fazhan/Computer Research and Development*, 56:566–575, 2019.

[11] Hossein Shapoorifard και Pirooz Shamsinjead Babaki. *Intrusion Detection using a Novel Hybrid Method Incorporating an Improved KNN*. *International Journal of Computer Applications*, 173:5–9, 2017.

[12] Gisung Kim, Seungmin Lee και Sehun Kim. *A novel hybrid intrusion detection method integrating anomaly detection with misuse detection*. *Expert Systems with Applications*, 41(4, Παρτ 2):1690–1700, 2014.

[13] Yanfang Fu, Yishuai Du, Zijian Cao, Qiang Li και Wei Xiang. *A Deep Learning Model for Network Intrusion Detection with Imbalanced Data*. *Electronics*, 11(6), 2022.

[14] Cosimo Ieracitano, Ahsan Adeel, Francesco Carlo Morabito και Amir Hussain. *A novel statistical analysis and autoencoder driven intelligent intrusion detection approach*. *Neurocomputing*, 387:51–62, 2020.

[15] Hongchao Song, Zhuqing Jiang, Aidong Men και Bo Yang. *A Hybrid Semi-Supervised Anomaly Detection Model for High-Dimensional Data*. *Computational Intelligence and Neuroscience*, 2017, 2017.

[16] Nathan Shone, Tran Nguyen Ngoc, Vu Dinh Phai και Qi Shi. *A Deep Learning Approach to Network Intrusion Detection*. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1):41–50, 2018.

[17] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus Robert Müller και Marius Kloft. *Deep Semi-Supervised Anomaly Detection*, 2019.

[18] Tongtong Su, Huazhi Sun, Jinqi Zhu, Sheng Wang και Yabo Li. *BAT: Deep Learning Methods on Network Intrusion Detection Using NSL-KDD Dataset*. *IEEE Access*, 8:29575–29585, 2020.

[19] Peng Lin, Kejiang Ye και Cheng Zhong Xu. *Dynamic Network Anomaly Detection System by Using Deep Learning Techniques*. *CLOUD*, σελίδα 161–176, Berlin, Heidelberg, 2019. Springer-Verlag.

[20] Vinayakumar Ravi, Soman Kp και Prabaharan Poornachandran. *A Comparative Analysis of Deep Learning Approaches for Network Intrusion Detection Systems (N-IDSs): Deep Learning for N-IDSs*. *International Journal of Digital Crime and Forensics*, 11:65–89, 2019.

[21] Pablo Torres, Carlos Catania, Sebastian Garcia και Carlos Garcia Garino. *An analysis of Recurrent Neural Networks for Botnet detection behavior*. *2016 IEEE Biennial Congress of Argentina (ARGENCON)*, σελίδες 1–6, 2016.

[22] Wei Wang, Ming Zhu, Xuewen Zeng, Xiaozhou Ye και Yiqiang Sheng. *Malware traffic classification using convolutional neural network for representation learning*. *2017 International Conference on Information Networking (ICOIN)*, σελίδες 712–717, 2017.

[23] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias και Georgios Tagaris. *Deep neural architectures for prediction in healthcare*. *Complex & Intelligent Systems*, 4(2):119–131, 2018.

[24] Athanasios Tagaris, Dimitrios Kollias και Andreas Stafylopatis. *Assessment of Parkinson's disease based on deep neural networks*. *International Conference on Engineering Applications of Neural Networks*, σελίδες 391–403. Springer, 2017.

[25] Athanasios Tagaris, Dimitrios Kollias, Andreas Stafylopatis, Georgios Tagaris και Stefanos Kollias. *Machine learning for neurodegenerative disorder diagnosis—survey of practices and launch of benchmark dataset*. *International Journal on Artificial Intelligence Tools*, 27(03):1850011, 2018.

[26] Ilianna Kollia, Andreas Georgios Stafylopatis και Stefanos Kollias. *Predicting Parkinson's disease using latent information extracted from deep neural networks*. *2019 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8. IEEE, 2019.

[27] James Wingate, Ilianna Kollia, Luc Bidaut και Stefanos Kollias. *Unified deep learning approach for prediction of Parkinson's disease*. *IET Image Processing*, 14(10):1980–1989, 2020.

[28] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian και Stefanos Kollias. *Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 537–544, 2021.

[29] Dimitrios Kollias, Anastasios Arsenos και Stefanos Kollias. *Ai-mia: Covid-19 detection & severity analysis through medical imaging*. *arXiv preprint arXiv:2206.04732*, 2022.

[30] Anastasios Arsenos, Dimitrios Kollias και Stefanos Kollias. *A Large Imaging Database and Novel Deep Neural Architecture for Covid-19 Diagnosis*. *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, σελίδες 1–5. IEEE, 2022.

[31] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis και Stefanos Kollias. *Adaptation and contextualization of deep neural network models*. *2017 IEEE symposium series on computational intelligence (SSCI)*, σελίδες 1–8. IEEE, 2017.

[32] D Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, I Kollia, L Sukissian, J Wingate και S Kollias. *Deep Transparent Prediction through Latent Representation Analysis*. *arXiv preprint arXiv:2009.07044*, 2020.

[33] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate και S Kollias. *Transparent adaptation in deep medical image diagnosis*. *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, σελίδες 251–267. Springer, 2020.

[34] Fabio De Sousa Ribeiro, Francesco Calivá, Mark Swainson, Kjartan Gudmundsson, Georgios Leontidis και Stefanos Kollias. *Deep bayesian self-training*. *Neural Computing and Applications*, 32(9):4275–4291, 2020.

[35] Fabio De Sousa Ribeiro, Georgios Leontidis και Stefanos D Kollias. *Capsule Routing via Variational Bayes*. *AAAI*, σελίδες 3749–3756, 2020.

[36] Fabio De Sousa Ribeiro, Georgios Leontidis και Stefanos Kollias. *Introducing routing uncertainty in capsule networks*. *Advances in Neural Information Processing Systems*, 33:6490–6502, 2020.

[37] Nikolaos Simou και Stefanos Kollias. *Fire: A fuzzy reasoning engine for imprecise knowledge*. Citeseer.

[38] Francesco Caliva, Fabio Sousa De Ribeiro, Antonios Mylonakis, Christophe Demazi'ere, Paolo Vinai, Georgios Leontidis και Stefanos Kollias. *A deep learning approach to anomaly detection in nuclear reactors*. *2018 International joint conference on neural networks (IJCNN)*, σελίδες 1–8. IEEE, 2018.

[39] Bashar Alhnaity, Stefanos Kollias, Georgios Leontidis, Shouyong Jiang, Bert Schamp και Simon Pearson. *An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth*. *Information Sciences*, 560:35–50, 2021.

[40] Bashar Alhnaity, Simon Pearson, Georgios Leontidis και Stefanos Kollias. *Using deep learning to predict plant growth and yield in greenhouse environments*. *International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019 1296*, σελίδες 425–432, 2019.

[41] Stefanos Kollias, Miao Yu, James Wingate, Aiden Durrant, Georgios Leontidis, Georgios Alexandridis, Andreas Stafylopatis, Antonios Mylonakis, Paolo Vinai και Christophe Demaziere. *Machine learning for analysis of real nuclear plant data in the frequency domain*. *Annals of Nuclear Energy*, 177:109293, 2022.

[42] Andreas Psaroudakis και Dimitrios Kollias. *MixAugment & Mixup: Augmentation Methods for Facial Expression Recognition*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 2367–2375, 2022.

[43] Dimitrios Kollias και Stefanos Zafeiriou. *Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm*. *2018 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8. IEEE, 2018.

[44] Dimitrios Kollias και Stefanos Zafeiriou. *Va-stargan: Continuous affect generation*. *International Conference on Advanced Concepts for Intelligent Vision Systems*, σελίδες 227–238. Springer, 2020.

[45] G Caridakis, A Raouzaiou, K Karpouzis και S Kollias. *Synthesizing Gesture Expressivity Based on Real Sequences*. *Workshop Programme*, τόμος 10, σελίδα 19, 2006.

[46] Phivos Mylonas, Evaggelos Spyrou, Yannis Avrithis και Stefanos Kollias. *Using visual context and region semantics for high-level concept detection*. *IEEE Transactions on Multimedia*, 11(2):229–243, 2009.

[47] Stefanos Kollias και Dimitris Anastassiou. *A unified neutral network approach to digital image halftoning*. *IEEE Transactions on signal processing*, 39(4):980–984, 1991.

[48] Paraskevi Tzouveli, Andreas Schmidt, Michael Schneider, Antonis Symvonis και Stefanos Kollias. *Adaptive reading assistance for the inclusion of students with dyslexia: The AGENT-DYSL approach*. *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, σελίδες 167–171. IEEE, 2008.

[49] Varun Chandola, Arindam Banerjee και Vipin Kumar. *Anomaly Detection: A Survey*. *ACM Comput. Surv.*, 41(3), 2009.

[50] Jiawei Han, Micheline Kamber και Jian Pei. *12 - Outlier Detection*. *Data Mining (Third Edition)*Jiawei Han, Micheline Kamber και Jian Pei, επιμελητές, The Morgan Kaufmann Series in Data Management Systems, σελίδες 543–584. Morgan Kaufmann, Boston, third editionη έκδοση, 2012.

[51] Victoria Hodge και Jim Austin. *A Survey of Outlier Detection Methodologies*. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[52] Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola και Robert Williamson. *Estimating Support of a High-Dimensional Distribution*. *Neural Computation*, 13:1443–1471, 2001.

[53] Fei Tony Liu, Kai Ming Ting και Zhi Hua Zhou. *Isolation forest*. *2008 Eighth IEEE International Conference on Data Mining*, σελίδες 413–422. IEEE, 2008.

[54] Fei Tony Liu, Kai Ming Ting και Zhi Hua Zhou. *Isolation-Based Anomaly Detection*. *ACM Trans. Knowl. Discov. Data*, 6(1), 2012.

[55] David E. Rumelhart, Geoffrey E. Hinton και Ronald J. Williams. *Learning Internal Representations by Error Propagation*. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*David E. Rumelhart και James L. Mcclelland, επιμελητές, σελίδες 318–362. MIT Press, Cambridge, MA, 1986.

[56] Geoffrey E Hinton και Richard Zemel. *Autoencoders, Minimum Description Length and Helmholtz Free Energy*. *Advances in Neural Information Processing Systems*J. Cowan, G. Tesauro και J. Alspector, επιμελητές, τόμος 6. Morgan-Kaufmann, 1993.

[57] Jürgen Schmidhuber. *Deep Learning in Neural Networks: An Overview*. *CoRR*, αβσ/1404.7828, 2014.

[58] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[59] H. Bourlard και Y. Kamp. *Auto-Association by Multilayer Perceptrons and Singular Value Decomposition*. *Biol. Cybern.*, 59(4–5):291–294, 1988.

[60] Andrew Ng και others. *Sparse autoencoder*. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[61] Devansh Arpit, Yingbo Zhou, Hung Ngo και Venu Govindaraju. *Why Regularized Auto-Encoders learn Sparse Representation?*, 2015.

[62] Jonathon Shlens. *Notes on Kullback-Leibler Divergence and Likelihood*. *CoRR*, αβσ/1404.2000, 2014.

[63] Harald Steck. *Autoencoders that don't overfit towards the Identity*. *Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 19598–19608. Curran Associates, Inc., 2020.

[64] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio και Pierre Antoine Manzagol. *Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion*. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010.

[65] Sercan O. Arik και Tomas Pfister. *TabNet: Attentive Interpretable Tabular Learning*, 2019.

[66] Jinsung Yoon, Yao Zhang, James Jordon και Mihaelavan der Schaar. *VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain*. *Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 11033–11043. Curran Associates, Inc., 2020.

[67] Ravid Shwartz-Ziv και Amitai Armon. *Tabular Data: Deep Learning is Not All You Need*, 2021.

[68] Llew Mason, Jonathan Baxter, Peter Bartlett και Marcus Frean. *Boosting Algorithms as Gradient Descent*. *Advances in Neural Information Processing Systems*S. Solla, T. Leen και K. Müller, επιμελητές, τόμος 12. MIT Press, 1999.

[69] Gareth James, Daniela Witten, Trevor Hastie και Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[70] Jerome H. Friedman. *Greedy Function Approximation: A Gradient Boosting Machine*. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[71] Tianqi Chen και Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, σελίδες 785–794, New York, NY, USA, 2016. ACM.

[72] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye και Tie Yan Liu. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. *Advances in Neural Information Processing Systems*I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan και R. Garnett, επιμελητές, τόμος 30. Curran Associates, Inc., 2017.

[73] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush και Andrey Gulin. *CatBoost: unbiased boosting with categorical features*. *Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[74] K. Pearson. *On Lines and Planes of Closest Fit to Systems of Points in Space*. *Philosophical Magazine*, 2:559–572, 1901.

[75] Harold Hotelling. *Analysis of a complex of statistical variables into principal components*. *Journal of Educational Psychology*, 24:498–520, 1933.

[76] Ian T. Jolliffe και Jorge Cadima. *Principal component analysis: a review and recent developments*. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[77] Bradley Efron και Robert J. Tibshirani. *An Introduction to the Bootstrap*. Αριθμός 57 στο Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[78] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo Prati, Bartosz Krawczyk και Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer Cham, 2018.

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

[80] Rok Blagus και Lara Lusa. *SMOTE for High-Dimensional Class-Imbalanced Data*. *BMC bioinformatics*, 14:106, 2013.

[81] Yotam Elor και Hadar Averbuch-Elor. *To SMOTE, or not to SMOTE?* *CoRR*, αβσ/2201.08528, 2022.

[82] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville και Yoshua Bengio. *Generative Adversarial Networks*, 2014.

[83] Martin Arjovsky, Soumith Chintala και Léon Bottou. *Wasserstein GAN*, 2017.

[84] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin και Aaron C. Courville. *Improved Training of Wasserstein GANs*. *CoRR*, αβσ/1704.00028, 2017.

[85] Mehdi Mirza και Simon Osindero. *Conditional Generative Adversarial Nets*. *CoRR*, αβσ/1411.1784, 2014.

[86] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante και Kalyan Veeramachaneni. *Modeling Tabular data using Conditional GAN*. *CoRR*, αβσ/1907.00503, 2019.

[87] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[88] Zinan Lin, Ashish Khetan, Giulia Fanti και Sewoong Oh. *PacGAN: The power of two samples in generative adversarial networks*. *Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[89] Charles Elkan. *The Foundations of Cost-Sensitive Learning*. *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle*, 1, 2001.

[90] Nguyen Thai-Nghe, Zeno Gantner και Lars Schmidt-Thieme. *Cost-sensitive learning methods for imbalanced data*. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8, 2010.

[91] James Bergstra, Rémi Bardenet, Yoshua Bengio και Balázs Kégl. *Algorithms for Hyper-Parameter Optimization*. *Advances in Neural Information Processing Systems*J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira και K.Q. Weinberger, επιμελητές, τόμος 24. Curran Associates, Inc., 2011.

[92] Jasper Snoek, Hugo Larochelle και Ryan P. Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*, 2012.

[93] Vu Nguyen. *Bayesian Optimization for Accelerating Hyper-Parameter Tuning*. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, σελίδες 302–305, 2019.

[94] Donald Jones. *A Taxonomy of Global Optimization Methods Based on Response Surfaces*. *J. of Global Optimization*, 21:345–383, 2001.

[95] Iman Sharafaldin, Arash Habibi Lashkari και Ali Ghorbani. *Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization*. *ICISSP*, σελίδες 108–116, 2018.

[96] Arash Habibi Lashkari, Amy Seo, Gerard Drapper Gil και Ali Ghorbani. *CIC-AB: Online ad blocker for browsers*. *2017 International Carnahan Conference on Security Technology (ICCST)*, σελίδες 1–7, 2017.

[97] Sture Holm. *A Simple Sequentially Rejective Multiple Test Procedure*. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[98] Alaa Tharwat. *Classification assessment methods*. *Applied Computing and Informatics*, 2020.

[99] Tom Fawcett. *An introduction to ROC analysis*. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[100] Jorge M. Lobo, Alberto Jiménez-Valverde και Raimundo Real. *AUC: a misleading measure of the performance of predictive distribution models*. *Global Ecology and Biogeography*, 17(2):145–151, 2008.

[101] John Muschelli. *ROC and AUC with a Binary Predictor: A Potentially Misleading Metric*. *J. Classif.*, 37(3):696–708, 2020.

[102] Jesse Davis και Mark Goadrich. *The Relationship Between Precision-Recall and ROC Curves*. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, τόμος 06, 2006.

[103] Takaya Saito και Marc Rehmsmeier. *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. *PLOS ONE*, 10(3):1–21, 2015.

[104] Kendrick Boyd, Kevin Eng και C. Page. *Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals*. *Machine Learning and Knowledge Discovery in Databases*, τόμος 8190, σελίδες 451–466, 2013.

[105] Guangzhe Fan και Mu Zhu. *Detection of rare items with TARGET*. *Statistics and Its Interface Volume*, 4:11–17, 2011.

[106] *Coogle. Colab*. https://colab.research.google.com.

[107] Wes McKinney. *Data Structures for Statistical Computing in Python*. *Proceedings of the 9th Python in Science Conference*Stéfan van der Walt και Jarrod Millman, επιμελητές, σελίδες 56 – 61, 2010.

[108] Charles R. Harris, K. Jarrod Millman, Stéfan J.van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H.van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernándezdel Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke και Travis E. Oliphant. *Array programming with NumPy*. *Nature*, 585(7825):357–362, 2020.

[109] J. D. Hunter. *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[110] Michael L. Waskom. *seaborn: statistical data visualization*. *Journal of Open Source Software*, 6(60):3021, 2021.

[111] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot και E. Duchesnay. *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[112] Guillaume Lemaître, Fernando Nogueira και Christos K. Aridas. *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[113] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu και Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015. Software available from tensorflow.org.

[114] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins και David D Cox. *Hyperopt: a python library for model selection and hyperparameter optimization*. *Computational Science & Discovery*, 8(1):014008, 2015.

[115] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta και Masanori Koyama. *Optuna: A Next-generation Hyperparameter Optimization Framework*. *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[116] Djork Arné Clevert, Thomas Unterthiner και Sepp Hochreiter. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*, 2015.

[117] Pierre Baldi και Peter J Sadowski. *Understanding Dropout*. *Advances in Neural Information Processing Systems*C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani και K.Q. Weinberger, επιμελητές, τόμος 26. Curran Associates, Inc., 2013.

[118] Stefan Wager, Sida Wang και Percy Liang. *Dropout Training as Adaptive Regularization*, 2013.

[119] *Porto Seguro's Safe Driver Prediction*. https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/discussion/44629. Ημερομηνία πρόσβασης: 03-02-2022 [Online].

[120] Scott Lundberg και Su In Lee. *A Unified Approach to Interpreting Model Predictions*, 2017.

[121] https://github.com/dpapamavros/AnomalyDetection. [Online].