ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Εκπαίδευση Ημιεπιβλεπόμενων Μοντέλων Βαθιάς Μάθησης με Χρήση Αντιπαραδειγμάτων για την Ανίχνευση Εισβολών Δικτύου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μάρκος Π. Δεληγιάννης

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Εκπαίδευση Ημιεπιβλεπόμενων Μοντέλων Βαθιάς Μάθησης με Χρήση Αντιπαραδειγμάτων για την Ανίχνευση Εισβολών Δικτύου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μάρκος Π. Δεληγιάννης

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Οκτωβρίου 2024.

Συμεών Παπαβασιλείου Καθηγητής Ε.Μ.Π	Ελένη Στάη Επ. Καθηγήτρια Ε.Μ.Π	Βασίλειος Καρυώτης Αναπ. Καθηγητής Ιόνιου Πανεπιστημίου

Αθήνα, Οκτώβριος 2024

Μάρκος Π. Δεληγιάννης

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

Περίληψη

Το φαινόμενο των κυβερνοεπιθέσεων αποτελεί ένα φλέγον ζήτημα που πλήττει τις σύγχρονες δικτυακές εγκασταστάσεις. Ένας από τους επιστημονικούς κλάδους που έχουν γνωρίσει έντονη ανάπτυξη ως απόρροια αυτού είναι αυτός της Ανίχνευσης Εισβολών σε Δίκτυα βασισμένος σε τεχνικές εντοπισμού ανωμαλιών. Ο τομέας αυτός εστιάζει στην ανάπτυξη ισχυρών μεθόδων που μπορούν να αποφανθούν για τον χαρακτήρα δικτυακής κίνησης με τρόπο γενικό, ανιχνεύοντας περιπτώσεις κακόβουλης κίνησης με την οποία δεν έχουν έρθει προηγουμένως σε επαφή. Στόχος αυτής της διπλωματικής εργασίας είναι η προσαρμογή και εξερεύνηση μίας υποχρησιμοποιούμενης μεθόδου για την ενίσχυση σύγχρονων αρχιτεκτονικών βαθιάς μάθησης, με εφαρμογή στο πρόβλημα της Ανίχνευσης Εισβολών σε Δίκτυα. Ο πυρήνας της μεθόδου αυτής είναι η προσαρμογή της εκπαιδευτικής διαδικασίας χωρίς αρχιτεκτονική αλλαγή σε ημιεπιβλεπόμενα μοντέλα. Τα μοντέλα αυτά, τα οποία βασίζονται στη μάθηση μέσω ανακατασκευής των δειγμάτων φυσιολογικής δικτυακής κίνησης, τροποποιούνται ώστε να συμπεριλαμβάνονται με οργανικό τρόπο και δείγματα ανώμαλης κίνησης ως αντιπαραδείγματα. Αυτό υλοποιείται ενθαρρύνοντας την κακή ανακατασκευή των ανώμαλων δειγμάτων, εισάγοντας στον στόχο ελαχιστοποίησης έναν όρο μέσου αντίστροφου σφάλματος ανακατασκευής. Έτσι, το μοντέλο μπορεί να ενσωματώσει γνώση για τη μορφή της κακόβουλης κίνησης, χωρίς όμως να θυσιάσει την ευελιξία και ικανότητα γενίκευσης των ημιεπιβλεπόμενων αρχιτεκτονικών. Για τον έλεγχο της επίδρασης που έχει η μέθοδος αυτή υλοποιούμε, τροποποιούμε κατάλληλα και παρουσιάζουμε τέσσερις αρχιτεκτονικές βαθιάς μάθησης, έναν απλό αυτοκωδικοποιητή, έναν παραλλακτικό αυτοκωδικοποιητή και δύο παραγωγικά αντιπαραθετικά δίκτυα της βιβλιογραφίας, ένα το οποίο έχει σχεδιαστεί για εντοπισμό ανωμαλιών σε εικόνες και ένα που έχει προταθεί για την ανίχνευση εισβολών στα δίκτυα. Αυτές οι αρχιτεκτονικές εκπαιδεύονται καταρχάς με χρήση μόνο δειγμάτων ομαλής κίνησης και στη συνέχεια με την προσθήκη αντιπαραδειγμάτων. Η εκπαίδευση και ο έλεγχος γίνονται στα σύνολα δεδομένων CIC-IDS-2018 και UNSW-NB15. Από την ανάλυσή μας προκύπτει ότι η προσθήκη αντιπαραδειγμάτων ανώμαλης κίνησης οδηγεί στην δραματική αύξηση της απόδοσης όλων των αρχιτεκτονικών και στα δύο σύνολα δεδομένων. Επιπλέον, εξετάζουμε τη μέθοδο των αντιπαραδειγμάτων σε μία νέα αρχιτεκτονική, η οποία βασίζεται στον υπολογισμό της συσχέτισης σε ένα παράθυρο δειγμάτων και τον εντοπισμό των δικτυακών επιθέσεων μέσω ενός δισδιάστατου συνελικτικού αυτοκωδικοποιητή. Παρόλα αυτά, ο πυρήνας του μοντέλου αυτού αποδεικνύεται ανεπαρκής, και έτσι αυτό αναφέρεται μόνο για λόγους πληρότητας.

Λέξεις-κλειδιά: Ανίχνευση Εισβολών, Δίκτυα, Εντοπισμός Ανωμαλιών, Μηχανική Μάθηση, Βαθιά Μάθηση, GAN, Autoencoder, AE, VAE, CNN, ConvAE

Summary

The phenomenon of cyberattacks constitutes a pressing issue that affects modern network infrastructures. One of the scientific fields that has experienced significant growth because of this is Network Intrusion Detection based on anomaly detection techniques. This field focuses on the development of robust methods capable of determining the nature of network traffic in a general manner, detecting instances of malicious traffic that have not been previously encountered. The aim of this thesis is to adapt and explore an underutilized method to enhance modern deep learning architectures in the field of Network Intrusion Detection. The core of this method lies in adapting the training process, without altering the architecture, for semi-supervised models. These models, which are based on learning by reconstructing samples of normal network traffic, are modified to organically incorporate samples of anomalous traffic as counterexamples. This is achieved by encouraging the erroneous reconstruction of anomalous samples by introducing a term for the average inverse reconstruction error into the minimization objective. In this way, the model can incorporate knowledge of the nature of malicious traffic without sacrificing the flexibility and generalization capabilities of semi-supervised architectures. To assess the impact of this method, we implement, appropriately modify, and present four deep learning architectures: a simple autoencoder, a variational autoencoder, and two generative adversarial networks from the literature—one designed for anomaly detection in images and another proposed for network intrusion detection. These architectures are initially trained using only normal traffic samples, followed by the introduction of counterexamples. Training and evaluation are conducted on the CIC-IDS-2018 and UNSW-NB15 datasets. Our analysis indicates that the inclusion of anomalous traffic counterexamples leads to a dramatic performance improvement across all architectures on both datasets. Additionally, we examine the counterexample method in a novel architecture, which is based on calculating correlations within a sliding window of samples and detecting network attacks using a two-dimensional convolutional autoencoder. The underlying model, however, proves inadequate and thus is included only for the sake of completeness.

Keywords: Intrusion Detection, Network, Anomaly detection, Machine Learning, Deep Learning, GAN, Autoencoder, AE, VAE, CNN, ConvAE

Ευχαριστίες

Η ολοκλήρωση αυτής της διπλωματικής εργασίας δεν θα ήταν εφικτή χωρίς την υποστήριξη και την καθοδήγηση κάποιων ξεχωριστών ανθρώπων, στους οποίους οφείλω βαθιές ευχαριστίες.

Πρώτα απ' όλα, θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη στον επιβλέποντα καθηγητή μου, κ. Βασίλειο Καρυώτη, για την πολύτιμη καθοδήγηση, την εμπιστοσύνη που μου έδειξε και τις ουσιαστικές συμβουλές του καθ' όλη τη διάρκεια της έρευνάς μου. Η υποστήριξή του υπήρξε καθοριστική για την ολοκλήρωση αυτής της εργασίας.

Θα ήθελα επίσης να ευχαριστήσω θερμά τους φίλους μου, οι οποίοι στάθηκαν δίπλα μου σε κάθε βήμα αυτής της διαδρομής. Ιδιαίτερη αναφορά αξίζει στον παιδικό μου φίλο, Τζόελ Γιάννη, που ήταν πάντα εκεί για να με στηρίζει και να μου δίνει δύναμη. Επίσης, ευχαριστώ τους συμφοιτητές και πολύ καλούς μου φίλους Πέτρο Μαράτο, Γεράσιμο Μουντάκη και Διονύση Κεφαλληνό για τη συντροφικότητα και τη βοήθεια τους, καθώς και για τις εποικοδομητικές συζητήσεις που είχαμε.

Τέλος, εκφράζω την απεριόριστη ευγνωμοσύνη μου στους γονείς μου, Αθανασία Τσιούγκου και Παναγιώτη Δεληγιάννη, που με στήριξαν αδιάκοπα όλα αυτά τα χρόνια, προσφέροντάς μου αγάπη, κατανόηση και ενθάρρυνση σε κάθε μου βήμα.

Η συμβολή όλων σας ήταν ανεκτίμητη και σας ευχαριστώ από καρδιάς.

Πίνακας περιεχομένων

Πίνακας σχημάτων

Πίνακας πινάκων

1. Εισαγωγή

Ο όρος *εντοπισμός ανωμαλιών* (anomaly detection) αποτελεί έναν γενικό όρο που συμπεριλαμβάνει μία πληθώρα διαφορετικών τομέων με κοινό παρονομαστή την εύρεση αποκλινόντων δειγμάτων σε ένα σύνολο δεδομένων. Τα αποκλίνοντα αυτά δεδομένα, ή αλλιώς *ανωμαλίες* είναι δεδομένα που διαφοροποιούνται από τα υπόλοιπα με κάποιον ουσιώδη τρόπο και αντιστοιχούν σε ανεπιθύμητες καταστάσεις, των οποίων ο εντοπισμός και αντιμετώπιση είναι επιθυμητή. Στη βιβλιογραφία υπάρχει πληθώρα εργασιών σε τομείς όπως η ιατρική για τον εντοπισμό ασθενειών [1], τα οικονομικά για τον εντοπισμό ξεπλύματος χρήματος και άλλων ειδών χρηματοοικονομικής απάτης [2], και η επιστήμη υλικών για τον εντοπισμό βλάβης σε δομικά υλικά [3].

Αυτή η διπλωματική εργασία επικεντρώνεται στον τομέα της *Ανίχνευσης Εισβολών* (Intrusion Detection). Πρόκειται για μία περιοχή της επιστήμης υπολογιστών που ασχολείται με τον εντοπισμό κακόβουλων ενεργειών σε μία υπολογιστική υποδομή. Με την αδιάκοπη εξέλιξη των υπολογιστικών τεχνολογιών, η πολυπλοκότητα των υποδομών που χρησιμοποιούνται από εταιρίες, ιδρύματα και οργανισμούς αυξάνεται ραγδαία. Η μεγαλύτερη πολυπλοκότητα των υποδομών και έντονη εξάρτηση από αυτές οδηγεί αναπόφευκτα στην έξαρση των κυβερνοεπιθέσεων, οι οποίες μάλιστα γίνονται ολοένα και πιο αδιόρατες. Ο εντοπισμός τους, λοιπόν, καθίσταται ένα μεγάλης κρισιμότητας πρόβλημα. Ο γενικός τομέας της ανίχνευσης εισβολών χωρίζεται σε δύο τομείς, τον εντοπισμό ανωμαλιών σε επίπεδο υπολογιστή (Host-based Intrusion Detection Systems / HIDS) και τον εντοπισμό ανωμαλιών σε επίπεδο δικτύου (Network-based Intrusion Detection Systems / NIDS). Στον πρώτο τομέα, το επίκεντρο της προσοχής είναι ο υπολογιστής. Ένα σύστημα HIDS συχνά λειτουργεί λαμβάνοντας στην είσοδό του αρχεία καταγραφής [4], ανιχνεύοντας αποκλίσεις στις εγγραφές τους που αποτελούν σημάδια κακόβουλης δραστηριότητας. Στον δεύτερο τομέα, το αντικείμενο της προσοχής είναι η δικτυακή δραστηριότητα. Πιο συγκεκριμένα, τα συστήματα NIDS δέχονται ως είσοδο δείγματα δικτυακής κίνησης και προσπαθούν να ανιχνεύσουν εάν αυτά αντιστοιχούν σε κανονική ή κακόβουλη κίνηση [5].

Σε αυτήν τη μελέτη θα μας απασχολήσει ο τομέας του εντοπισμού ανωμαλιών σε δίκτυα. Αξίζει να σημειωθεί ότι ο τομέας αυτός δεν εμπίπτει στο σύνολό του στον κλάδο του εντοπισμού ανωμαλιών. Τα συστήματα NIDS χωρίζονται σε δύο μεγάλες οικογένειες, τα NIDS βασισμένα σε υπογραφές (Signature based NIDS), και τα Anomaly Detection NIDS (εντοπισμού ανωμαλιών). Τα πρώτα ανιχνεύουν την κακόβουλη κίνηση συγκρίνοντας την «υπογραφή» της, δηλαδή το αποτύπωμά της, με αποτυπώματα που είναι γνωστό ότι αντιστοιχούν σε ανώμαλη κίνηση [6]. Αντιθέτως, τα NIDS εντοπισμού ανωμαλιών προσπαθούν να ανακαλύψουν τους ενδότερους παράγοντες που διαχωρίζουν την ομαλή από ανώμαλη κίνηση, ώστε να είναι σε θέση να εντοπίσουν επιθέσεις τις οποίες δεν έχουν ξαναδεί [7]. Τα τελευταία είναι αυτά στα οποία θα εστιάσουμε, τα οποία αποτελούν και την τομή του κλάδου του Network Intrusion Detection με το Anomaly Detection.

Αναφορικά με τις τεχνολογίες που χρησιμοποιούνται για Anomaly Based NIDS, η πιο συχνά χρησιμοποιούμενη τεχνολογία στη βιβλιογραφία είναι η απλή μηχανική μάθηση, με τεχνολογίες όπως τα Support Vector Machines (SVMs), δέντρα απόφασης και K-means [8]. Παρόλα αυτά οι μέθοδοι βαθιάς μάθησης διαθέτουν πολύ μεγαλύτερη εκφραστική δύναμη, οπότε και αποτελούν την πιο ελκυστική επιλογή. Αναφορικά με τη μέθοδο εκπαίδευσης, παρατηρούμε ότι επικρατούν οι επιβλεπόμενες μέθοδοι μάθησης [8], στις οποίες τα μοντέλα εκπαιδεύονται σε ομαλά και ανώμαλα δεδομένα με σκοπό την ταξινόμηση κάθε δείγματος στη σωστή κλάση. Εντούτοις, αυτή η χρήση αυτής της μεθόδου δεν είναι ιδιαίτερα επιθυμητή, καθώς περιορίζει το μοντέλο στα δείγματα με τη μορφή των οποίων έχει ήδη έρθει σε επαφή. Άλλωστε, οι επιθέσεις, ως «ανωμαλίες», δεν έχουν πάντα σταθερή μορφή και ενδέχεται η εμφάνισή τους σε ένα ρεαλιστικό σενάριο να μην ανιχνευτεί. Είναι απαραίτητο να υπάρχουν επιλογές για ευέλικτες μεθόδους, οι οποίες μπορούν να γενικεύσουν αποτελεσματικά στο σύνολο των επιθέσεων. Για την επίτευξη αυτού του στόχου, μοντέλα έχουν προταθεί για την ημιεπιβλεπόμενη εκπαίδευση πάνω σε δεδομένα που περιέχουν μόνο ομαλή κίνηση [7, 9]. Εντούτοις, αυτά τα μοντέλα αναμενόμενα εμφανίζουν αρκετά υποδεέστερη απόδοση από αυτά που αξιοποιούν δείγματα ανώμαλης κίνησης, αφού συχνά αγνοούν λεπτές αποκλίσεις των δειγμάτων που σηματοδοτούν ανώμαλη κίνηση.

Για την αντιμετώπιση των παραπάνω προβλημάτων, προτείνουμε μία μέθοδο εκπαίδευσης η οποία έχει τη δυνατότητα να συμφιλιώσει την απαίτηση ευέλικτης μάθησης του μοντέλου με την απαίτηση υψηλής απόδοσης. Αυτή η μέθοδος, η οποία προτάθηκε για πρώτη φορά στο [10] για τη βελτίωση της απόδοσης του μοντέλου Deep SVDD, προσαρμόζεται από εμάς για την ενίσχυση μοντέλων βαθιάς μάθησης στα πλαίσια του Anomaly based NIDS. Ο πυρήνας αυτής της μεθόδου είναι να αποθαρρυνθεί η καλή ανακατασκευή ανώμαλων δειγμάτων. Με αυτόν τον τρόπο, το μοντέλο αποφεύγει να μάθει ευθέως την κατανομή των ανωμαλιών, μαθαίνοντας αντί αυτού την κατανομή των ομαλών δειγμάτων, με τα ανώμαλα δείγματα ως αντιπαραδείγματα. Για να αξιολογήσουμε με ακρίβεια την επίδραση της προσθήκης αντιπαραδειγμάτων με τη μέθοδο αυτή παρατηρούμε τη βελτίωση της απόδοσης σε τέσσερις αρχιτεκτονικές ημιεπιβλεπόμενης μάθησης. Πιο συγκεκριμένα, υλοποιούμε έναν αυτοκωδικοποιητή, έναν παραλλακτικό αυτοκωδικοποιητή με δύο συναρτήσεις απώλειας, και δύο αρχιτεκτονικές παραγωγικών αντιπαραθετικών δικτύων με ενσωματωμένα δίκτυα κωδικοποίησης, μία βασισμένη στο GANomaly [11] και μία ελαφρώς τροποποιημένη έκδοση της αρχιτεκτονικής που παρουσιάζεται στο [12]. Εξετάζουμε τα μοντέλα στα datasets CIC-IDS-2018 [13] και UNSW-NB15 [14, 15, 16, 17, 18]. Από την ανάλυσή μας προκύπτει ότι η προσθήκη αντιπαραδειγμάτων ανώμαλης κίνησης οδηγεί στην δραματική αύξηση της απόδοσης όλων των αρχιτεκτονικών και στα δύο dataset. Επιπλέον, εξετάζουμε τη μέθοδο των αντιπαραδειγμάτων σε μία νέα αρχιτεκτονική, η οποία βασίζεται στον υπολογισμό της συσχέτισης ενός παραθύρου δειγμάτων και τον εντοπισμό των δικτυακών επιθέσεων μέσω μίας αρχιτεκτονικής δισδιάστατου συνελικτικού αυτοκωδικοποιητή. Αυτή η αρχιτεκτονική αποδεικνύεται ανεπαρκής, θεωρούμε όμως απαραίτητο να την αναφέρουμε για λόγους πληρότητας.

Η εργασία αυτή αποτελεί μέρος μίας ευρύτερης ερευνητικής ενέργειας του εργαστηρίου Network Management & Optimal Design Lab (NETMODE) του Εθνικού Μετσόβιου Πολυτεχνείου, μαζί με τις δημοσιεύσεις [19, 20].

Το έργο αυτό διαρθρώνεται ως εξής: Στο κεφάλαιο 2 αναλύονται προσεγγίσεις που έχουν ακολουθηθεί στη σχετική βιβλιογραφία, στο κεφάλαιο 3 παρουσιάζεται μία γενική θεωρητική ανάλυση των εννοιών που είναι κρίσιμες για την κατανόηση της συνεισφοράς μας, στο κεφάλαιο 4 παρατίθεται μία γενική περιγραφή των αρχιτεκτονικών που θα χρησιμοποιήσουμε, στο κεφάλαιο 5 παρέχονται οι λεπτομέρειες υλοποίησης, στο κεφάλαιο 6 αναλύονται τα αποτελέσματα που προκύπτουν, και τέλος στο κεφάλαιο 7 εξάγονται τα τελικά συμπεράσματα και προτείνονται μελλοντικές κατευθύνσεις.

Ο κώδικας που αναπτύχθηκε είναι διαθέσιμος στο:
https://github.com/mark-deligiannis/Diploma-thesis-NIDS

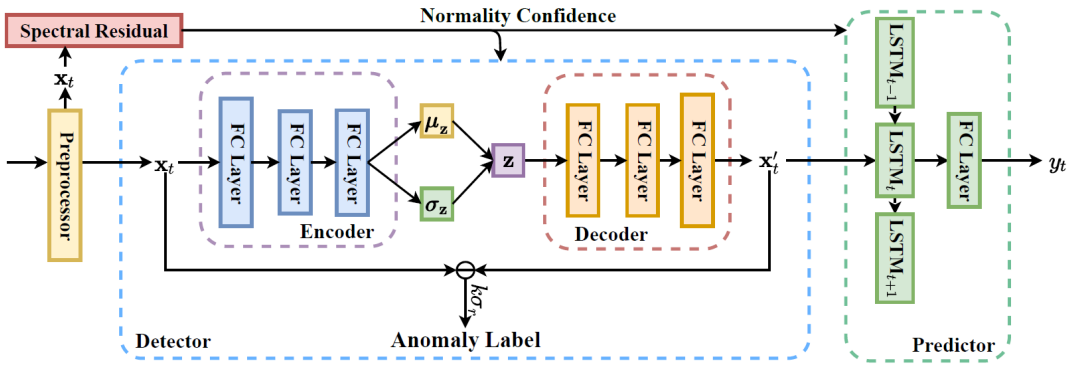
2. Σχετική Βιβλιογραφία

2.1. Γενικό πρόβλημα εντοπισμού ανωμαλιών

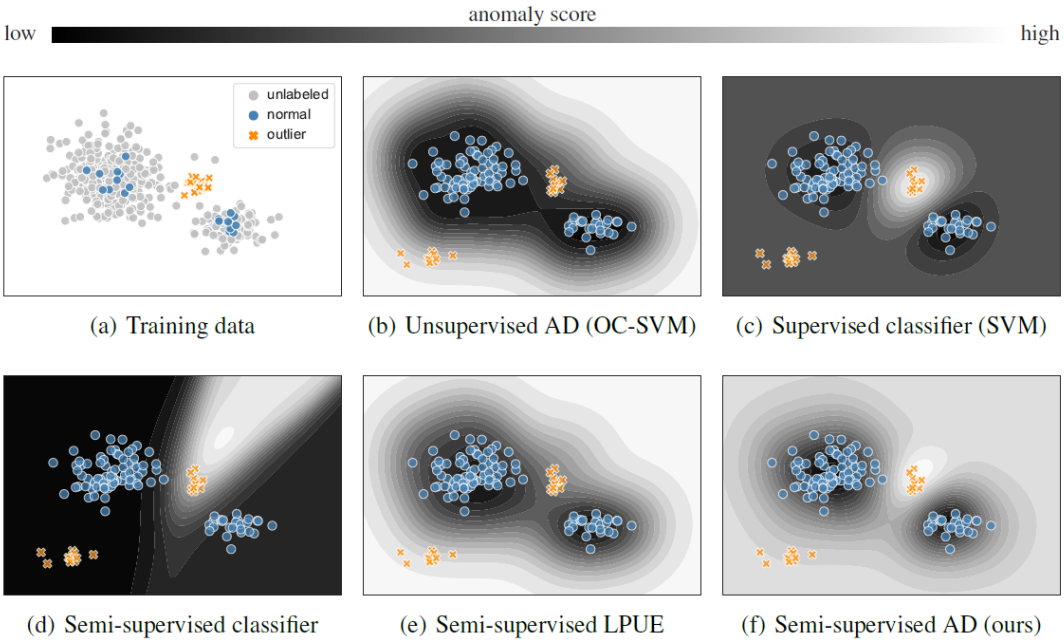
Όπως προαναφέρθηκε, ο εντοπισμός ανωμαλιών αποτελεί έναν ιδιαιτέρως ευρύ ερευνητικό τομέα, με πληθώρα εφαρμογών και μεθόδων που έχουν αναπτυχθεί. Εντούτοις, η γενικότερη νοηματική συνάφεια των διαφορετικών εφαρμογών και η κυριαρχία της τεχνητής νοημοσύνης δίνουν αξία στο να ερευνηθούν μέθοδοι από διαφορετικούς κλάδους εντοπισμού ανωμαλιών και να προσαρμοστούν ώστε να εφαρμόζονται στο Network Intrusion Detection. Προς αυτόν τον σκοπό, κρίνουμε απαραίτητο να παρουσιάσουμε μερικές μεθόδους από διαφορετικά πεδία εφαρμογών που επιτέλεσαν καθοριστικό ρόλο στη διαμόρφωση της παρούσας μεθοδολογίας.

Στο [21], οι ερευνητές παρουσιάζουν ένα Deep Autoencoding Gaussian Mixture Model (DAGMM) για μη επιβλεπόμενο εντοπισμό ανωμαλιών. Το μοντέλο αποτελείται από έναν Autoencoder (AE) για τη συμπίεση των δεδομένων εισόδου, καθώς και ένα δίκτυο εκτίμησης, το οποίο μαθαίνει να προσεγγίζει την πιθανότητα της εμφάνισης του δείγματος στο πλαίσιο των Gaussian Mixture Models (GMM). Καινοτόμα προσέγγιση αποτελεί η αξιοποίηση τόσο του σφάλματος ανακατασκευής, όσο και του λανθάνοντος χώρου για την αξιολόγηση της ομαλότητας του δείγματος. Επιπλέον, το δίκτυο εκτίμησης υλοποιείται ως Feed Forward Neural Network, εξαλείφοντας την ανάγκη για εφαρμογή του αλγόριθμου Estimation Maximization και επιτρέποντας την από άκρο-σε-άκρο εκπαίδευση του μοντέλου. Ως συνέπεια, ο AE ωθείται να μάθει αναπαραστάσεις κατάλληλες για τον διαχωρισμό των ανωμαλιών από τα ομαλά δείγματα. Το DAGMM εξετάζεται σε μία ευρεία συλλογή datasets, και αποδεικνύεται ανώτερο από απλούστερες παραλλαγές του (ablation study), καθώς και από την προηγούμενη τεχνολογία αιχμής.

Στο [22], οι ερευνητές παρουσιάζουν το πρώτο ενοποιημένο μοντέλο για ταυτόχρονο robust prediction και unsupervised Anomaly Detection σε σειρές IT operations. Η ακολουθία ενεργειών που πραγματοποιείται (pipeline) είναι η ακόλουθη: Τα δεδομένα εισόδου κανονικοποιούνται και χωρίζονται αρχικά σε τεμάχια. Έπειτα, ένας Variational AutoEncoder (VAE) χρησιμοποιείται για την ανακατασκευή της εισόδου, και τέλος ένα LSTM για την πρόβλεψη του επόμενου δείγματος. Spectral Residual (SR) χρησιμοποιείται πριν τον VAE μόνο κατά την εκπαίδευση, για την ανάθεση ενός απλού αρχικού anomaly score στο δείγμα, το οποίο αξιοποιείται από τα VAE και Long Short-Term Memory Recurrent Neural Network (LSTM) για robustness. Αυτά απεικονίζονται σχηματικά στο Σχήμα 1. Τα VAE και LSTM εκπαιδεύονται από κοινού με χρήση μίας ενοποιημένης συνάρτησης απώλειας. Το μοντέλο που προκύπτει αξιολογείται ως προς το Prediction (KPI dataset) και το Anomaly Detection (KPI & Yahoo dataset), και επιτυγχάνει αποτελέσματα συγκρίσιμα ή και ανώτερα από απλούστερες παραλλαγές του, καθώς και από το προηγούμενο state-of-the-art, υποδεικνύοντας ότι τα VAE και LSTM λειτουργούν αποδοτικότερα όταν χρησιμοποιούνται μαζί, παρά χωριστά.



Σχήμα 1: Η δομή του μοντέλου που προτείνεται στο [22]

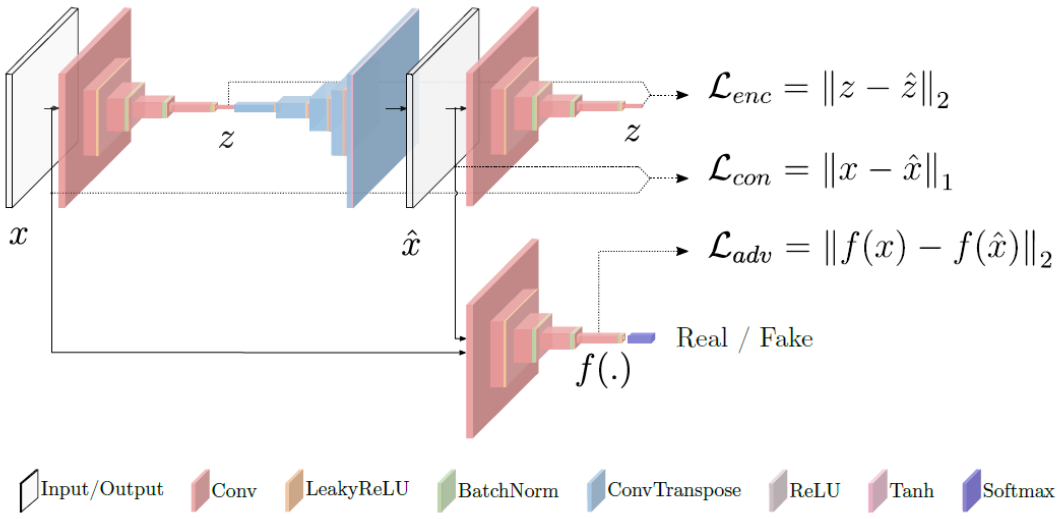


Σχήμα 2: Σύγκριση της επίδοσης του Deep SAD με άλλα μοντέλα της βιβλιογραφίας. Παρατηρούμε ότι το Deep SAD επιτυγχάνει ικανοποιητική ισορροπία μεταξύ της μη επιβλεπόμενης και επιβλεπόμενης μάθησης, αφού εντοπίζει ορθά τις περισσότερες γνωστές ανωμαλίες, χωρίς όμως να περιορίζεται μόνο σε αυτές.

Στο [10], οι ερευνητές παρουσιάζουν το Deep SAD, ένα μοντέλο βαθιάς μάθησης για τον γενικευμένο ημιεπιβλεπόμενο (semi-supervised) εντοπισμό ανωμαλιών. Αρχικά, παρέχεται μία θεωρητική θεμελίωση της μεθόδου βασισμένη στη θεωρία πληροφορίας, σύμφωνα με την οποία στόχος είναι η μεγιστοποίηση της αμοιβαίας πληροφορίας μεταξύ των δεδομένων και της λανθάνουσας αναπαράστασής τους, υπό έναν περιορισμό ομαλοποίησης της τελευταίας που ορίζεται από την εντροπία των δειγμάτων. Το υπό ανάλυση μοντέλο αποτελεί γενίκευση του Deep SVDD με τη δυνατότητα να εκμεταλλεύεται τόσο μη επισημασμένα όσο και επισημασμένα δεδομένα από κάθε κλάση. Αναλυτικότερα, ο στόχος του μοντέλου είναι η πυκνή απεικόνιση των δεδομένων εισόδου σε έναν λανθάνοντα χώρο, στον οποίον ελαχιστοποιείται η απόσταση από ένα σταθερό (μη εκπαιδεύσιμο) σημείο για κανονικά δείγματα ή δείγματα χωρίς ετικέτα, και μεγιστοποιείται για ανώμαλα δείγματα. Το Deep SAD αξιολογείται σε πολλαπλά Anomaly Detection datasets, επιτυγχάνοντας παραπλήσια έως και

ανώτερη επίδοση από το προηγούμενο state-of-the-art. Σχηματική απεικόνιση της επίδοσης του μοντέλου βρίσκεται στο Σχήμα 2.

Τέλος, στο [11] παρουσιάζεται το GANomaly (βλ. Σχήμα 3), ένα μοντέλο γενικευμένου semi-supervised AD, το οποίο είναι τόσο αποτελεσματικό, όσο και χρονικά αποδοτικό. Το μοντέλο αυτό αποτελείται από convolutional autoencoders αρχιτεκτονικής encoder-decoder-encoder οι οποίοι εκπαιδεύονται σε adversarial περιβάλλον με τη χρήση DCGAN (Deep Convolutional GAN). Κατά την εκπαίδευση στον autoencoder ελαχιστοποιείται ο συνδυασμός της απόστασης μεταξύ της εισόδου και της ανακατασκευασμένης εισόδου (contextual loss), της απόστασης μεταξύ των δύο αναπαραστάσεων στον χαμηλοδιάστατο χώρο (encoder loss), και της απόστασης των ενδιάμεσων αναπαραστάσεων του GAN για την είσοδο και την ανακατασκευασμένη είσοδο (adversarial loss). Με αυτόν τον τρόπο, το μοντέλο μαθαίνει ποιοτικές αναπαραστάσεις των δεδομένων. Κατά την πρόβλεψη, το anomaly score υπολογίζεται ως η L1 απόσταση μεταξύ των δύο λανθανόντων αναπαραστάσεων της εισόδου, αξιοποιώντας το γεγονός ότι το μοντέλο δεν μπορεί να αναπαραστήσει ορθά τις ανωμαλίες έχοντας εκπαιδευτεί σε αμιγώς ομαλά δείγματα. Το GANomaly αποδεικνύεται ανώτερο από το προηγούμενο state-of-the-art, τόσο ως προς την ικανότητα πρόβλεψης, όσο και στην υπολογιστική επίδοση.



Σχήμα 3: Η αρχιτεκτονική του μοντέλου GANomaly

2.2. Εντοπισμός ανωμαλιών σε Network Intrusion Detection (NIDS)

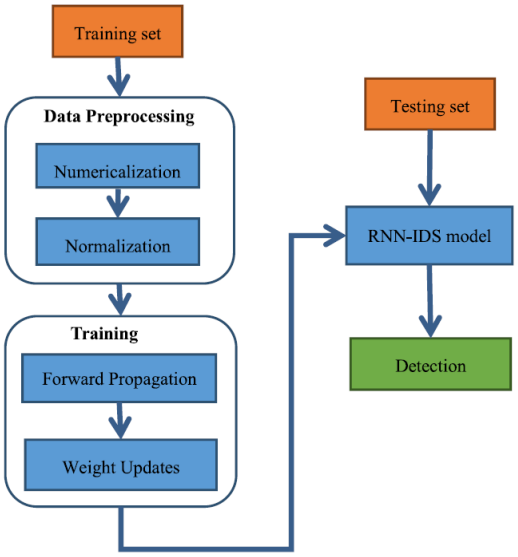
Εστιάζοντας στον τομέα του ενδιαφέροντός μας, εντοπίζουμε πολλά επιδραστικά ερευνητικά έργα στη βιβλιογραφία.

2.2.1. Γενικά μοντέλα μηχανικής μάθησης

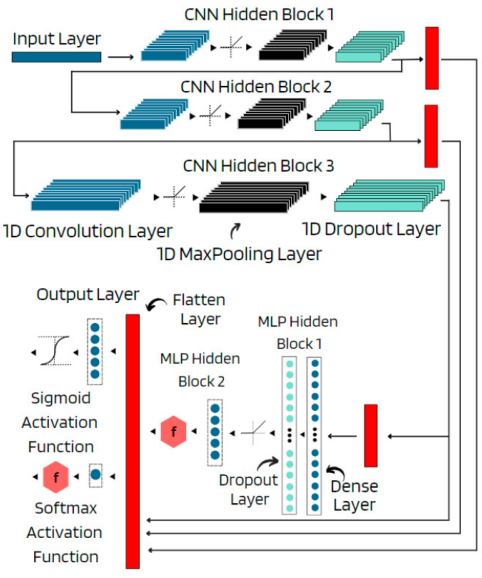
Στο [23], οι ερευνητές μελετούν την απόδοση αλγορίθμων παραδοσιακής επιβλεπόμενης μηχανικής μάθησης στο πρόβλημα του εντοπισμού ανωμαλιών σε δικτυακή κίνηση NetFlow. Επιπλέον, αναλύεται η επίδραση παραγόντων όπως η αναλογία του training προς testing set

και η μέθοδος κωδικοποίησης κατηγορικών χαρακτηριστικών στην απόδοση του μοντέλου. Τα μοντέλα υπό εξέταση είναι τα Stochastic Gradient Descent (SGD), Support Vector Machines (SVM), K-Nearest Neighbor (K-NN), Gaussian Naive Bayes (GNB), Decision Tree (DT), Random Forest (RF), AdaBoost (AB). Επιπλέον, για την κωδικοποίηση κατηγορικών χαρακτηριστικών αξιολογούνται οι μέθοδοι Label Encoding και One Hot encoding, ενώ οι τιμές του λόγου train / test είναι 0.2, 0.3, 0.33, 0.4 και 0.5. Το dataset αξιολόγησης είναι το UNSWNB15, από το οποίο εξάγονται μόνο τα 7 χαρακτηριστικά τα οποία απαντώνται σε κίνηση NetFlow. Για την αξιολόγηση χρησιμοποιούνται πολλές μετρικές, από τις οποίες οι F2-score and AUC αναφέρονται ως οι καταλληλότερες, λόγω του έντονα biased dataset. Από τα πειράματα προκύπτει πως το Label Encoding είναι η πιο αποτελεσματική και χωρικά αποδοτική μέθοδος κωδικοποίησης, το 0.4 είναι ο βέλτιστος λόγος train / test και το μοντέλο με την καλύτερη επίδοση είναι το Random Forest Classifier.

Στο [24], οι ερευνητές παρουσιάζουν το RNN-IDS, ένα μοντέλο επιβλεπόμενης μάθησης για NIDS που αξιοποιεί επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks / RNN). Στο προτεινόμενο pipeline, τα δεδομένα εισόδου υφίστανται αρχικά κατάλληλο preprocessing (numericalization και normalization) και τίθενται στην είσοδο του RNN, το οποίο τα ταξινομεί σε κατηγορίες. Για την εκπαίδευση του μοντέλου χρησιμοποιείται backpropagation. Το τελικό αποτέλεσμα είναι ένα RNN που έχει μάθει όχι μόνο μέσω της διαδικασίας του backpropagation, αλλά και γνωρίζοντας την «ιστορία» των δειγμάτων εισόδου με τη βοήθεια της εσωτερικής του κατάστασης. Η αξιολόγηση γίνεται στο NSL-KDD dataset, καθώς και μία παραλλαγή του που περιέχει τα πιο απαιτητικά δείγματα. Δύο είδη ταξινόμησης εξετάζονται: δυαδική ταξινόμηση (ομαλό / ανώμαλο δείγμα) και ταξινόμηση πολλών κλάσεων (ομαλό / DoS, R2L, U2R, Probe). Το RNN-IDS επιτυγχάνει και στα δύο σενάρια ανώτερη επίδοση από τα προηγούμενα νευρωνικά δίκτυα της βιβλιογραφίας, καθώς και από πολλά μοντέλα παραδοσιακής μηχανικής μάθησης. Σχηματική αναπαράσταση του μοντέλου βρίσκεται στο Σχήμα 4.



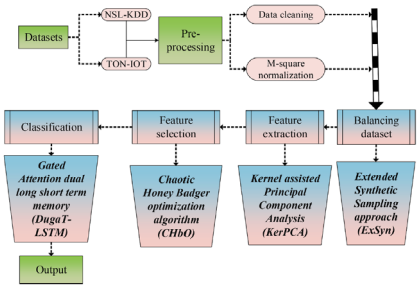
Σχήμα 4: Διάγραμμα block του RNN-IDS



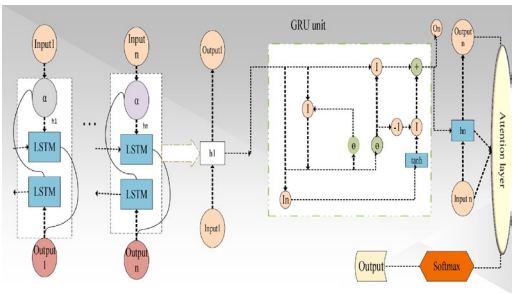
Σχήμα 5: Η δομή του μοντέλου του [25]

Στο [25] (βλ. Σχήμα 5) παρουσιάζεται μία αρχιτεκτονική επιβλεπόμενης μάθησης για τον εντοπισμό ανωμαλιών. Αναλυτικότερα, η αρχιτεκτονική αυτή αξιοποιεί τρία συνελικτικά επίπεδα μίας διάστασης (CNN1D) για την εξαγωγή ιεραρχικών αναπαραστάσεων των δεδομένων φυσιολογικής δικτυακής κίνησης. Τα εξαγόμενα χαρακτηριστικά του τρίτου συνελικτικού επιπέδου διέρχονται μέσα από δύο επιπλέον πλήρως συνδεδεμένα επίπεδα και τα τελικά χαρακτηριστικά σε συνδυασμό με τα χαρακτηριστικά των δύο πρώτων CNN1D επιπέδων τίθενται ως είσοδος στο τελευταίο επίπεδο, στο οποίο λαμβάνεται η απόφαση για το είδος του δείγματος. Για τη διαχείριση της ανισότητας στον αριθμό δειγμάτων κάθε κλάσης αξιοποιείται προσαρμοστική συνθετική δειγματοληψία (ADASYN), καθώς και βάρη κλάσεων, ώστε να ενισχυθούν τα μειοψηφικά δείγματα. Η αρχιτεκτονική αυτή αξιολογείται στο πλήρες NSL-KDD, καθώς και στο υποσύνολό του που περιέχει τα πιο «δύσκολα» δείγματα, επιτυγχάνοντας μετρικές ίσες με 93% και 89% αντίστοιχα, και ξεπερνώντας το σχετικό state-of-the-art.

Στο [26] παρουσιάζεται το Dugat-LSTM, ένα μοντέλο βαθιάς επιβλεπόμενης μάθησης που χρησιμοποιεί τροποποιημένα LSTM σε συνδυασμό με μετα-ευριστικές μεθόδους για εντοπισμό δικτυακών ανωμαλιών. Τα δεδομένα εισόδου υφίστανται καταρχάς M-squared normalization για την αφαίρεση θορύβου και τη διασφάλιση στιβαρότητας. Έπειτα, εφαρμόζεται Extended Synthetic Sampling, μία τεχνική βασισμένη στον αλγόριθμο SMOTE για την αντιμετώπιση της ανισότητας δειγμάτων διαφορετικών κλάσεων. Έπειτα, ανάλυση κύριων συνιστωσών (PCA) με πυρήνα εφαρμόζεται για την εξαγωγή χαρακτηριστικών. Για την τελική επιλογή χαρακτηριστικών χρησιμοποιείται η μετα-ευριστική τεχνική chaotic honey badger optimization, η οποία επιτυγχάνει ικανοποιητική ισορροπία μεταξύ εξερεύνησης και εκμετάλλευσης. Τέλος, τα δεδομένα ταξινομούνται σε κλάσεις με χρήση διπλού LSTM με gated attention, το οποίο χρησιμοποιεί επίπεδα LSTM και GRU. Το pipeline ελέγχεται στα datasets TON-IOT και NSL-KDD, επιτυγχάνοντας ακρίβεια 98.76% και 99.65% αντίστοιχα, και ξεπερνώντας το state-of-the-art. Το pipeline που περιγράψαμε εικονίζεται στο Σχήμα 6 και η αρχιτεκτονική του Dugat-LSTM αναλύεται στο Σχήμα 7.

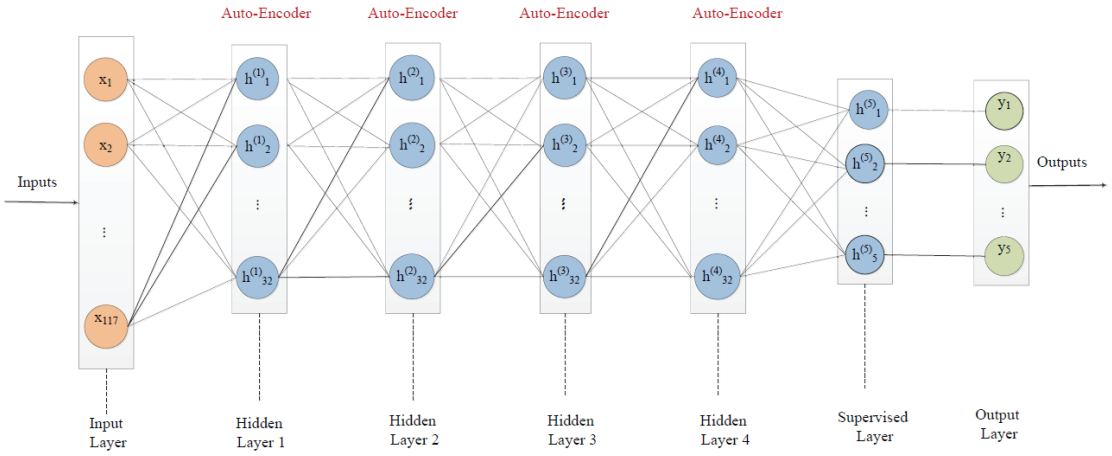


Σχήμα 6: Περιγραφή του pipeline του Dugat-LSTM



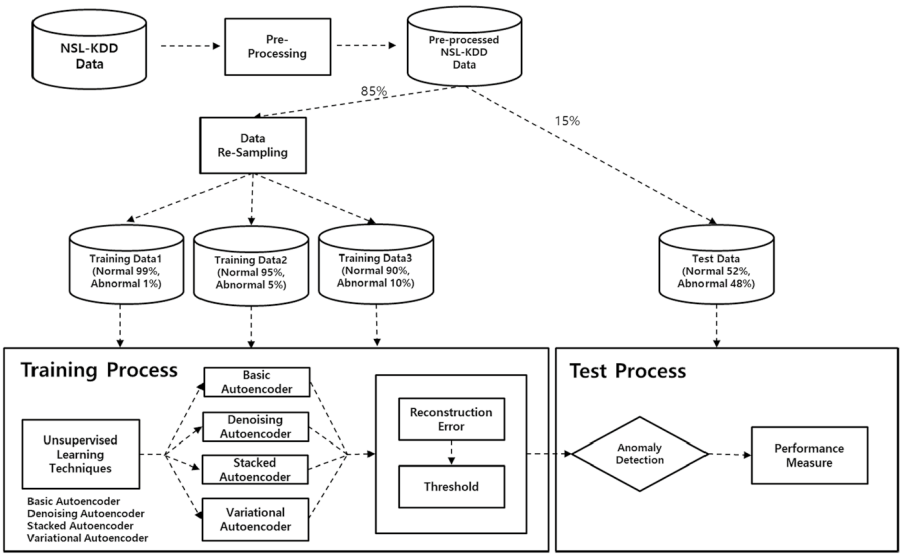
Σχήμα 7: Αρχιτεκτονική του Dugat-LSTM

2.2.2. Με χρήση Auto-Encoders



Σχήμα 8: Η αρχιτεκτονική Deep Auto-Encoder του [27]

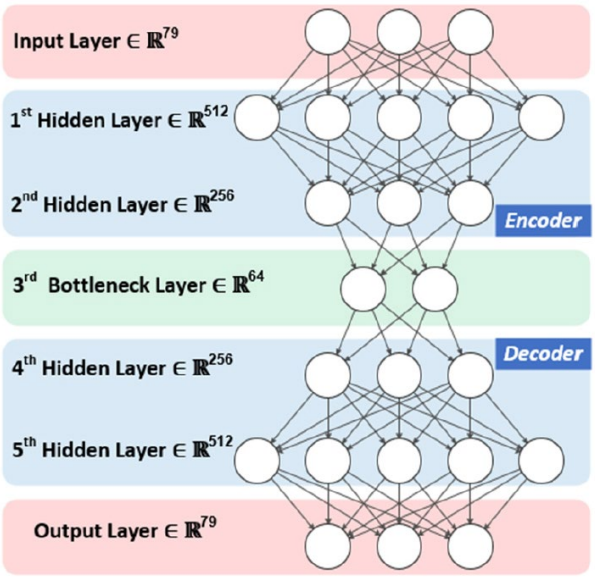
Στο [27] (βλ. Σχήμα 8) προτείνεται η χρήση ενός Deep Auto-Encoder (DAE) για τον εντοπισμό ανωμαλιών. Αναλυτικότερα, τα δείγματα εισόδου διέρχονται από μία σειρά autoencoders, οι οποίοι βαθμιαία συρρικνώνουν τη διάστασή τους. Κάθε εσωτερικός autoencoder εκπαιδεύεται στα δεδομένα λανθάνοντος χώρου του προηγούμενου autoencoder, ανεξάρτητα από αυτόν, με greedy τρόπο. Έτσι επιτυγχάνεται καλύτερη και ταχύτερη σύγκλιση σε ικανοποιητικές παραμέτρους. Η εκπαίδευση των AE γίνεται με μη επιβλεπόμενο τρόπο πάνω στα δεδομένα. Τα δείγματα του λανθάνοντος χώρου του τελευταίου autoencoder τίθενται ως είσοδος σε ένα πυκνό επίπεδο, το οποίο τα ταξινομεί σε κλάσεις και εκπαιδεύεται με επιβλεπόμενο τρόπο. Τέλος, οι παράμετροι όλου του μοντέλου ρυθμίζονται (fine-tuning) μέσω backpropagation, με επιβλεπόμενο τρόπο. Το μοντέλο αυτό αξιολογείται στο dataset KDD-CUP'99, επιτυγχάνοντας ικανοποιητική απόδοση.



Σχήμα 9: Η ερευνητική μέθοδος του [28]

Στο [28] (βλ. Σχήμα 9), οι συγγραφείς ερευνούν την απόδοση τεσσάρων μοντέλων Autoencoders (AE) στον εντοπισμό ανωμαλιών σε μη επιβλεπόμενη ρύθμιση. Αναλυτικότερα,

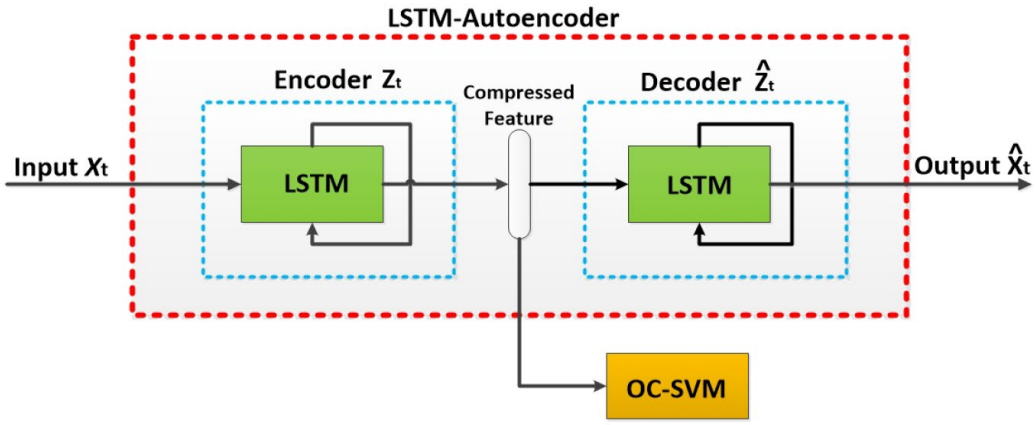
εξετάζονται οι αρχιτεκτονικές απλού Autoencoder, Denoising Autoencoder, ο οποίος προσθέτει γκαουσιανό θόρυβο στα δείγματα και εκπαιδεύει τον AE ώστε να ανακατασκευάζει τα αρχικά δείγματα, Stacked Autoencoder, ο οποίος εκπαιδεύει κάθε επίπεδο του AE χωριστά, και Variational Autoencoder, στον οποίον ο κωδικοποιητής παράγει μία κατανομή στον λανθάνοντα χώρο, αντί για συγκεκριμένο δείγμα. Τα μοντέλα αυτά εκπαιδεύονται στο NSL-KDD dataset, με τρείς ρυθμίσεις: 1%, 5% και 10% περιεκτικότητα σε ανώμαλα δείγματα. Σε κάθε περίπτωση, κατά την αξιολόγηση, το 48% των δειγμάτων είναι γνωστό ότι είναι ανώμαλα, οπότε αυτό το ποσοστό των δειγμάτων με το μεγαλύτερο σφάλμα ανακατασκευής χαρακτηρίζεται ως ανώμαλο. Τα μοντέλα αυτά, και ιδιαίτερα το Stacked και Variational Autoencoder επέδειξαν ικανοποιητική διακριτική ικανότητα, ξεπερνώντας προηγούμενες μη επιβλεπόμενες μεθόδους.



Σχήμα 10: Η αρχιτεκτονική AE και VAE του [9]

Στο [9], οι ερευνητές εξετάζουν τρεις αρχιτεκτονικές μη επιβλεπόμενων μεθόδων βαθιάς μάθησης σε ημι-επιβλεπόμενη ρύθμιση (μόνο ομαλά δείγματα χρησιμοποιούνται για την εκπαίδευση) για τον εντοπισμό ανωμαλιών σε Network Flows. Οι αρχιτεκτονικές αυτές είναι VAE, AE και One Class SVM (OC-SVM). Πιο συγκεκριμένα, στις δύο πρώτες αρχιτεκτονικές, ένας Autoencoder με encoder και decoder 2 επιπέδων (βλ. Σχήμα 10) χρησιμοποιείται για την απεικόνιση της εισόδου στον λανθάνοντα χώρο, και η πιθανότητα ανακατασκευής (VAE) / σφάλμα ανακατασκευής (AE) χρησιμοποιείται ως anomaly score. Στην τρίτη περίπτωση, το OC-SVM χρησιμοποιείται για την απεικόνιση των δεδομένων σε έναν λανθάνοντα χώρο, ελαχιστοποιώντας την απόστασή τους από ένα σημείο. Η απόσταση της αναπαράστασης ενός δείγματος στον λανθάνοντα χώρο από αυτό το σημείο χρησιμοποιείται ως anomaly score. Για την αξιολόγηση αξιοποιείται μία πληθώρα datasets, στα οποία τα μοντέλα καλούνται να διαχωρίσουν την ομαλή κλάση από κάθε είδος επίθεσης χωριστά, καθώς επίσης και από το σύνολο των επιθέσεων (δυαδική ταξινόμηση κάθε φορά). Το VAE εμφανίζει

σαφώς ανώτερη επίδοση στην πλειοψηφία των σεναρίων, το AE παρουσιάζει ικανοποιητική επίδοση, ενώ το OC-SVM εμφανίζει κατηγορηματικά τη χειρότερη επίδοση.

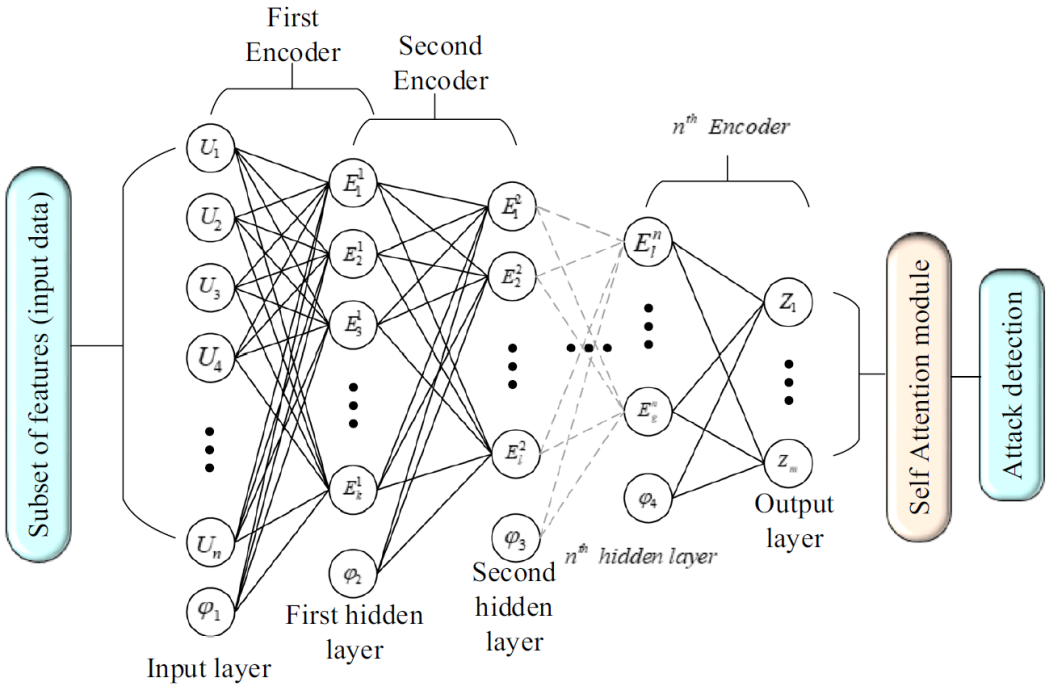


Σχήμα 11: Το μοντέλο LSTM-AE/OC-SVM του [7]

Στο [7] οι ερευνητές προτείνουν την εφαρμογή ενός pipeline που αποτελείται από ένα LSTM Autoencoder σε συνδυασμό με ένα OC-SVM για τον εντοπισμό ανωμαλιών σημείου σε ημι-επιβλεπόμενη ρύθμιση (βλ. Σχήμα 11). Ο LSTM Autoencoder αξιοποιείται για τη μείωση των διαστάσεων της εισόδου, επιτρέποντας στο OC-SVM να μάθει πιο αποτελεσματικά την κατανομή των ομαλών δειγμάτων πάνω σε έναν πυκνότερο λανθάνοντα χώρο. Αυτό γίνεται λαμβάνοντας υπόψη τη χρονική συσχέτιση μεταξύ των δειγμάτων, η οποία αποτελεί χαρακτηριστικό των δικτύων. Όταν ένα ανώμαλο δείγμα παρουσιαστεί, η απόκλιση του από τα ομαλά δείγματα θα αντικατοπτριστεί στη λανθάνουσα αναπαράστασή του και κατ' επέκταση στην απεικόνιση του OC-SVM, επιτρέποντας τον εντοπισμό του. Οι ερευνητές αξιολογούν το μοντέλο στο InSDN, ένα πρόσφατο dataset για Software Defined Networks με δείγματα που αντανακλούν την πραγματική δικτυακή κίνηση σε ικανοποιητικό βαθμό. Εξετάζεται το σενάριο της δυαδικής ταξινόμησης (normal / anomalous traffic). Το μοντέλο εμφανίζει ικανοποιητική απόδοση (ROC-AUC = 0.906), ενώ σε σύγκριση με το LSTM-AE χωρίς OC-SVM και το OC-SVM χωρίς LSTM-AE εμφανίζει ανώτερη απόδοση, αιτιολογώντας την επιλογή των ερευνητών.

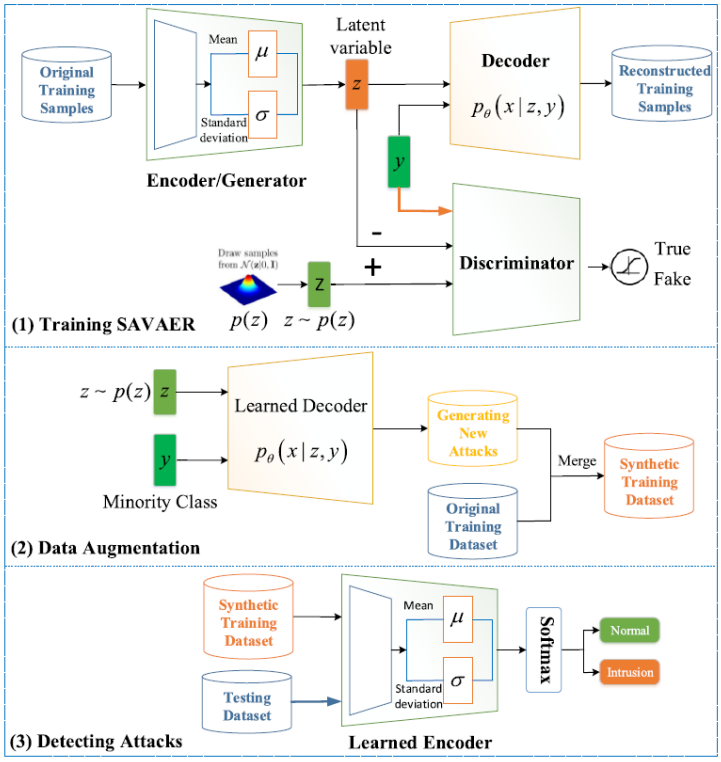
Στο [29] παρουσιάζεται ένα pipeline για Anomaly Detection σε NIDS, το οποίο αποτελείται από μία μέθοδο επιλογής χαρακτηριστικών (extended Pelican Optimization Algorithm / Ex-Pel) και έναν Weighted Autoencoder υποβοηθημένο από Self-Attention (SAttn_WAE) (βλ. Σχήμα 12). Πιο συγκεκριμένα, ο Ex-Pel είναι μία μετα-ευριστική τεχνική επηρεασμένη από το κυνήγι των πελεκάνων. Η λύση που «κυνηγάται» είναι το βέλτιστο υποσύνολο χαρακτηριστικών από το dataset, αυτό δηλαδή που φέρει την πιο χρήσιμη πληροφορία για την ταξινόμηση. Οι τιμές των χαρακτηριστικών που επιλέγονται τελικά προωθούνται στον SAttn_WAE, ο οποίος αποτελεί επέκταση του AE που χρησιμοποιεί Self-Attention για την εστίαση στα σημαντικά δεδομένα, καθώς και έναν όρο ποινικοποίησης της έντονης ενεργοποίησης των νευρώνων στο κρυφό επίπεδο. Για την εκπαίδευση του AE ένα αρχικό greedy pre-training ανά επίπεδο ακολουθείται από fine tuning από άκρο-σε-άκρο. Για την

αξιολόγηση χρησιμοποιούνται τα KDD-CUP 99, NSL-KDD και UNSW-NB15 datasets σε περιβάλλον ταξινόμησης πολλών κλάσεων. Η προτεινόμενη προσέγγιση των ερευνητών αποδεικνύεται ανώτερη από τα προϋπάρχοντα μοντέλα της βιβλιογραφίας.



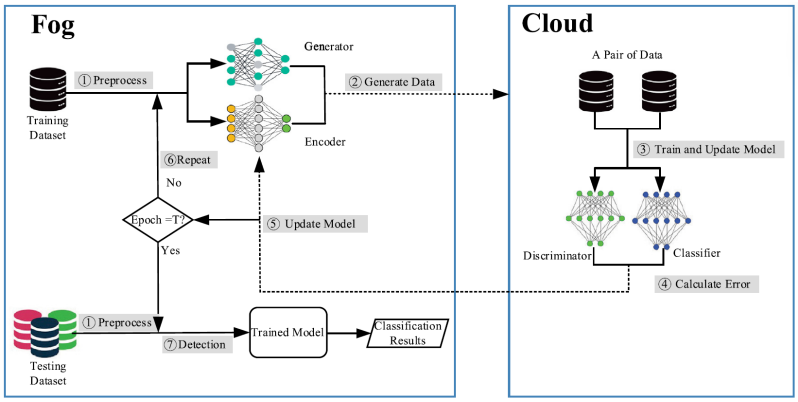
Σχήμα 12: Το Self-Attention assisted Weighted Autoencoder του [29]

Στο [5] παρουσιάζεται το SAVAER, ένα μοντέλο επιβλεπόμενης μάθησης για την αναπλήρωση δειγμάτων ανωμαλιών με ανεπαρκή αναπαράσταση. Το SAVAER απαρτίζεται από έναν Variational AutoEncoder (VAE), ο οποίος αξιοποιεί για την εκπαίδευση τις ετικέτες των δειγμάτων εισόδου και εκπαιδεύεται με adversarial τρόπο. Τα δεδομένα εισόδου κωδικοποιούνται στον λανθάνοντα χώρο από τον κωδικοποιητή (encoder) του VAE, και ανακατασκευάζονται από τον αποκωδικοποιητή (decoder), ο οποίος λαμβάνει υπόψη και την ετικέτα. Επιπλέον, ο κωδικοποιητής λειτουργεί και ως adversarial generator ενός WGAN-GP, με τα δείγματα του λανθάνοντος χώρου να εισάγονται σε έναν διαχωριστή, ο οποίος τα διακρίνει από δείγματα κανονικής κατανομής, κανονικοποιώντας έτσι τις αναπαραστάσεις. Ο αποκωδικοποιητής που προκύπτει από την εκπαίδευση χρησιμοποιείται για την επαύξηση των δεδομένων, αφού μπορεί να παράγει δείγματα από την επιθυμητή κλάση. Τέλος, ο κωδικοποιητής μαζί με ένα επίπεδο softmax πραγματοποιεί τον εντοπισμό ανωμαλιών. Το μοντέλο αξιολογείται στα datasets NSL-KDD και UNSW-NB15, επιτυγχάνοντας ανώτερη απόδοση από το state-of-the-art. Η σχηματική αναπαράσταση του μοντέλου παρουσιάζεται στο Σχήμα 13.



Σχήμα 13: Το σύστημα ανίχνευσης εισβολών SAVAER-DNN του [5]

2.2.3. Με χρήση GAN

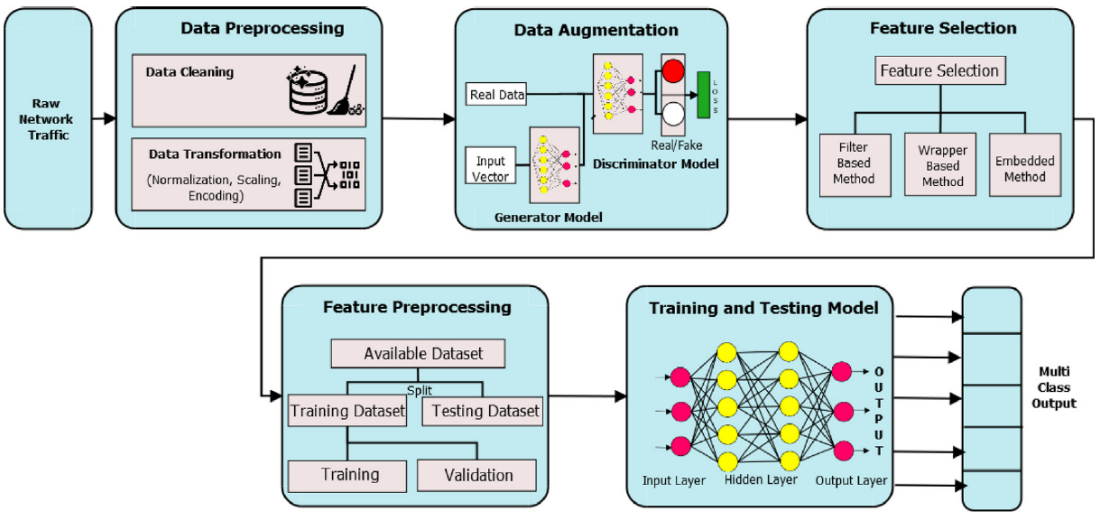


Σχήμα 14: Το pipeline εκπαίδευσης και αξιολόγησης του μοντέλου [12]

Στο [12] παρουσιάζεται ένα σύστημα εντοπισμού ανωμαλιών για το Internet of Things (IoT), το οποίο αξιοποιεί Bidirectional Generative Adversarial Networks (BiGAN) σε περιβάλλον υπολογισμού ομίχλης. Το BiGAN βελτιώνεται με χρήση απόστασης Wasserstein και ποινικοποίηση κλίσης (gradient penalty) ώστε να επιτύχει σταθερότερη εκπαίδευση και καλύτερη απόδοση. Οι ερευνητές προσθέτουν ένα επιπλέον μοντέλο διαχωριστή που εστιάζει μόνο στη λανθάνουσα αναπαράσταση, ενώ το ζεύγος encoder/decoder εκπαιδεύεται ώστε να μειώνει τη διαχωριστική ικανότητα των δύο διαχωριστών, διατηρώντας παράλληλα την ποιότητα ανακατασκευής των δειγμάτων (cycle consistency

loss). Επιπλέον, η διαδικασία εκπαίδευσης είναι υβριδική, με τους κόμβους ομίχλης να εκπαιδεύουν τους encoder / decoder και τον εξυπηρετητή νέφους να εκπαιδεύει τους δύο διαχωριστές. Το anomaly score υπολογίζεται χρησιμοποιώντας την ανακατασκευή από τους encoder / decoder και τις εσωτερικές ενεργοποιήσεις του πρωτεύοντος διαχωριστή. Η διαδικασία εκπαίδευσης και αξιολόγησης φαίνεται και στο Σχήμα 14. Το μοντέλο αξιολογείται στα dataset UNSW-NB15 και CIC-IDS2017, επιτυγχάνοντας 4% βελτίωση στην ακρίβεια και 4% μείωση στο false alarm rate σε σχέση με το προηγούμενο state-of-the-art.

Στο [30] αξιοποιούνται GAN για την καλύτερη επαύξηση δεδομένων (data augmentation), ώστε να αντιμετωπιστεί το πρόβλημα της ανισότητας κλάσεων αποφεύγοντας μεθόδους οι οποίες αλλοιώνουν την κατανομή των μειοψηφικών δειγμάτων. Το pipeline (βλ. Σχήμα 15) περιλαμβάνει αρχικά την προεπεξεργασία των δεδομένων εισόδου, τη χρήση GAN για data augmentation και την υλοποίηση ενός μηχανισμού Feature Selection για την απόρριψη των χαρακτηριστικών που δεν συνεισφέρουν σημαντική επιπλέον πληροφορία για την ταξινόμηση. Αυτό επιτυγχάνεται με χρήση Filter based Method, και ειδικότερα με την απόρριψη χαρακτηριστικών τα οποία έχουν υψηλή συσχέτιση Pearson με άλλα χαρακτηριστικά του dataset. Τέλος, ένα βαθύ νευρωνικό δίκτυο (DNN) 3 πλήρως συνδεδεμένων επιπέδων χρησιμοποιείται για την επιβλεπόμενη μάθηση πάνω στο επεξεργασμένο dataset. Η αρχιτεκτονική αυτή αξιολογείται στο UNSW-NB15 dataset και επιτυγχάνει αύξηση της ακρίβειας από 84% χωρίς χρήση GAN σε 91%, καθώς και την αύξηση όλων των μετρικών για τις μειοψηφικές κλάσεις, αναδεικνύοντας τις δυνατότητες των GAN για την επαύξηση δεδομένων.



Σχήμα 15: Η ροή εργασιών που περιγράφεται στο [30]

Στο [31] οι ερευνητές παρουσιάζουν το IGAN-IDS (Imbalanced GAN), ένα σύστημα εντοπισμού ανωμαλιών με τη δυνατότητα να αντιμετωπίσει το πρόβλημα της ανισότητας κλάσεων σε ad-hoc δίκτυα. Αναλυτικότερα, τα δεδομένα εισόδου, αφού τροφοδοτηθούν σε ένα Feed-Forward Neural Network για την εξαγωγή χαρακτηριστικών, επαυξάνονται με τη χρήση του IGAN. Η εξισορρόπηση των κλάσεων γίνεται εφαρμόζοντας ένα φίλτρο για την

επιλογή μειοψηφικών κλάσεων. Οι ετικέτες της κλάσης που επιλέγεται εισάγονται στον γεννήτορα (generator) του GAN μαζί με τον γκαουσιανό θόρυβο. Ο διαχωριστής (discriminator) επίσης αξιοποιεί την πληροφορία της κλάσης του δείγματος κατά τη λήψη της απόφασής του για τη γνησιότητά του. Έτσι, υπάρχει η δυνατότητα για παραγωγή ποιοτικών δειγμάτων οποιασδήποτε μειοψηφικής κλάσης. Επιπλέον, ο γεννήτορας περιέχει μία μίξη πυκνών και συνελικτικών επιπέδων, ενισχύοντας τη χωρητικότητά του και επιτρέποντάς του την καλύτερη ανακατασκευή των δειγμάτων. Τέλος, τα επαυξημένα δεδομένα οδηγούνται σε ένα βαθύ νευρωνικό δίκτυο, το οποίο λαμβάνει την απόφαση για την κλάση κάθε δείγματος. Η αρχιτεκτονική εξετάζεται στα datasets NSL-KDD, UNSW-NB15 και CICIDS2017, επιτυγχάνοντας ικανοποιητικές αποδόσεις (μέσο Area Under the Curve 95.55%, 97.09% και 99.98% αντίστοιχα) και ξεπερνώντας το state-of-the-art. Στο Σχήμα 16 εικονίζεται η εσωτερική αρχιτεκτονική των μονάδων του IGAN, ενώ στο Σχήμα 17 παρουσιάζεται εποπτικά το πλήρες pipeline του συστήματος ανίχνευσης εισβολών.

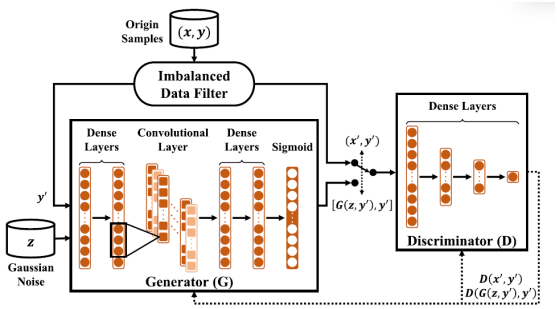
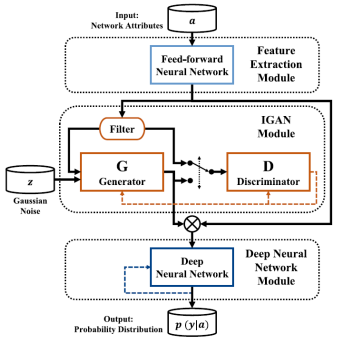
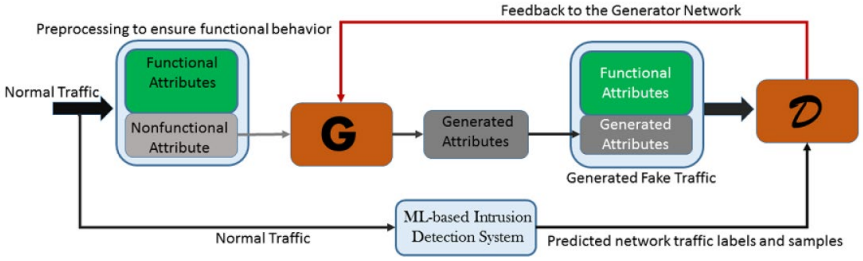


Fig. 1. Model architecture of IGAN.

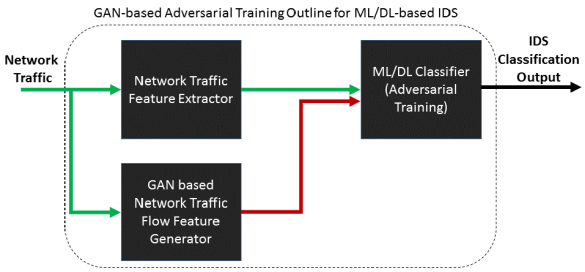
Σχήμα 16: Η αρχιτεκτονική του IGAN



Σχήμα 17: Πλήρες pipeline του συστήματος IGAN-IDS

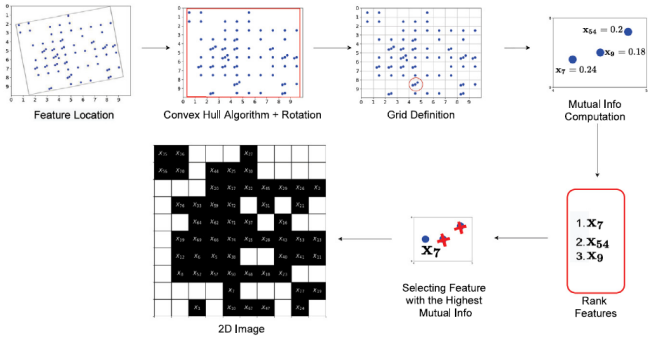


Σχήμα 18: Εκπαίδευση του GAN του [32] για την αποφυγή εντοπισμού από συστήματα ανίχνευσης εισβολών

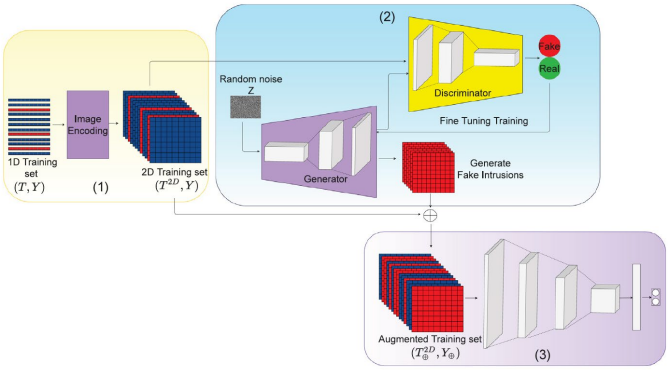


Σχήμα 19: Εποπτεία της μεθόδου αντιπαραθετικής εκπαίδευσης με GAN για τη θωράκιση συστημάτων ανίχνευσης εισβολών του [32]

Στο [32] οι ερευνητές εξετάζουν τον βαθμό στον οποίον μοντέλα παραδοσιακής, αλλά και βαθιάς μηχανικής μάθησης είναι ευάλωτα σε adversarial επιθέσεις. Προς αυτόν τον σκοπό, σχεδιάζεται μία αρχιτεκτονική GAN η οποία μπορεί να παράγει συνθετικά δείγματα τα οποία πλήττουν την απόδοση οποιουδήποτε δικτύου εντοπισμού ανωμαλιών, χωρίς γνώση της εσωτερικής του αρχιτεκτονικής. Καταρχάς, τα χαρακτηριστικά της δικτυακής ροής χωρίζονται σε λειτουργικά και μη λειτουργικά, όπου μόνο τα τελευταία μπορούν να υποστούν αλλαγές χωρίς τα λογικά χαρακτηριστικά της ροής να αλλάξουν. Έπειτα, ο γεννήτορας του GAN εκπαιδεύεται ώστε να δέχεται ως είσοδο μη λειτουργικά χαρακτηριστικά ανώμαλης κίνησης και να παράγει συνθετικά δείγματα που παραμένουν ανώμαλα, αλλά ταξινομούνται λανθασμένα από τον διαχωριστή. Ο τελευταίος εκπαιδεύεται ώστε, πέρα από τον εντοπισμό των δειγμάτων που παράγει ο γεννήτορας, να προσομοιώνει την είσοδο του μοντέλου «μαύρου κουτιού» που στοχοποιείται. Η προαναφερθείσα αρχιτεκτονική, η οποία φαίνεται σχηματικά στο Σχήμα 18, αξιολογείται στο NSL-KDD και επιτυγχάνει σημαντική επιδείνωση των μετρικών απόδοσης όλων των μοντέλων που εξετάζονται. Επιπλέον, οι ερευνητές παρουσιάζουν μία μέθοδο για τη θωράκιση των συστημάτων αυτών, στην οποία το υποψήφιο μοντέλο «μαύρο κουτί» εκπαιδεύεται σε υβριδικά δεδομένα τα οποία περιλαμβάνουν δεδομένα που παράγει το προηγούμενο GAN μοντέλο. Η εκπαίδευση με αυτόν τον τρόπο, η οποία εικονίζεται στο Σχήμα 19, πρακτικά επαναφέρει την απόδοση των μοντέλων, και αποδεικνύεται πιο αποτελεσματική από την απλή adversarial εκπαίδευση.



Σχήμα 20: Η τεχνική του MAGNETO για τη μετατροπή μονοδιάστατων χαρακτηριστικών σε δισδιάστατα



Σχήμα 21: Η πλήρης αρχιτεκτονική του MAGNETO. Η μονάδα (1) μετατρέπει τα μονοδιάστατα σε δισδιάστατα δεδομένα, η μονάδα (2) παράγει νέα δεδομένα επίθεσης για επαύξηση και εξισορρόπηση του συνόλου δεδομένων, και η μονάδα (3) εκπαιδεύεται πάνω στα επαυξημένα δεδομένα για τον εντοπισμό των ανωμαλιών.

Στο [33] οι ερευνητές παρουσιάζουν το MAGNETO, μία μέθοδο επαύξησης δεδομένων μειοψηφίας για την καλύτερη αναγνώριση όλων των επιθέσεων, με χρήση αρχιτεκτονικών GAN και CNN δύο διαστάσεων. Καινοτομία αποτελεί η μεθοδολογία μετατροπής των δεδομένων εισόδου σε δύο διαστάσεις, κατά την οποία εφαρμόζεται tSNE στον ανάστροφο πίνακα δεδομένων, απεικονίζοντας κάθε χαρακτηριστικό σε ένα σημείο στον δισδιάστατο χώρο. Έπειτα, το μικρότερο περιστραμμένο παραλληλόγραμμο προσαρμόζεται στα δείγματα του χώρου, και χωρίζεται σε pixels. Κάθε χαρακτηριστικό απεικονίζεται στο αντίστοιχο pixel του δισδιάστατου χώρου. Η διαδικασία αυτή εικονίζεται στο Σχήμα 20. Με αυτόν τον τρόπο, η δύο διαστάσεων αναπαράσταση του δείγματος που προκύπτει κωδικοποιεί λογική πληροφορία στη θέση των σημείων, επιτρέποντας την αποτελεσματική εφαρμογή συνελικτικών μοντέλων. Μετά τη μετατροπή, ένα επιβλεπόμενο GAN με συνελικτικά δίκτυα εφαρμόζεται για την αναπλήρωση ελλειμματικών δειγμάτων και ένα CNN2D πραγματοποιεί την ταξινόμηση. Το πλήρες pipeline φαίνεται και στο Σχήμα 21. Η αρχιτεκτονική αυτή αξιολογείται με επιτυχία στα datasets KDDCUP99, UNSW-NB15, CICIDS17 και AAGM17.

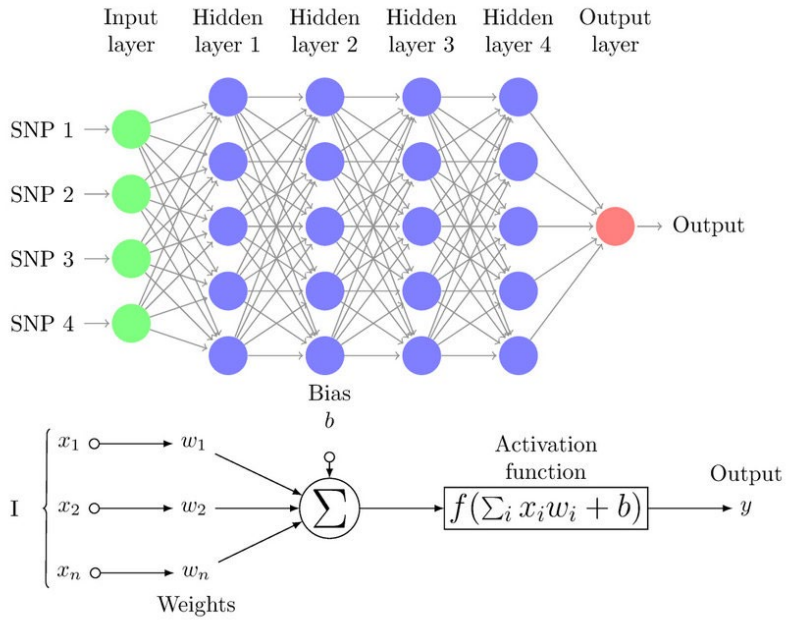
3. Θεωρητικό Πλαίσιο Ανάλυσης

3.1. Βασικές αρχιτεκτονικές

Τα σύγχρονα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται χαρακτηρίζονται από έντονη αρχιτεκτονική πολυπλοκότητα, έχουν όμως ορισμένα κοινά χαρακτηριστικά. Σε αυτήν την ενότητα θα περιγράψουμε συνοπτικά δύο βασικά συστατικά των μοντέλων βαθιάς μάθησης, το πολυστρωματικό αντίληπτρο, και τα συνελικτικά νευρωνικά δίκτυα. Έπειτα, θα έχουμε κατοχυρώσει το απαραίτητο θεωρητικό υπόβαθρο για να θίξουμε αρχιτεκτονικές υψηλότερης πολυπλοκότητας, οι οποίες θα χρησιμοποιηθούν σε αυτή τη διπλωματική εργασία.

3.1.1. Πολυστρωματικό αντίληπτρο (Multi-Layer Perceptron / MLP)

Στον πυρήνα σχεδόν κάθε αρχιτεκτονικής βαθιάς μάθησης βρίσκεται το πολυστρωματικό αντίληπτρο (MLP) [34, pp. 116-120]. Το MLP αποτελεί ένα δίκτυο οργανωμένο σε επίπεδα. Κάθε επίπεδο περιέχει πολλαπλές μονάδες (ή αλλιώς νευρώνες), οι οποίες λαμβάνουν στην είσοδό τους το διάνυσμα τιμών εξόδου του προηγούμενου επιπέδου, υπολογίζουν το εσωτερικό του γινόμενο με ένα διάνυσμα προσαρμοστικών βαρών, προσθέτουν μία σταθερή τιμή και υπολογίζουν την έξοδο μίας συνάρτησης, της *συνάρτησης ενεργοποίησης*, για την τιμή που προκύπτει. Επειδή η ροή της πληροφορίας γίνεται από την είσοδο προς τα διαφορετικά επίπεδα διαδοχικά, χωρίς ανατροφοδοτήσεις, τα δίκτυα αυτά ονομάζονται και δίκτυα πρόσθιας τροφοδότησης (Feed-Forward Neural Networks). Σχηματικά έχουμε την εξής δομή:



Σχήμα 22: Η δομή του MLP [35]

Η μαθηματική αναπαράσταση γίνεται ως εξής: Έστω $\boldsymbol{x}$ το διάνυσμα τιμών εισόδου. Για κάθε επίπεδο k, συμβολίζουμε με $\boldsymbol{z_k} = [z_{k,1}, z_{k,2}, \dots, z_{k,d_k}]$ το διάνυσμα εξόδου και με $\boldsymbol{a^{(k)}} = \left[a_1^{(k)}, a_2^{(k)}, \dots, a_{d_k}^{(k)}\right]$ την έξοδο του επιπέδου πριν την εφαρμογή της συνάρτησης ενεργοποίησης. d_k είναι το πλήθος των μονάδων εξόδου του k-οστού επιπέδου. Συμβολίζουμε επίσης με $w_{ji}^{(k)}, i > 0$ το βάρος της σύνδεσης της j-οστής μονάδας του k-οστού επιπέδου με την i-οστή μονάδα του (k-1)-οστού επιπέδου, και $w_{j0}^{(k)}$ τον σταθερό όρο (bias) που αντιστοιχεί στη j-οστή μονάδα του k-οστού επιπέδου. Ορίζουμε επίσης $\boldsymbol{z_0} = \boldsymbol{x}$.

Με δεδομένα τα παραπάνω, έχουμε ότι για το k-οστό επίπεδο με είσοδο $\boldsymbol{z_{k-1}}$ η έξοδος $\boldsymbol{z_k}$ υπολογίζεται ως εξής:

$$a_j^{(k)} = \sum_{i=1}^{d} w_{ji}^{(k)} \cdot z_{k-1,i} + w_{j0}^{(k)} \quad \text{και}$$

$$\boldsymbol{z_k} = g\left(\boldsymbol{a^{(k)}}\right),$$

όπου g είναι η συνάρτηση ενεργοποίησης, η οποία εφαρμόζεται χωριστά σε κάθε στοιχείο του διανύσματος $\boldsymbol{a^{(k)}}$.

3.1.2. Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Network / CNN)

Τα συνελικτικά νευρωνικά δίκτυα (CNNs) αποτελούν ειδική κατηγορία πολυστρωματικών αντίληπτρων. Η αρχιτεκτονική αυτή όπως είναι γνωστή σήμερα, με εκπαίδευση με backpropagation [36], προτάθηκε για πρώτη φορά από τους LeCun κ.ά. [37] για το πεδίο της αναγνώρισης εικόνων, όμως γρήγορα βρήκαν εφαρμογή σε πολλαπλούς τομείς, μεταξύ των οποίων και ο εντοπισμός ανωμαλιών.

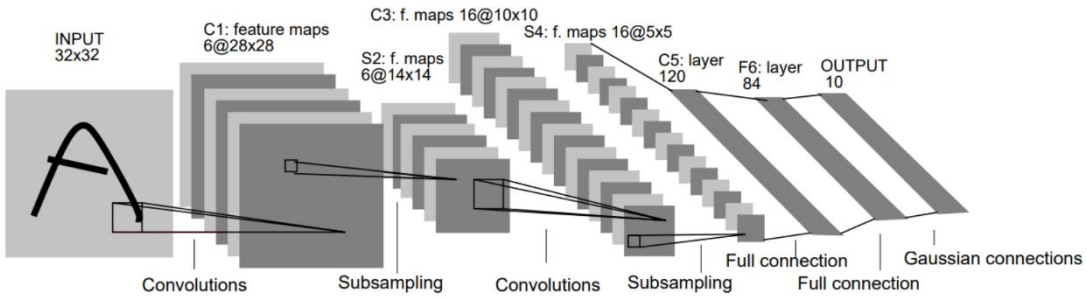
Σύμφωνα με το [38, pp. 276-281], ένα CNN συνήθως λαμβάνει τρισδιάστατη είσοδο, με τις διαστάσεις να αντιστοιχούν στις σειρές, στήλες, και «κανάλια» της εισόδου (συνήθως 3 για έγχρωμες εικόνες). Η είσοδος στη συνέχεια υφίσταται μία σειρά βημάτων επεξεργασίας, τα οποία ονομάζονται επίπεδα ή στρώματα. Ένα τυπικό συνελικτικό δίκτυο εναλλάσσει στρώματα συνέλιξης και υποδειγματοληψίας. Όταν η έξοδος συρρικνωθεί αρκετά διέρχεται από ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα. Τα στρώματα συνέλιξης και υποδειγματοληψίας εξάγουν ιεραρχικά χαρακτηριστικά από την είσοδο και τα πλήρως συνδεδεμένα επίπεδα τα αξιοποιούν για τη λήψη του τελικού αποτελέσματος.

Χάριν συντομίας, θα αναλύσουμε μόνο τη λειτουργία του συνελικτικού στρώματος, καθώς αυτό αποτελεί τον πυρήνα του συνελικτικού δικτύου.

Κάθε συνελικτικό στρώμα αποτελείται από ένα σύνολο C *χαρτών χαρακτηριστικών*, ουσιαστικά δισδιάστατων πλεγμάτων νευρώνων διαστάσεων $N \times N$. Ο νευρώνας στη θέση i, j του k-οστού χάρτη χαρακτηριστικών υλοποιεί τη συνάρτηση:

$$y_{ij}^{k} = f\left(\sum_{l=1}^{C'}\sum_{a=1}^{m}\sum_{\beta=1}^{m} w_{a,\beta,l}^{k} x_{i-a,j-\beta}^{l} + b^{k}\right),$$

όπου C' το πλήθος χαρτών χαρακτηριστικών του προηγούμενου στρώματος. Κάθε νευρώνας δέχεται πληροφορίες από μία μικρή περιοχή $m \times m$ με κέντρο τη θέση του i, j. Ο πίνακας $\boldsymbol{W^k} = \left[w_{a,\beta,l}^{k}\right]$ καλείται *συνελικτική μάσκα* για τον k-οστό χάρτη χαρακτηριστικών. Η πράξη μέσα στην παρένθεση ονομάζεται *συνέλιξη*. Το αποτέλεσμα της συνέλιξης διέρχεται από τη *συνάρτηση ενεργοποίησης* $f(\cdot)$, η οποία εισάγει μη γραμμικότητα στη διαδικασία. Σημαντική παρατήρηση αποτελεί ότι, σε αντίθεση με τα πλήρως συνδεδεμένα δίκτυα, στα οποία κάθε ζεύγος νευρώνων διαδοχικών επιπέδων έχουν διαφορετικό βάρος, στα συνελικτικά δίκτυα τα βάρη αφορούν μετατοπίσεις στη θέση των νευρώνων, με αποτέλεσμα να επαναλαμβάνονται έντονα. Αποτέλεσμα αυτού είναι η έντονη μείωση παραμέτρων του μοντέλου, καθώς και η ανίχνευση χαρακτηριστικών ανεξάρτητα από τη θέση τους στην εικόνα εισόδου.

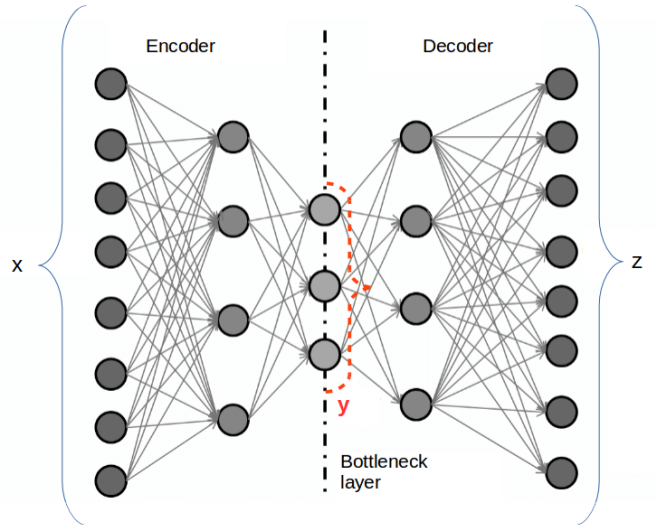


Σχήμα 23: Η αρχιτεκτονική του LeNET-5 [39]

3.2. Σύνθετες αρχιτεκτονικές - Αυτοκωδικοποιητές

3.2.1. Αυτοκωδικοποιητής (Autoencoder / AE)

Σύμφωνα με το [40] οι αυτοκωδικοποιητές προτάθηκαν αρχικά από τον LeCun για τη διδακτορική του διατριβή [41]. Ο αυτοκωδικοποιητής αποτελείται από δύο στοιχεία: τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Ο κωδικοποιητής απεικονίζει την είσοδο σε έναν συνήθως χαμηλοδιάστατο χώρο, ο οποίος καλείται *λανθάνων χώρος*, ενώ ο αποκωδικοποιητής απεικονίζει δείγματα του λανθάνοντος χώρου στον χώρο της εισόδου.



Σχήμα 24: Σχηματική αναπαράσταση αυτοκωδικοποιητή. Πηγή: [42]

Οι αυτοκωδικοποιητές ωθούνται να αναπαράγουν στην έξοδό τους τα δείγματα εισόδου. Τότε ο λανθάνων χώρος περιλαμβάνει την απαραίτητη πληροφορία για ανακατασκευή των δειγμάτων, γεγονός που είναι χρήσιμο για πληθώρα εφαρμογών, όπως συμπίεση δεδομένων και εξαγωγή χαρακτηριστικών.

Περιγράφουμε τώρα τα ανωτέρω με χρήση άλγεβρας, όπως στο [40]:

Έστω σύνολο εκπαίδευσης $S = \{x_i | x_i \in R^d\}, 1 \leq i \leq n$, τότε ο αυτοκωδικοποιητής μοντελοποιείται από την εξίσωση:

$$\begin{cases} z = f(w_e, b_e; x) \\ r = g(w_d, b_d; z) \end{cases}$$

Όπου $f(\cdot), g(\cdot)$ είναι οι συναρτήσεις του κωδικοποιητή και αποκωδικοποιητή αντίστοιχα, οι οποίες υλοποιούνται με χρήση νευρωνικών δικτύων. Οι παράμετροι w_e, b_e είναι ρυθμίσιμες παράμετροι του κωδικοποιητή, οι οποίες αντιστοιχούν στα βάρη των παραμέτρων και τους σταθερούς όρους, ενώ οι w_d, b_d είναι οι αντίστοιχες παράμετροι του αποκωδικοποιητή. Σκοπός της εκπαίδευσης είναι η ρύθμιση των παραμέτρων αυτών για την βέλτιστη ανακατασκευή των δειγμάτων εισόδου, και πιο συγκεκριμένα για την ελαχιστοποίηση της συνάρτησης:

$$J(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} \|x_i - r_i\|_2^2 , \quad \theta = (w_e, b_e; w_d, b_d)$$

Το παραπάνω πρόβλημα ελαχιστοποίησης επιλύεται με χρήση back-propagation και εφαρμογή κάποιου γνωστού βελτιστοποιητή (όπως Adam [43] ή SGD [44]).

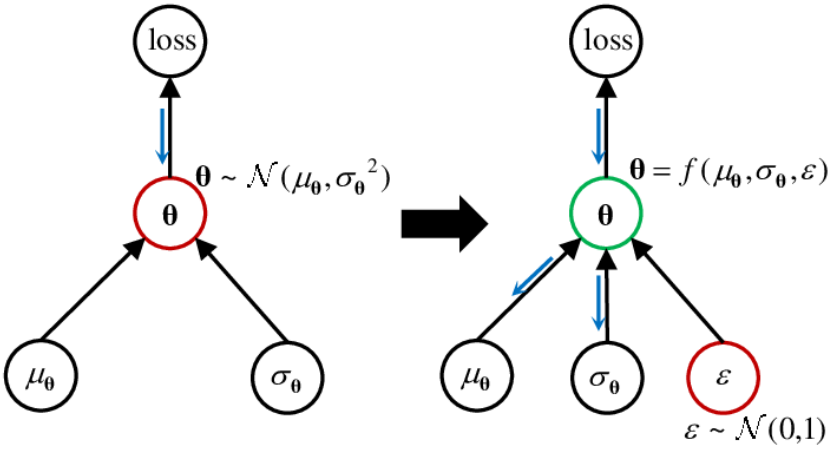
3.2.2. Παραλλακτικός αυτοκωδικοποιητής (Variational Autoencoder / VAE)

3.2.2.1. Βασική αρχιτεκτονική

Μία παραλλαγή του αυτοκωδικοποιητή με πληθώρα επιτυχών εφαρμογών είναι ο παραλλακτικός αυτοκωδικοποιητής (VAE). Σύμφωνα με το [45, pp. 278-283], οι παραλλακτικοί αυτοκωδικοποιητές επιβάλλουν μία στοχαστική δομή στις κρυφές μονάδες. Πιο συγκεκριμένα, το σύνολο των δειγμάτων του λανθάνοντα χώρου ωθούνται να ακολουθούν μία συγκεκριμένη κατανομή, η οποία τις περισσότερες φορές είναι πολυμεταβλητή γκαουσιανή. Στην παρακάτω ανάλυση υποθέτουμε γκαουσιανή κατανομή. Ο κωδικοποιητής αντί για τον υπολογισμό συγκεκριμένων δειγμάτων του λανθάνοντα χώρου υπολογίζει τις παραμέτρους της γκαουσιανής κατανομής, δηλαδή δύο διανύσματα που αντιστοιχούν στις μέσες τιμές και τις τυπικές αποκλίσεις της κατανομής. Ο αποκωδικοποιητής καλείται να ανακατασκευάσει την είσοδο του κωδικοποιητή λαμβάνοντας ως είσοδό του ένα δείγμα του λανθάνοντα χώρου. Συνεπώς, κατά τη σύνδεση των δύο στοιχείων πρέπει να πραγματοποιηθεί δειγματοληψία στον λανθάνοντα χώρο. Συμβολίζοντας με k τη διάσταση του λανθάνοντα χώρου, $\bar{\mu}(\bar{X})$, $\bar{\sigma}(\bar{X})$ τα υπολογισμένα από τον κωδικοποιητή διανύσματα μέσων τιμών και τυπικών αποκλίσεων του δείγματος $\bar{X}$, η δειγματοληψία πραγματοποιείται εκλέγοντας ένα $\varepsilon \sim N(0, I)$, όπου I είναι ο $k \times k$ πίνακας ταυτότητας, και υπολογίζοντας την ακόλουθη ποσότητα:

$$\bar{h}(\bar{X}) = \varepsilon \odot \bar{\sigma}(\bar{X}) + \bar{\mu}(\bar{X}).$$

Η μέθοδος αυτή ονομάζεται «τέχνασμα επαναπαραμετροποίησης». Ο περιορισμός της τυχαιότητας στην μεταβλητή ε, αντί για την απευθείας εκλογή $\bar{h}(\bar{X}) \sim N\left(\bar{\mu}(\bar{X}), \bar{\sigma}^2(\bar{X})\right)$, επιτρέπει τη ροή των κλίσεων στον κωδικοποιητή κατά το backpropagation, και κατ'επέκταση την από άκρο-σε-άκρο εκπαίδευση του μοντέλου.



Σχήμα 25: Το τέχνασμα επαναπαραμετροποίησης. Αριστερά – Στοχαστικός κόμβος με απευθείας εκλογή μεταβλητής. Η πληροφορία των κλίσεων δεν μπορεί να διέλθει στον κωδικοποιητή. Δεξιά – Ντετερμινιστικός κόμβος με χρήση επαναπαραμετροποίησης. Η πληροφορία των κλίσεων μπορεί να περάσει στον κωδικοποιητή, θεωρώντας το ε σταθερό. Πηγή: [46]

Η εκπαίδευση του παραλλακτικού αυτοκωδικοποιητή γίνεται με τον συνδυασμό δύο στόχων, τη μεγιστοποίηση της πιθανότητας εμφάνισης των δειγμάτων εκπαίδευσης, και την ώθηση

του λανθάνοντος χώρου να ακολουθεί την πολυμεταβλητή τυποποιημένη κανονική κατανομή $N(0, I)$. Ο πρώτος στόχος αντιστοιχεί στη «μάθηση» των δειγμάτων που πρέπει να ανακατασκευάζονται. Ο δεύτερος στόχος εξυπηρετεί ως ένας στοχαστικός περιορισμός κανονικοποίησης του μοντέλου, που ωθεί σε ουσιαστικότερη μάθηση, αλλά και δυνατότητα δημιουργίας έγκυρων δειγμάτων χρησιμοποιώντας μόνο τον αποκωδικοποιητή και τροφοδοτώντας τον με δείγματα κανονικής κατανομής. Η μεγιστοποίηση της πιθανότητας εμφάνισης των δειγμάτων εκπαίδευσης συνήθως υλοποιείται ως η ελαχιστοποίηση της τετραγωνικής απώλειας ανακατασκευής:

$$L = \|\bar{X} - \bar{X}'\|^2,$$

όπου το $\bar{X}'$ είναι το ανακατασκευασμένο δείγμα εισόδου από τον αποκωδικοποιητή. Η απώλεια κανονικοποίησης R μαθηματικοποιείται μέσω της απόκλισης Kullback-Leibler (KL) της υπό συνθήκη πιθανότητας με τις παραμέτρους που παράγει ο κωδικοποιητής, σε σχέση με την k-διάστατη τυποποιημένη κανονική κατανομή. Η τιμή της είναι:

$$R = \frac{1}{2}\left(\|\bar{\mu}(\bar{X})\|^2 + \|\bar{\sigma}(\bar{X})\|^2 - 2\sum_{i=1}^{k} \ln(\bar{\sigma}(\bar{X})_i) - k\right).$$

Η συνολική απώλεια δίνεται από τον τύπο:

$$J = L + \lambda R,$$

όπου $\lambda > 0$ είναι η παράμετρος κανονικοποίησης. Ο τυπικός VAE έχει $\lambda = 1$, ενώ όταν χρησιμοποιείται γενικευμένο λ αναφερόμαστε στον αποκωδικοποιητή ως β-VAE [47].

3.2.2.2. Βελτιωμένη συνάρτηση απώλειας

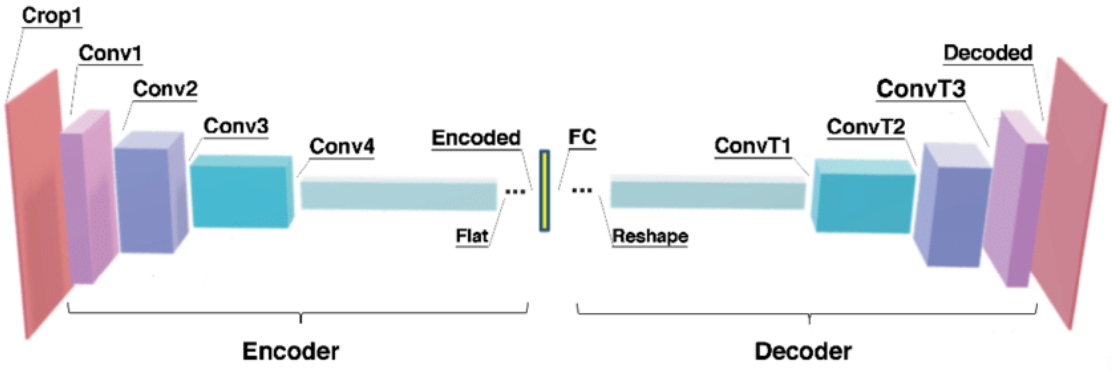
Οι C.P. Burgess κ.α. [48] αναλύουν θεωρητικά την απόδοση του β-VAE, παρατηρώντας ότι η αύξηση της τιμής β (λ στον δικό μας συμβολισμό) αυξάνει τη δυνατότητα για εκμάθηση ποιοτικότερων, αποπολυπλεγμένων αναπαραστάσεων θυσιάζοντας την ποιότητα ανακατασκευής, ενώ η μείωση του β έχει την αντίστροφη επίδραση. Για την καταπολέμηση αυτής της αντίστροφης σχέσης ποιότητας αναπαραστάσεων και ποιότητας ανακατασκευής, οι συγγραφείς προτείνουν την εξής απώλεια (απλοποιημένη ώστε να ενσωματώνει τις προηγούμενες παραδοχές μας):

$$L(\bar{X}, \bar{X}') = \|\bar{X} - \bar{X}'\|^2 + \gamma\left|\frac{1}{2}\left(\|\bar{\mu}(\bar{X})\|^2 + \|\bar{\sigma}(\bar{X})\|^2 - 2\sum_{i=1}^{k} \ln(\bar{\sigma}(\bar{X})_i) - k\right) - C\right|,$$

όπου το C είναι μία ελεγχόμενη παράμετρος πληροφοριακής χωρητικότητας, η οποία αυξάνεται γραμμικά από μία ελάχιστη μέχρι μία μέγιστη τιμή. Η ενθάρρυνση του μοντέλου να αυξάνει σταδιακά τη χωρητικότητά του με αυτόν τον τρόπο συμφιλιώνει τις δύο επιθυμητές ιδιότητες που προηγουμένως ήταν αντικρουόμενες, παράγοντας μία πιο στιβαρή έκδοση του παραλλακτικού αυτοκωδικοποιητή.

3.2.3. Συνελικτικός αυτοκωδικοποιητής (Convolutional Autoencoder)

Ο συνελικτικός αυτοκωδικοποιητής (CAE ή ConvAE), σύμφωνα με το [40] είναι μια λογική επέκταση του απλού αυτοκωδικοποιητή, με μόνη διαφορά ότι ο κωδικοποιητής και αποκωδικοποιητής αντί για πλήρως συνδεδεμένα επίπεδα αξιοποιούν συνελικτικά «μπλοκ», τα οποία αποτελούνται συνήθως από ένα συνελικτικό επίπεδο, ακολουθούμενο από ένα δίκτυο υποδειγματοληψίας. Στο επίπεδο «στενωπού» (bottleneck layer) συνήθως εντοπίζονται πλήρως συνδεδεμένα επίπεδα, όπως στα παραδοσιακά συνελικτικά δίκτυα. Οι λεπτομέρειες της πράξης της συνέλιξης έχουν ήδη περιγραφεί στη σχετική ενότητα. Σημειώνουμε ότι, επειδή τα συνελικτικά νευρωνικά δίκτυα εμπίπτουν στην κατηγορία των δικτύων πρόσθιας τροφοδότησης, ο αλγόριθμος back-propagation μπορεί να εφαρμοστεί κανονικά για την εκπαίδευσή τους. Οι συνελικτικοί αυτοκωδικοποιητές είναι καταλληλότεροι σε εφαρμογές όπως η αναγνώριση εικόνας, στις οποίες ο μηχανισμός εξαγωγής ιεραρχικών χαρακτηριστικών αμετάβλητων προς τις μετατοπίσεις, που παρέχουν τα συνελικτικά δίκτυα, είναι πολύτιμος.



Σχήμα 26: Μία τυπική αρχιτεκτονική συνελικτικού αυτοκωδικοποιητή. Πηγή: [49]

3.2.4. Σύνθετες αρχιτεκτονικές – Παραγωγικά αντιπαραθετικά δίκτυα (Generative Adversarial Networks / GAN)

Σε αυτήν την ενότητα αναλύουμε μία άλλη οικογένεια αρχιτεκτονικών βαθιάς μάθησης, που ξεκίνησε από την αναγνώριση εικόνας, αλλά αξιοποιήθηκε σε πληθώρα εφαρμογών με μεγάλη επιτυχία, τα παραγωγικά αντιπαραθετικά δίκτυα (GAN).

3.2.4.1. Απλή αρχιτεκτονική GAN

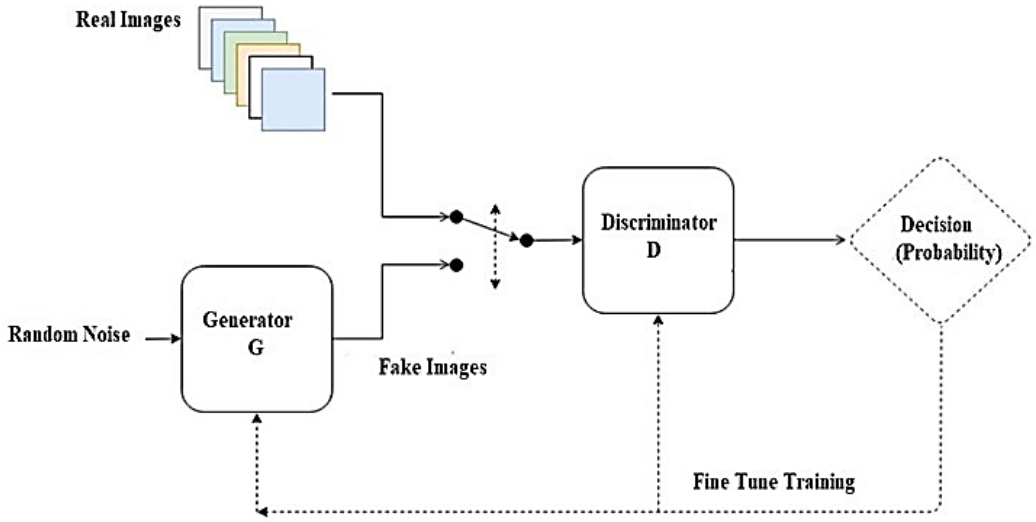
Τα παραγωγικά αντιπαραθετικά δίκτυα δημιουργήθηκαν από τους Ian J. Goodfellow κ.ά. [50], παρέχοντας έναν καινοτόμο τρόπο για την αποδοτική μοντελοποίηση της κατανομής των δεδομένων εισόδου. Ένα απλό GAN αποτελείται από δύο συστατικά μοντέλα, τον γεννήτορα (generator) και τον διαχωριστή (discriminator), τα οποία μοντελοποιούνται ως δίκτυα

πρόσθιας τροφοδότησης. Ο γεννήτορας λαμβάνει στην είσοδό του θόρυβο a priori κατανομής $p_z(z)$, συνήθως πολυμεταβλητής τυποποιημένης κανονικής κατανομής, και παράγει υποψήφια δείγματα. Ο διαχωριστής καλείται να διαχωρίσει τα δείγματα που παράγονται από τον γεννήτορα από πραγματικά δείγματα.

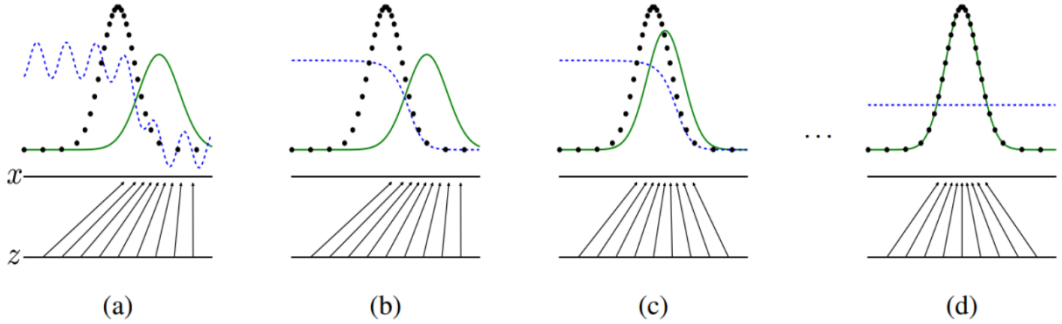
Η εκπαίδευση μπορεί να υλοποιηθεί ως το ακόλουθο min-max παιχνίδι μεταξύ του γεννήτορα G και του διαχωριστή D, με συνάρτηση $V(G,D)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\big(G(z)\big)\right)\right]$$

Αναλυτικότερα, ο γεννήτορας και διαχωριστής εναλλάξ τροποποιούν τις παραμέτρους τους, ώστε ο διαχωριστής να μπορέσει να διακρίνει καλύτερα τα παραγόμενα από τον G δείγματα, και ο γεννήτορας να μπορέσει να προσαρμόζεται ώστε να μην ανιχνεύεται από τον διαχωριστή.



Σχήμα 27: Τυπική αρχιτεκτονική GAN. Πηγή: [51]



Σχήμα 28: Οι έξοδοι των μονάδων ενός GAN σε διάφορα στάδια εκπαίδευσης. Οι μαύρες κουκκίδες αντιστοιχούν στην κατανομή των πραγματικών δεδομένων, η πράσινη γραμμή στην παραγόμενη κατανομή του γεννήτορα, και η μπλε γραμμή στην αυτοπεποίθηση του διαχωριστή για τη γνησιότητα του δείγματος. Στην αρχή (a) ο γεννήτορας και ο διαχωριστής έχουν ουσιαστικά τυχαία συμπεριφορά. Στη συνέχεια (b) - (c), ο γεννήτορας πλησιάζει την κατανομή των δεδομένων, ενώ ο διαχωριστής ορθά προσαρμόζεται ώστε να διακρίνει τις δύο κατανομές. Στο τέλος (d), παρουσιάζεται το ιδανικό σενάριο σύγκλισης, στο οποίο η κατανομή των δειγμάτων του γεννήτορα ταυτίζεται με αυτή των πραγματικών δειγμάτων, και ο διαχωριστής αναγκάζεται να επιλέξει τυχαία. Πηγή: [50]

Δεδομένου ότι ο αλγόριθμος εκπαίδευσης αποκλίνει από το απλό backpropagation των προηγούμενων αρχιτεκτονικών, κρίνουμε απαραίτητο να τον παρουσιάσουμε, όπως ακριβώς αυτός περιγράφεται από τους Goodfellow κ.ά.:

Αλγόριθμος εκπαίδευσης:
for *πλήθος εποχών*:
  for *k βήματα*:
    1) Λάβε μίνι παρτίδα *m* δειγμάτων θορύβου, $\{z^{(1)}, \dots, z^{(m)}\}$ από κατανομή $p_g(z)$
    2) Λάβε μίνι παρτίδα *m* δειγμάτων $\{x^{(1)}, \dots, x^{(m)}\}$ από την πραγματική κατανομή $p_{data}(x)$
    3) Ανανέωσε τις παραμέτρους του διαχωριστή ανεβαίνοντας την κλίση:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \right]$$

  End-for
  Λάβε μίνι παρτίδα *m* δειγμάτων θορύβου από κατανομή $p_g(z)$
  Ανανέωσε τις παραμέτρους του γεννήτορα κατεβαίνοντας την κλίση:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)$$

End-for

Οι Goodfellow κ.ά. αποδεικνύουν ότι, με την προϋπόθεση ότι τα δύο συστατικά μοντέλα έχουν την απαιτούμενη χωρητικότητα, σε κάθε βήμα ο διαχωριστής είναι βέλτιστος για τον τρέχοντα γεννήτορα, και σε κάθε βήμα του γεννήτορα αυτός μειώνει την απόδοση του βέλτιστου θεωρητικού διαχωριστή, τότε η κατανομή που επάγει ο γεννήτορας συγκλίνει στην κατανομή των δειγμάτων.

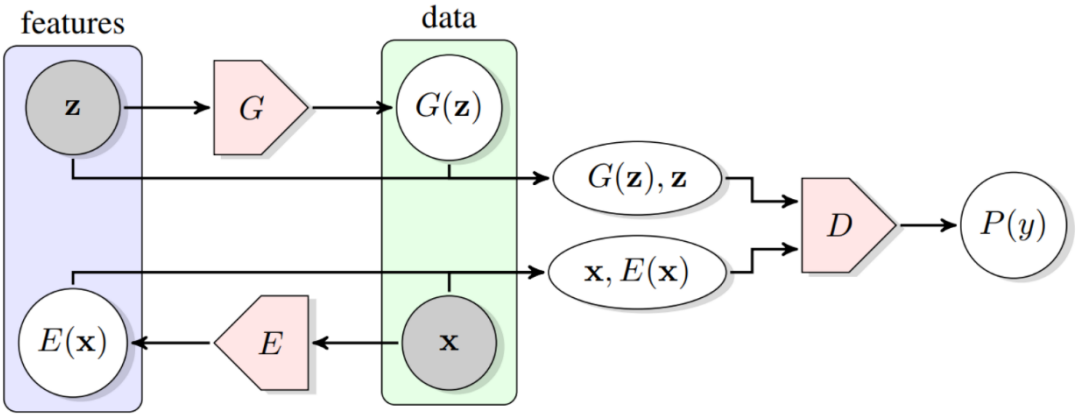
Παρά τις υψηλές δυνατότητες της αρχιτεκτονικής που περιγράψαμε, υπάρχουν ορισμένοι περιορισμοί. Καταρχάς, σε αντίθεση με τους αυτοκωδικοποιητές, δεν υπάρχει εύκολος τρόπος γρήγορης ανακατασκευής ενός δείγματος εισόδου. Έτσι, στο πρόβλημα του εντοπισμού ανωμαλιών, αν κανείς θέλει να ελέγξει εάν ένα δείγμα είναι ομαλό, δηλαδή αν ανήκει στην κατανομή των ομαλών δειγμάτων, και έχει στη διάθεσή του ένα GAN που έχει «μάθει» την κατανομή αυτή, πρέπει να ακολουθήσει την ακόλουθη διαδικασία [52]:

Να λάβει ένα σύνολο δειγμάτων του λανθάνοντος χώρου και να υπολογίσει την έξοδο του γεννήτορα για κάθε ένα από αυτά. Έπειτα, να επιλέξει το δείγμα από αυτά που έχει την ελάχιστη απόσταση από το δείγμα προς ανακατασκευή, και μετά να εκτελέσει αλγόριθμο κατάβασης κλίσης στον γεννήτορα ώστε να βρει το δείγμα του λανθάνοντος χώρου που αντιστοιχεί στο παραγόμενο από τον γεννήτορα δείγμα με την ελάχιστη απόσταση από το επιθυμητό. Τέλος, να ελέγξει αν αυτή η ανακατασκευή είναι ικανοποιητική (μικρό σφάλμα ανακατασκευής) ή όχι (μεγάλο σφάλμα ανακατασκευής), και να αποφανθεί για τον χαρακτήρα του δείγματος υπό εξέταση. Αυτό είναι μια χρονοβόρα διαδικασία με συχνά απαγορευτικό κόστος. Στην επόμενη ενότητα περιγράφουμε μία παραλλαγή της αρχιτεκτονικής που μας επιτρέπει να υπερβούμε αυτόν τον περιορισμό.

Δύο πρόσθετες αδυναμίες της τρέχουσας αρχιτεκτονικής είναι το πρόβλημα της κατάρρευσης τρόπου, κατά την οποία ο γεννήτορας αρκείται στο να εστιάζει σε ένα περιορισμένο υποσύνολο δειγμάτων υψηλής ποιότητας, αγνοώντας την πλήρη κατανομή των δεδομένων [53], και το πρόβλημα της αστάθειας της εκπαίδευσης, κατά την οποία ο γεννήτορας και ο διαχωριστής διαρκώς ταλαντώνονται μεταξύ μη ικανοποιητικών τιμών παραμέτρων, χωρίς να επιτυγχάνεται σύγκλιση της κατανομής του γεννήτορα στην πραγματική κατανομή των δεδομένων [53]. Στις επόμενες ενότητες θα παρουσιάσουμε μοντέλα που καταπολεμούν αυτά τα προβλήματα.

3.2.4.2. Αμφίδρομα GAN (BiGAN)

Στο [54], οι ερευνητές αναγνωρίζουν την αδυναμία των GAN να παράγουν αποδοτικά τον αντίστροφο μετασχηματισμό, αυτόν δηλαδή από τον χώρο των δεδομένων στον λανθάνοντα χώρο, και προτείνουν την αρχιτεκτονική των αμφίδρομων GAN.



Σχήμα 29: Η δομή των αμφίδρομων GAN. Πηγή: [54]

Τα BiGAN περιλαμβάνουν, πέρα από τον γεννήτορα και τον διαχωριστή, μία επιπλέον μονάδα, τον κωδικοποιητή (encoder). Ο κωδικοποιητής είναι ένα δίκτυο πρόσθιας τροφοδότησης, υπεύθυνο για τον αντίστροφο μετασχηματισμό του γεννήτορα. Ο διαχωριστής τροποποιείται ώστε να λαμβάνει στην είσοδο τον συνδυασμό δεδομένων και αντίστοιχων δειγμάτων του λανθάνοντα χώρου, και να αποφαίνεται για την πιθανότητα αυτά να έχουν παραχθεί από πραγματικά δεδομένα που κωδικοποιήθηκαν, ή δείγματα του λανθάνοντα χώρου που αποκωδικοποιήθηκαν. Παρά το γεγονός ότι, βάσει της παραπάνω περιγραφής, ο γεννήτορας και ο κωδικοποιητής δεν επικοινωνούν ευθέως, οι συγγραφείς αποδεικνύουν ότι οι δύο αυτές μονάδες πρέπει να «μάθουν» να αντιστρέφει η μία την άλλη ώστε να ελαχιστοποιήσουν την ικανότητα του διαχωριστή να διακρίνει τα δείγματα των δύο κατανομών.

Με μαθηματικούς όρους, έστω G, E, D ο γεννήτορας, κωδικοποιητής και διαχωριστής αντίστοιχα, p_x, p_z οι κατανομές των δεδομένων και δειγμάτων του λανθάνοντα χώρου αντίστοιχα, και $p_E(z|x), p_G(x|z)$ οι κατανομές που επάγουν ο κωδικοποιητής και γεννήτορας αντίστοιχα. Ο τροποποιημένος στόχος εκπαίδευσης του BiGAN είναι ο ακόλουθος στόχος minimax:

$$\min_{G,E} \max_{D} V(D, E, G),$$

όπου

$$V(D, E, G) \coloneqq \mathbb{E}_{x \sim p_x}\left[\mathbb{E}_{z \sim p_E(\cdot|x)}[\log D(x, z)]\right] + \mathbb{E}_{z \sim p_z}\left[\mathbb{E}_{x \sim p_G(\cdot|z)}\left[\log\left(1 - D(x, z)\right)\right]\right].$$

Η διαδικασία εκπαίδευσης είναι παρόμοια με το απλό GAN.

3.2.4.3. Wasserstein GAN (WGAN)

Τα Wasserstein GAN προτάθηκαν από τους Martin Arjovksky κ.ά. [53] για την αντιμετώπιση προβλημάτων του παραδοσιακού GAN, όπως η κατάρρευση τρόπου και η ασταθής εκπαίδευση, όπως προαναφέραμε. Σε αυτά, η αρχιτεκτονική του δικτύου παραμένει ίδια και τροποποιείται μόνο ο στόχος εκπαίδευσης. Αντί για την προαναφερθείσα συνάρτηση στόχο, η οποία αντιστοιχεί στην απόκλιση Jensen-Shannon (JS), χρησιμοποιείται η παρακάτω αντικειμενική συνάρτηση, υπό την οποία ο βέλτιστος διαχωριστής υπολογίζει την Earth-Mover (EM) απόσταση, ή Wasserstein-1 απόσταση:

$$W\left(\mathbb{P}_r, \mathbb{P}_g\right) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]\right),$$

όπου το supremum είναι πάνω στις 1-Lipschitz συναρτήσεις $f: \mathrm{X} \to \mathbb{R}$.

Για την επίτευξη του παραπάνω υλοποιούμε την f ως ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης, τον διαχωριστή, και εκπαιδεύουμε ώστε να επιτύχουμε μεγιστοποίηση του όρου εντός του supremum. Για την υλοποίηση του περιορισμού η f να είναι συνάρτηση K-Lipschitz (η χαλάρωση σε K-Lipschitz δεν επηρεάζει τα αποτελέσματα) οι συγγραφείς προτείνουν την *αποκοπή βαρών*, ώστε τα βάρη του μοντέλου να παραμένουν στο διάστημα $[-c, c]$, όπου c υπερπαράμετρος. Ο αλγόριθμος εκπαίδευσης είναι ίδιος με το απλό GAN, με αλλαγή του στόχου εκπαίδευσης ώστε να είναι ο στατιστικός όρος εντός του supremum, όπως προαναφέραμε, και να χρησιμοποιείται αποκοπή βαρών. Οι τροποποιήσεις αυτές αποδεικνύονται αρκετές για να περιορίσουν σε σημαντικό βαθμό το πρόβλημα της κατάρρευσης τρόπου, αλλά και της ασταθούς εκπαίδευσης. Παρά ταύτα, τα προβλήματα αυτά δεν αντιμετωπίζονται πλήρως. Αυτό, σε συνδυασμό με το γεγονός ότι οι συγγραφείς σημειώνουν ότι η αποκοπή βαρών είναι σχετικά αδρό μέτρο υλοποίησης του περιορισμού που χρησιμοποιείται μόνο επειδή δεν βρέθηκαν καλύτερες λύσεις, οδηγεί στην ανάπτυξη της αρχιτεκτονικής που θα παρουσιάσουμε στην επόμενη υποενότητα.

3.2.4.4. Wasserstein GAN με ποινή κλίσεων (WGAN-GP)

Για την αντιμετώπιση των προβλημάτων του απλού WGAN που προαναφέραμε, οι Ishaan Gulrajani κ.ά. [55] προτείνουν το WGAN-GP, μία αρχιτεκτονική που αντικαθιστά την προβληματική αποκοπή βαρών του παραδοσιακού WGAN με την *ποινή κλίσεων* (gradient penalty). Αυτή υλοποιείται με τον ακόλουθο τρόπο:

Πέρα από τις κατανομές των πραγματικών δεδομένων $\mathbb{P}_r$ και αυτών που παράγονται από τον γεννήτορα $\mathbb{P}_g$ ορίζεται μία τρίτη κατανομή, η $\mathbb{P}_{\hat{x}}$ δειγματοληπτώντας ομοιόμορφα κατά

μήκος ευθύγραμμων τμημάτων μεταξύ ζευγών σημείων από τις δύο πρώτες κατανομές. Η κλίση του διαχωριστή σε σημεία αυτής της κατανομής ενθαρρύνεται να λαμβάνει τιμές κοντά στη μονάδα. Με μαθηματικούς όρους, η τροποποιημένη συνάρτηση απώλειας που ελαχιστοποιεί ο διαχωριστής είναι:

$$L = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda \cdot \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}}\left(D(\hat{x})\right)\right\|_2 - 1\right)^2\right],$$

όπου λ υπερπαράμετρος εξισορρόπησης του όρου ποινής κλίσεων.

Σημειώνουμε πως, σε αντίθεση με τις συναρτήσεις-στόχους που αναφέραμε στις προηγούμενες υποενότητες, η συνάρτηση αυτή είναι συνάρτηση απώλειας, οπότε ο διαχωριστής επιδιώκει να την ελαχιστοποιήσει.

Ο γεννήτορας αγνοεί την ποινή κλίσης, και προσπαθεί μόνο να μεγιστοποιήσει τον πρώτο όρο της εξίσωσης που παραθέτουμε.

Σημαντική παρατήρηση: Οι συγγραφείς αναφέρουν πως η δομή του μοντέλου καθιστά τη χρήση ομαλοποίησης παρτίδας (batch normalization [56]*) προβληματική. Αντί αυτού, προτείνουν την αποφυγή κανονικοποίησης ή χρήση ομαλοποίησης επιπέδου (layer normalization* [57]*).*

4. Η συνεισφορά μας

Όπως αναδείχθηκε στην βιβλιογραφική μας επισκόπηση, ο τομέας του εντοπισμού ανωμαλιών σε δίκτυα υπολογιστών έχει δει πληθώρα προσεγγίσεων με χρήση πολλών εργαλείων, όπως αυτοκωδικοποιητές, παραλλακτικούς αυτοκωδικοποιητές, παραγωγικά αντιπαραθετικά δίκτυα και επαναλαμβανόμενα δίκτυα (RNNs). Παρόλα αυτά, εντοπίσαμε μία πόλωση στις προσεγγίσεις, και πιο συγκεκριμένα στον τρόπο εκπαίδευσης. Όπως προαναφέραμε, παρά τη χρήση όλων των μεθόδων εκπαίδευσης (με επίβλεψη, με μερική επίβλεψη, χωρίς επίβλεψη), η εκπαίδευση γίνεται είτε πάνω στα δείγματα με μη επιβλεπόμενο τρόπο, είτε πάνω στα ομαλά δείγματα, είτε πάνω σε όλα τα δείγματα με πλήρως επιβλεπόμενο τρόπο. Η τελευταία προσέγγιση μάλιστα βρίσκεται σε ισχυρή αντίφαση με τον χαρακτήρα της «ανωμαλίας», αφού η ανωμαλία είναι η μη ομαλότητα, και ως εκ τούτου δεν έχει συγκεκριμένο χαρακτήρα. Δείγματα κίνησης μπορούν να είναι ανώμαλα με εντελώς διαφορετικό τρόπο, καθιστώντας την προσπάθεια μάθησης της μη ομαλότητας μη παραγωγική. Εντούτοις, η εκπαίδευση μόνο πάνω στα ομαλά δείγματα είναι δύσκολο να πετύχει να συλλάβει την ουσία της «ομαλότητας», ιδιαίτερα στον τομέα της ανίχνευσης εισβολών σε δίκτυα, όπου οι ανωμαλίες συχνά έχουν πολύ λεπτές διαφορές από την ομαλή κίνηση.

Σε αυτή τη διπλωματική εργασία, εξερευνούμε έναν τρόπο να ενισχύσουμε την ικανότητα ενός μοντέλου ημιεπιβλεπόμενης μάθησης να αποφεύγει την σωστή ανακατασκευή ανώμαλων δειγμάτων. Αυτό γίνεται μέσω αντιπαραδειγμάτων, χωρίς όμως να περιορίζεται η αντιληπτική ικανότητα του μοντέλου μόνο στα αντιπαραδείγματα αυτά. Αυτή η μέθοδος έχει χρησιμοποιηθεί νωρίτερα στο [10], όμως δεν γνωρίζουμε κανένα έργο στη βιβλιογραφία που έχει εξετάσει την επίδραση αυτής της τροποποίησης σε πιο επίκαιρα μοντέλα βαθιάς μάθησης (όπως τους αυτοκωδικοποιητές) στον τομέα του εντοπισμού ανωμαλιών δικτύου. Θα δείξουμε πως η μέθοδος αυτή μπορεί να οδηγήσει σε έντονη βελτίωση της απόδοσης του μοντέλου σε όλες τις μετρικές.

Επιπλέον, παραθέτουμε μία νέα μέθοδο εντοπισμού ανωμαλιών, η οποία λαμβάνει υπόψη τη χρονική διάσταση του προβλήματος με απλούστερο τρόπο από τις υπάρχουσες αρχιτεκτονικές επαναλαμβανόμενων δικτύων. Η αρχιτεκτονική αυτή δεν επιτυγχάνει ικανοποιητικά αποτελέσματα, όμως κρίνουμε απαραίτητο να την παρουσιάσουμε για λόγους πληρότητας.

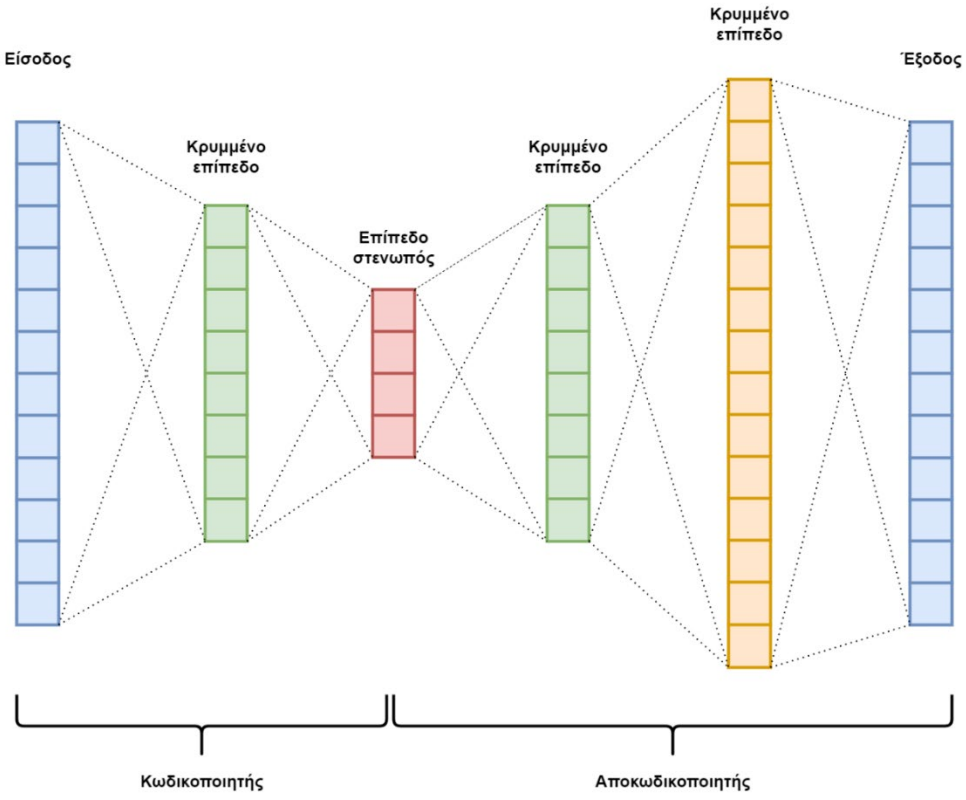
Στις επόμενες ενότητες αναλύουμε την αρχιτεκτονική δομή των μοντέλων που θα αξιοποιηθούν.

4.1. Απλός Αυτοκωδικοποιητής

Όπως προαναφέραμε, ο σκοπός αυτού του έργου είναι η ανάλυση της επίδρασης ενός βελτιωμένου μηχανισμού εκπαίδευσης ενός μοντέλου βαθιάς μάθησης που αξιοποιεί ανακατασκευή. Συνεπώς, για να αναδείξουμε την αξία αυτής της τροποποίησης, αρκεί να ξεκινήσουμε με ένα σχετικά απλό μοντέλο βαθιάς μάθησης, τον παραδοσιακό αυτοκωδικοποιητή.

Υλοποιούμε μία ασύμμετρη αρχιτεκτονική αυτοκωδικοποιητή, η οποία έχει εφαρμοστεί με επιτυχία στο [12] (ο αποκωδικοποιητής αντιστοιχεί στον γεννήτορα του GAN), με μικρές διαφοροποιήσεις. Επιλέγουμε να μην τροποποιήσουμε έντονα την αρχιτεκτονική αυτή, για δύο λόγους. Πρώτον, μας ενδιαφέρει να αναδείξουμε την ικανότητα της ημι-επιβλεπόμενης προσθήκης μας να βελτιώσει σημαντικά την απόδοση ακόμα και απλών αρχιτεκτονικών. Δεύτερον, η δομή αυτοκωδικοποιητή εμφανίζεται σε όλα τα μοντέλα που παρουσιάζουμε. Η διατήρηση μίας ενιαίας δομής μας επιτρέπει να συγκρίνουμε αποτελεσματικά και ουσιαστικά τα διαφορετικά μοντέλα μεταξύ τους.

Ο κωδικοποιητής περιλαμβάνει δύο γραμμικά επίπεδα για τη συμπίεση των δεδομένων, ενώ ο αποκωδικοποιητής περιλαμβάνει τρία γραμμικά επίπεδα. Σχηματικά έχουμε το παρακάτω:



Σχήμα 30: Η αρχιτεκτονική του αυτοκωδικοποιητή μας

Για την εκπαίδευση του αυτοκωδικοποιητή συγκρίνουμε δύο μεθόδους:

Πρώτη μέθοδος αποτελεί η παραδοσιακή μέθοδος ημιεπιβλεπόμενης μάθησης, κατά την οποία το μοντέλο εκπαιδεύεται μόνο στα ομαλά δεδομένα εκπαίδευσης, μαθαίνοντας την

κατανομή τους, και χρησιμοποιεί το σφάλμα ανακατασκευής ως *σκορ ανωμαλίας*. Πιο συγκεκριμένα, έστω E, D ο κωδικοποιητής και αποκωδικοποιητής αντίστοιχα, και $\{X^i\}_{i=1}^n$ τα ομαλά δείγματα εκπαίδευσης. Ο σκοπός εκπαίδευσης είναι η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος:

$$L = \frac{1}{m \cdot n} \sum_{i=1}^{n} \left\| X^i - D\left(E\left(X^i\right)\right) \right\|_2^2,$$

όπου m είναι το πλήθος χαρακτηριστικών της εισόδου. Να σημειωθεί ότι διαιρούμε με το m για λόγους κανονικοποίησης και ευκολότερης σύγκρισης της απόδοσης του μοντέλου στα διαφορετικά σύνολα δεδομένων, τα οποία έχουν διαφορετικό πλήθος χαρακτηριστικών. Θα τηρήσουμε τη σύμβαση σε όλη την έκταση της διπλωματικής εργασίας για λόγους συνέπειας.

Κατά την αξιολόγηση του μοντέλου, το σκορ ανωμαλίας ενός δείγματος $\boldsymbol{X} = [X_1, X_2, \dots, X_m]$ υπολογίζεται αντίστοιχα ως:

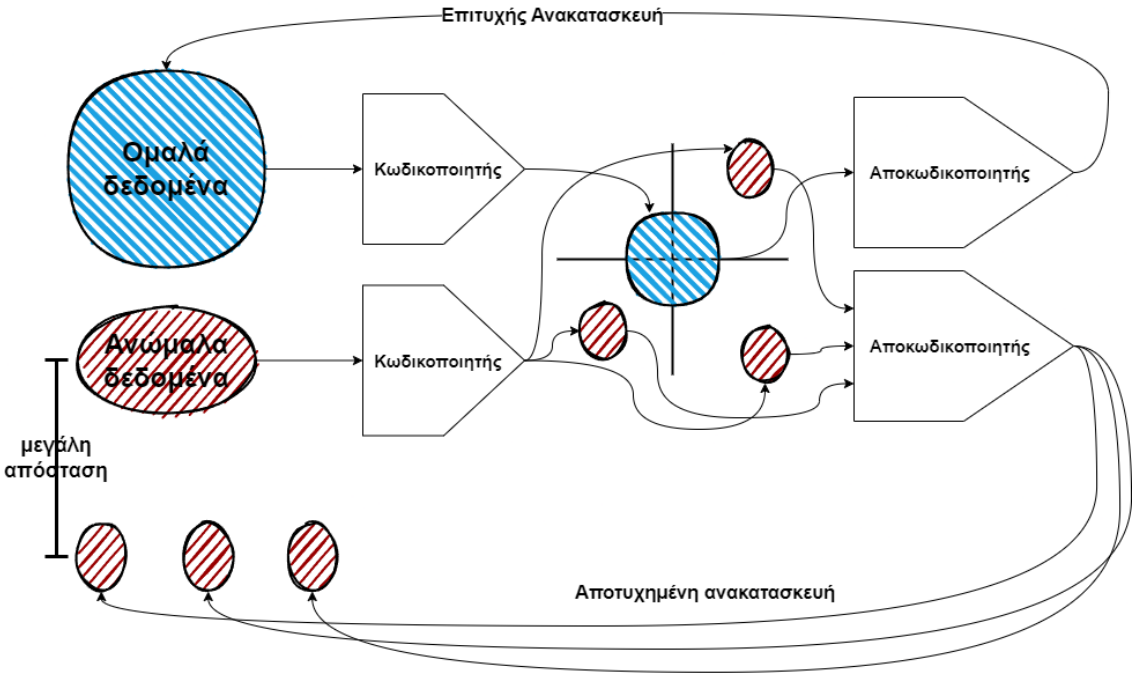
$$A(\boldsymbol{X}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(X_i - D\big(E(\boldsymbol{X})\big)_i \right)^2.$$

Δεύτερη μέθοδος είναι η προτεινόμενη από εμάς μέθοδος, κατά την οποία το μοντέλο εκπαιδεύεται αξιοποιώντας και τα δείγματα ανώμαλης δικτυακής κίνησης ως αντιπαραδείγματα. Αναλυτικότερα, ο απλός αυτοκωδικοποιητής μαθαίνει να ανακατασκευάζει σωστά τα ομαλά δεδομένα εκπαίδευσης, όμως είναι εύκολο να ανακατασκευάζει σωστά και ορισμένα ανώμαλα δείγματα δικτυακής κίνησης τα οποία φαινομενικά ομοιάζουν με ομαλά, αποτυγχάνοντας στον εντοπισμό τους. Για αυτόν τον λόγο, εισάγουμε στην απώλεια εκπαίδευσης του αυτοκωδικοποιητή έναν όρο μέσου αντίστροφου τετραγωνικού σφάλματος ανακατασκευής των ανώμαλων αντιπαραδειγμάτων, η ελαχιστοποίηση του οποίου οδηγεί σε κακή ανακατασκευή των δειγμάτων αυτών. Αναλυτικότερα, έστω ότι έχουμε επιπλέον k αντιπαραδείγματα που συμβολίζουμε με $\{\tilde{X}\}_{i=1}^k$, τότε η απώλεια εκπαίδευσης μετασχηματίζεται σε:

$$L = \frac{1}{m \cdot n} \sum_{i=1}^{n} \left\| X^i - D\left(E\left(X^i\right)\right) \right\|_2^2 + \theta \cdot \frac{1}{k} \sum_{i=1}^{k} \left(\frac{\left\| \tilde{X}^i - D\left(E\left(\tilde{X}^i\right)\right) \right\|_2^2}{m} \right)^{-1},$$

όπου θ μία υπερπαράμετρος εξισορρόπησης των δύο απωλειών.

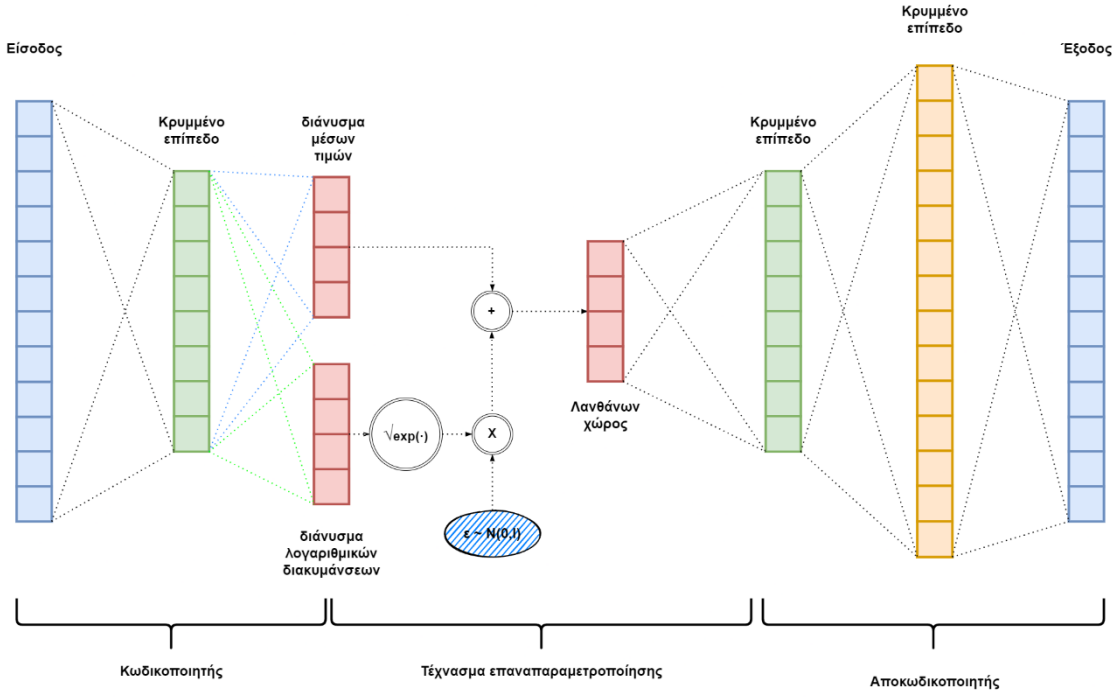
Η αξία της παραπάνω προσθήκης έγκειται στο ότι δεν περιορίζει την οπτική του μοντέλου στις συγκεκριμένες ανωμαλίες που παρουσιάζονται. Το μοντέλο δεν ενθαρρύνεται να μάθει τις ανωμαλίες ως κλάση, το οποίο υπονοεί ομοιότητα μεταξύ τους. Αντιθέτως, γίνεται πιο ευαίσθητο στις λεπτομέρειες που διαφοροποιούν τα ομαλά από τα μη ομαλά δείγματα, γενικεύοντας καλύτερα στο σύνολο των ανωμαλιών.



Σχήμα 31: Επιθυμητή συμπεριφορά του αυτοκωδικοποιητή μετά την προσθήκη μας

4.2. Παραλλακτικός Αυτοκωδικοποιητής

Κλιμακώνοντας ελαφρώς την πολυπλοκότητα των αρχιτεκτονικών, εξερευνούμε την επίδραση της ίδιας προσθήκης στον παραλλακτικό αυτοκωδικοποιητή. Για λόγους σύγκρισης, αφήνουμε την αρχιτεκτονική ίδια με τον προηγούμενο αυτοκωδικοποιητή, με μόνη διαφορά τον διπλασιασμό της διάστασης εξόδου του κωδικοποιητή, καθώς χρειαζόμαστε τη μέση τιμή και τυπική απόκλιση για κάθε χαρακτηριστικό.



Σχήμα 32: Η αρχιτεκτονική του παραλλακτικού αυτοκωδικοποιητή μας

Θα δοκιμάσουμε τον αυτοκωδικοποιητή μας με δύο τρόπους εκπαίδευσης πάνω σε αμιγώς ομαλά δεδομένα, και τους ίδιους δύο τρόπους πάνω σε μεικτά δεδομένα με αντιπαραδείγματα. Οι δύο τρόποι εκπαίδευσης υπό συζήτηση είναι οι εξής:

Πρώτον, η χρήση της απώλειας που παρουσιάσαμε στο κομμάτι της θεωρητικής ανάλυσης, η οποία αντιστοιχεί στον β-VAE, και παρατίθεται παρακάτω, με μικρές προσαρμογές που αντιστοιχούν στη δική μας υλοποίηση:

$$J = L + \lambda \cdot R,$$

$$L = \frac{\|\bar{X} - \bar{X}'\|^2}{m},$$

$$R = \frac{1}{2 \cdot m}\left(\|\bar{\mu}(\bar{X})\|^2 + \|\bar{\sigma}(\bar{X})\|^2 - 2\sum_{i=1}^{m} \ln(\bar{\sigma}(\bar{X})_i) - m\right),$$

όπου $\bar{X}' = \text{Decoder}\big(\text{Encoder}(\bar{X})\big)$.

Δεύτερον, η βελτιωμένη συνάρτηση απώλειας που αναφέραμε στη θεωρητική ενότητα, την οποία παραθέτουμε παρακάτω:

$$J = L + \lambda \cdot R,$$

$$L = \frac{\|\bar{X} - \bar{X}'\|^2}{m},$$

$$R = \left| \frac{1}{2 \cdot m} \left(\|\bar{\mu}(\bar{X})\|^2 + \|\bar{\sigma}(\bar{X})\|^2 - 2 \sum_{i=1}^{m} \ln(\bar{\sigma}(\bar{X})_i) - m \right) - C \right|.$$

Οι παραπάνω απώλειες αφορούν την περίπτωση εκπαίδευσης μόνο στα ομαλά δείγματα. Για την εκπαίδευση πάνω στα ανώμαλα δεδομένα αρκεί να τροποποιήσουμε την συνάρτηση απώλειας του παραλλακτικού αυτοκωδικοποιητή ως εξής:

$$J' = J + \theta \cdot \tilde{L},$$

όπου

$$\tilde{L} = \left(\frac{\left\| \tilde{X} - \tilde{X}' \right\|_2^2}{m} \right)^{-1},$$

με $\tilde{X}$ τα ανώμαλα δεδομένα εισόδου, $\tilde{X}' = \text{Αποκώδ.}\left(\text{Κωδ.}\left(\tilde{X}\right)\right)$ οι ανακατασκευές τους, και θ η υπερπαράμετρος βάρους των αντιπαραδειγμάτων.

Με τον παραπάνω τρόπο προστίθενται δύο ακόμα ρυθμίσεις του μοντέλου προς εξέταση.

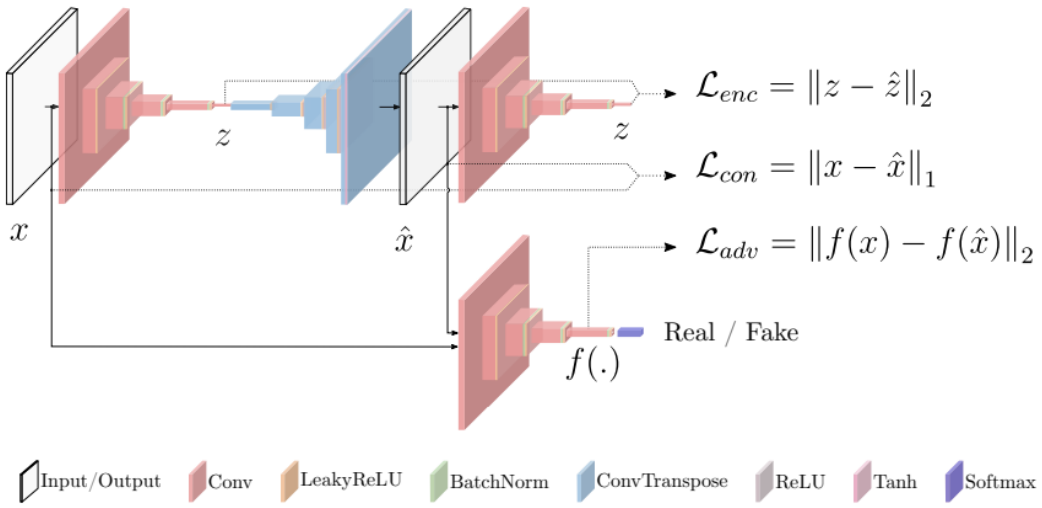
Το σκορ ανωμαλίας των δειγμάτων κατά την αξιολόγηση του μοντέλου υπολογίζεται με ανάλογο τρόπο με τον παραδοσιακό αυτοκωδικοποιητή. Αναλυτικότερα, για δείγμα $\boldsymbol{X} = [X_1, X_2, \dots, X_m]$ έχουμε:

$$A(\boldsymbol{X}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(X_i - D\left(E(\boldsymbol{X})\right)_i \right)^2,$$

όπου για την λήψη των δειγμάτων z του λανθάνοντα χώρου χρησιμοποιούμε επαναπαραμετροποίηση.

4.3. Προσαρμογή του μοντέλου GANomaly

Ως επόμενη αρχιτεκτονική θα αναλύσουμε το GANomaly, που παρουσιάζεται στο [11]. Όπως αναλύσαμε και στην ενότητα του προηγούμενου έργου, το συγκεκριμένο μοντέλο συνδυάζει την αντιπαραθετική μάθηση (adversarial learning) με τους αυτοκωδικοποιητές παράγοντας μία στιβαρή αρχιτεκτονική με υψηλή απόδοση στον εντοπισμό ανωμαλιών σε εικόνες. Η δομή αυτού του μοντέλου το καθιστά καλό υποψήφιο για την εφαρμογή την εφαρμογή της μεθόδου χρήσης αντιπαραδειγμάτων που προαναφέραμε. Για την αξιολόγηση της επίδρασης της τροποποίησής μας κρίνουμε απαραίτητο να υλοποιήσουμε πρώτα το ίδιο το GANomaly χωρίς αυτήν. Εντούτοις, το μοντέλο έχει σχεδιαστεί για την εφαρμογή σε εικόνες, οπότε είναι απαραίτητο να γίνουν αρκετές τροποποιήσεις. Η σημαντικότερη από αυτές είναι η αντικατάσταση των συνελικτικών μπλοκ από πλήρως συνδεδεμένα επίπεδα, καθώς η είσοδος είναι μονοδιάστατη και η εξαγωγή χαρακτηριστικών με απευθείας εφαρμογή της συνέλιξης δεν υποστηρίζεται από κάποιο διαισθητικό επιχείρημα. Οι υπόλοιπες τεχνικές λεπτομέρειες θα συζητηθούν αναλυτικά στο κεφάλαιο της υλοποίησης.



Σχήμα 33: Η αρχιτεκτονική του GANomaly [11]. Τα δείγματα εισόδου διέρχονται από τον αυτοκωδικοποιητή γεννήτορα και ανακατασκευάζονται. Ένας επιπρόσθετος κωδικοποιητής εφαρμόζεται στην έξοδο των συνθετικών δειγμάτων, καθώς οι συγγραφείς αναφέρουν πώς βοηθά στη σταθερότερη εκπαίδευση του μοντέλου. Τέλος, ένας διαχωριστής λαμβάνει κάθε ζεύγος δείγματος / ανακατασκευασμένου δείγματος και αποφαίνεται για τον χαρακτήρα του.

Η αντικειμενική συνάρτηση του γεννήτορα του GANomaly, την οποία κρατούμε ίδια στην τροποποιημένη υλοποίησή μας, είναι η ακόλουθη:

$$L = w_{adv}L_{adv} + w_{con}L_{con} + w_{enc}L_{enc},$$

όπου $w_{adv}, w_{con}, w_{enc}$ είναι υπερπαράμετροι που ρυθμίζουν την επίδραση κάθε όρου.

Πρόκειται για σύνθεση τριών διαφορετικών απωλειών. Η πρώτη απώλεια είναι η αντιθετική απώλεια (adversarial loss). Σε αντίθεση με το παραδοσιακό GAN, όπου ο γεννήτορας ανανεώνεται με βάση την έξοδο του διαχωριστή, χρησιμοποιείται η εσωτερική αναπαράσταση των δεδομένων εισόδου από τον διαχωριστή, η οποία έχει δειχθεί ότι οδηγεί σε καλύτερη απόδοση. Πιο συγκεκριμένα:

$$L_{adv} = \frac{1}{k}\mathbb{E}_{x\sim p_x}\left\|f(x) - f\big(G(x)\big)\right\|_2^2,$$

όπου f είναι η εσωτερική αναπαράσταση του διαχωριστή για το δείγμα x, k η διάστασή της, και G ο γεννήτορας (αυτοκωδικοποιητής). Ο γεννήτορας, λοιπόν, προσπαθεί να μειώσει την αναμενόμενη απόσταση των αναπαραστάσεων του διαχωριστή για τα δύο δείγματα, δυσκολεύοντάς τον να τα διακρίνει.

Ο δεύτερος όρος της συνάρτησης απώλειας είναι η απώλεια πλαισίου (contextual loss). Πιο συγκεκριμένα, ο γεννήτορας ενισχύεται με την απώλεια αυτοκωδικοποιητή ώστε να διασφαλίζεται η ομοιότητα των δειγμάτων εξόδου με αυτά της εισόδου. Με μαθηματικούς όρους:

$$L_{con} = \frac{1}{m}\mathbb{E}_{x\sim p_x}\|x - G(x)\|_1,$$

όπου m *όπως πριν είναι το πλήθος χαρακτηριστικών της εισόδου.*

Τέλος, ο γεννήτορας ενθαρρύνεται να ελαχιστοποιήσει την απώλεια κωδικοποίησης (encoder loss). Πρόκειται για τη μέση τετραγωνική απόσταση των δειγμάτων του λανθάνοντα χώρου με τα ανακατασκευασμένα ανάλογά τους. Εξυπηρετεί τη σταθερή εκπαίδευση του μοντέλου. Έχουμε λοιπόν:

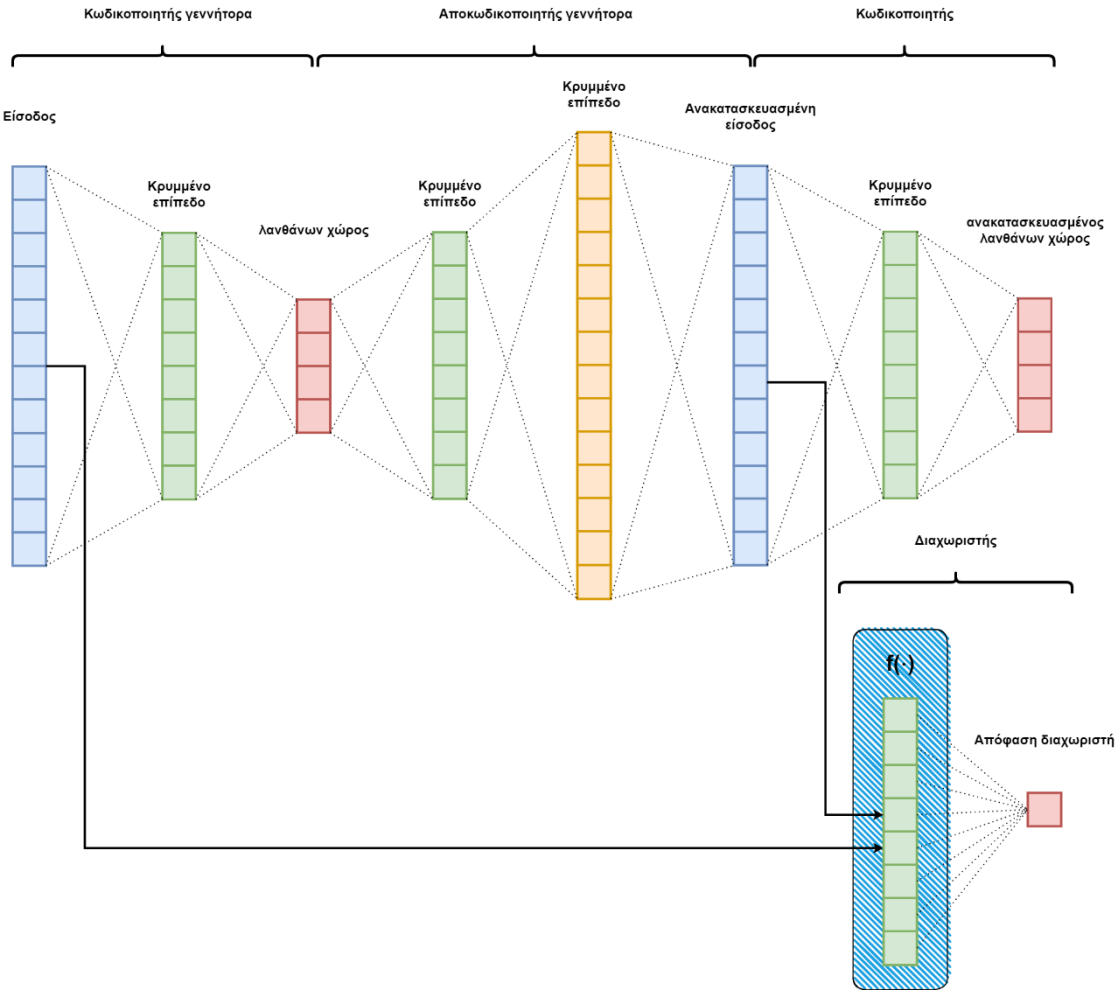
$$L_{enc} = \frac{1}{\lambda}\mathbb{E}_{x\sim p_x}\left\|G_E(x) - E\big(G(x)\big)\right\|_2^2,$$

όπου λ η διάσταση του λανθάνοντα χώρου.

Θα αναλύσουμε τώρα την τροποποίηση που προτείνουμε για την αύξηση της απόδοσης του GANomaly μέσω χρήσης αντιπαραδειγμάτων. Υπάρχουν πολλές επιλογές, όπως η τροποποίηση της απώλειας κωδικοποίησης και η ενίσχυση του διαχωριστή, εντούτοις τροποποιούμε συνειδητά μόνο την απώλεια L_{con}, καθώς ο αυτοκωδικοποιητής αποτελεί τον πυρήνα του δικτύου. Η τροποποίηση είναι ίδια με την περίπτωση του αυτοκωδικοποιητή. Η απώλεια γίνεται λοιπόν:

$$\mathrm{L} = w_{adv} L_{adv} + w_{con} \hat{L}_{con} + w_{enc} L_{enc},$$

$$\hat{L}_{con} = \mathbb{E}_{x \sim p_x} \frac{\|x - G(x)\|_1}{m} + \theta \cdot \mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}} \left(\frac{\|\tilde{x} - G(\tilde{x})\|_1}{m} \right)^{-1}.$$



Σχήμα 34: Αρχιτεκτονική του GANomaly_variant, τροποποίησης του GANomaly. Τα συνελικτικά επίπεδα έχουν αντικατασταθεί με πλήρως συνδεδεμένα. Χρησιμοποιούμε την ίδια ασύμμετρη αρχιτεκτονική του αυτοκωδικοποιητή αφενός καθώς παρουσιάζει ικανοποιητική απόδοση, και αφετέρου για λόγους σύγκρισης με τις άλλες αρχιτεκτονικές που εξετάζουμε.

Για την αξιολόγηση του GANomaly οι ερευνητές χρησιμοποιούν το σφάλμα κωδικοποίησης L_{enc}. Κρίνουμε πως αυτή η επιλογή είναι επαρκής για τις ανάγκες του προβλήματός μας και συμβατή με τις προηγούμενές μας επιλογές, καθώς η κακή ανακατασκευή των ανώμαλων δειγμάτων από τον αυτοκωδικοποιητή θα οδηγήσει σε αντίστοιχη αδυναμία ανακατασκευής των αναπαραστάσεών τους στον λανθάνοντα χώρο. Το σκορ ανωμαλίας, λοιπόν, ενός δείγματος $\hat{x}$ είναι:

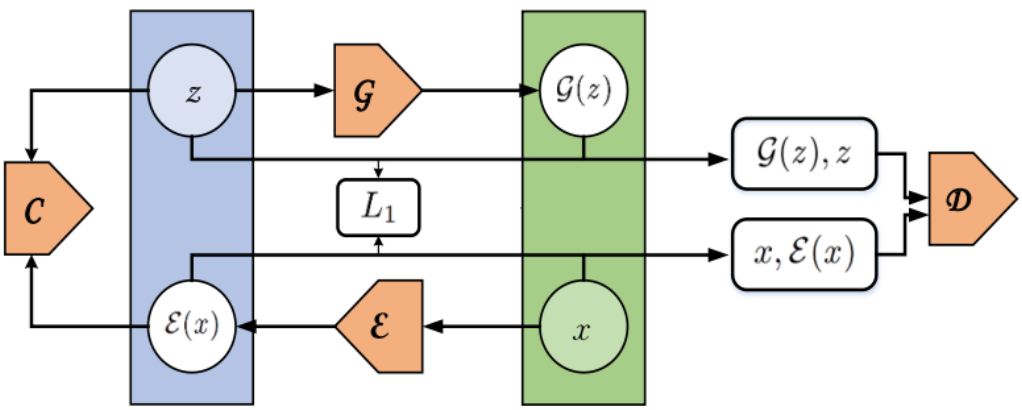
$$A(\hat{x}) = \frac{1}{\lambda} \left\| G_E(\hat{x}) - E\big(G(\hat{x})\big) \right\|_1.$$

4.4. Αμφίδρομο WGAN-GP (BiWGAN-GP)

Εξετάζουμε τώρα μία αρχιτεκτονική παραγωγικού αντιπαραθετικού δικτύου με πιο συμβατική μορφή. Το συγκεκριμένο μοντέλο προτάθηκε από τους Wei Yao κ.ά. [12] και θα αναφερόμαστε σε αυτό ως *BiWGAN-GP*, αφού συνδυάζει τις καινοτόμες προσθήκες των WGAN-GP και BiGAN. Ακολουθεί η περιγραφή του τρόπου λειτουργίας του αρχικού BiWGAN-GP που προτάθηκε από τους συγγραφείς.

Σημείωση: Στη δημοσίευση του BiWGAN-GP διαδραματίζει έντονο ρόλο η σχέση της αρχιτεκτονικής με την υποδομή νέφους και η ανάπτυξή της στα πλαίσια αυτά. Δεν αναλύουμε τα σχετικά με το νέφος κομμάτια, καθώς δεν σχετίζονται με αυτήν την εργασία. Υλοποιούμε, συνεπώς, αυτήν την αρχιτεκτονική αμιγώς τοπικά, αγνοώντας τη διάσταση της υποδομής νέφους.

Το μοντέλο BiWGAN-GP αποτελείται από τέσσερις μονάδες: Τον κωδικοποιητή, τον γεννήτορα, τον διαχωριστή και τον ταξινομητή. Ο γεννήτορας, κωδικοποιητής και διαχωριστής λειτουργούν όπως στο παραδοσιακό BiGAN. Αναλυτικότερα, ο κωδικοποιητής λαμβάνει στην είσοδό του δείγματα από τα σύνολα δεδομένων και παράγει την αναπαράστασή τους στον λανθάνοντα χώρο, ενώ ο γεννήτορας υλοποιεί τον αντίστροφο μετασχηματισμό. Ο διαχωριστής δέχεται τη συνένωση δεδομένων με τη λανθάνουσα αναπαράστασή τους και αποφαίνεται για το αν αυτά αντιστοιχούν στον γεννήτορα ή στον κωδικοποιητή. Ο ταξινομητής είναι μία καινούρια προσθήκη, η οποία λειτουργεί ως δεύτερος ελάσσων διαχωριστής, μόνο για τα δείγματα του λανθάνοντα χώρου. Είναι ένα μέτρο για τον ισχυρότερο περιορισμό του γεννήτορα και κωδικοποιητή, ώστε αυτοί να ενθαρρυνθούν να παράγουν πιο ποιοτικές εξόδους.



Σχήμα 35: Η αρχιτεκτονική του BiWGAN-GP [12].

Η συνάρτηση απώλειας είναι σύνθετη, με κάθε συνιστώσα να εξυπηρετεί διαφορετικό ρόλο. Αναλύουμε πρώτα την κάθε συνιστώσα χωριστά.

Η αντιπαραθετική απώλεια αντιστοιχεί στην απώλεια του min-max αντιπαλικού παιχνιδιού μεταξύ των γεννήτορα/κωδικοποιητή και του διαχωριστή. Η μαθηματική της έκφραση είναι η εξής:

$$\min_{G,E} \max_{D \in W_D} V_{adv}(G,E,D) = \mathbb{E}_{x \sim p_x}\left[D\big(x, E(x)\big)\right] - E_{z \sim p_z}[D(G(z), z)] + \mu \cdot \mathrm{GP}(\tilde{x}, \tilde{z}),$$

όπου W_D είναι το σύνολο των 1-Lipschitz συναρτήσεων. Ισχύει επιπλέον ότι:

$$\mathrm{GP}(\cdot) = \mathbb{E}[(\|\nabla D(\cdot)\|_2 - 1)^2]$$

ο όρος ποινής κλίσεων που προτάθηκε από τους Gulrajani κ.ά. [55], μ είναι το βάρος της ποινής κλίσεων, και το $(\tilde{x}, \tilde{z})$ αντιστοιχεί στην ομοιόμορφη δειγματοληψία κατά μήκος ευθύγραμμων τμημάτων που ενώνουν ζεύγη σημείων της κατανομής των δειγμάτων του γεννήτορα και δειγμάτων του κωδικοποιητή. Παρατηρούμε ότι σε αντίθεση με το παραδοσιακό BiGAN, οι όροι της μέσης τιμής δεν περιλαμβάνουν λογάριθμους. Αυτό οφείλεται στη χρήση απώλειας Wasserstein αντί για Jensen-Shannon.

Η απώλεια κωδικοποίησης είναι πλήρως αντίστοιχη με την αντιπαραθετική απώλεια, με μόνη διαφορά ότι αφορά τον ταξινομητή και τα δείγματα του λανθάνοντα χώρου. Ο σκοπός της απώλειας αυτής είναι η συμπερίληψη του ταξινομητή στο min-max παίγνιο με τον κωδικοποιητή και η σταθεροποίηση της εκπαίδευσης. Η μαθηματική της έκφραση είναι:

$$\min_{E} \max_{C \in W_C} V_{cod}(C,E) = \mathbb{E}_{x \sim p_x}\big[C\big(E(x)\big)\big] - \mathbb{E}_{z \sim p_z}[C(z)] + \mu \cdot \mathrm{GP}(\tilde{z}),$$

όπου W_C είναι το σύνολο των 1-Lipschitz συναρτήσεων.

Τέλος, οι συγγραφείς εισάγουν την απώλεια κυκλικής συνέπειας, έναν επιπλέον μηχανισμό περιορισμού του ζεύγους γεννήτορα και κωδικοποιητή, για την εξασφάλιση της αντίστροφης σχέσης μεταξύ τους. Το παραδοσιακό BiGAN [54] δεν επιβάλλει τέτοιον περιορισμό, καθώς θεωρητικά αυτός δεν είναι απαραίτητος για την επιτυχή αντιστροφή του γεννήτορα από τον κωδικοποιητή. Εντούτοις, οι συγγραφείς του BiWGAN παρατηρούν ότι στην πράξη η θεωρητικά αναμενόμενη σύγκλιση συχνά δεν επιτυγχάνεται, γεγονός που τους οδηγεί στην υιοθέτηση αυτής της λύσης. Με μαθηματικούς όρους:

$$V_{cyc}(G,E) = \frac{1}{m}\mathbb{E}_{x \sim p_x}\left[\left\|x - G\big(E(x)\big)\right\|_1\right],$$

όπου m το πλήθος χαρακτηριστικών της εισόδου όπως και πριν.

Ο συνολικός στόχος εκπαίδευσης είναι ο ακόλουθος:

$$\min_{G,E} \max_{D \in W_D, C \in W_C} V(G,E,D,C) = V_{adv}(G,E,D) + V_{cod}(C,E) + \sigma V_{cyc}(G,E),$$

όπου η υπερπαράμετρος σ ελέγχει τη σημασία της απώλειας συνέπειας.

Όταν το μοντέλο έχει εκπαιδευτεί, το σκορ ανωμαλίας υπολογίζεται συνδυαστικά από τον γεννήτορα, κωδικοποιητή και διαχωριστή. Πιο συγκεκριμένα, έστω x ένα δείγμα εισόδου. Το σκορ ανωμαλίας του δίνεται από τον τύπο:

$$A(x) = \mathcal{L}_{f_D}(x) = \frac{1}{k} \left\| f_D\big(x, E(x)\big) - f_D\left(G\big(E(x)\big), E(x)\right) \right\|_1,$$

όπου $f_D(\cdot,\cdot)$ είναι οι ενεργοποιήσεις του επιπέδου αμέσως πριν την απόφαση του διαχωριστή, και k η διάστασή τους, όπως και στο GANomaly που περιγράψαμε προηγουμένως. Διαισθητικά, για ένα ομαλό δείγμα δικτυακής κίνησης, η ανακατασκευή του ζεύγους γεννήτορα και κωδικοποιητή θα είναι ικανοποιητική, και ο διαχωριστής δεν θα μπορεί να διακρίνει τα δύο δείγματα, δηλαδή οι ενεργοποιήσεις f_D θα μοιάζουν, με αποτέλεσμα το σκορ ανωμαλίας να λαμβάνει χαμηλές τιμές. Αντίθετα, για ανώμαλα δείγματα, ο γεννήτορας και κωδικοποιητής δεν θα πετυχαίνουν ακριβή ανακατασκευή, και ο διαχωριστής θα μεγεθύνει τις διαφορές τους, αυξάνοντας το σκορ ανωμαλίας.

Παραπάνω περιγράφεται η αρχιτεκτονική BiWGAN-GP που προτάθηκε από τους Wei Yao κ.ά. [12]. Τώρα θα περιγράψουμε τη δική μας τροποποίηση για την ενσωμάτωση αντιπαραδειγμάτων στην εκπαιδευτική διαδικασία.

Όπως προηγουμένως, υπολογίζουμε τον αντίστροφο του σφάλματος ανακατασκευής των ανώμαλων δειγμάτων:

$$\widehat{V_{cyc}}(G, E) = \mathbb{E}_{x \sim p_x}\left[\frac{\left\|x - G\big(E(x)\big)\right\|_1}{m}\right] + \theta \cdot \mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}}\left[\left(\frac{\left\|\tilde{x} - G\big(E(\tilde{x})\big)\right\|_1}{m}\right)^{-1}\right],$$

όπου θ ο συντελεστής βάρους για τα αντιπαραδείγματα.

Όπως αναλύσαμε και προηγουμένως, αυτή η προσθήκη διευκολύνει τη διάκριση των ανωμαλιών κατά την αξιολόγηση, αφού η ανακατασκευή τους είναι χειρότερη, γεγονός που εντοπίζεται από τον διαχωριστή.

4.5. Συνελικτικός Αυτοκωδικοποιητής (ConvAE)

Βάσει της βιβλιογραφικής μας έρευνας, η πλειοψηφία των μοντέλων που χρησιμοποιούν συνελικτικά επίπεδα λειτουργούν με συνέλιξη μίας διάστασης ευθέως πάνω στα δεδομένα εισόδου [25, 58]. Επιπλέον, στο [33] χρησιμοποιείται μία μέθοδος δισδιάστατης συνέλιξης η οποία, παρά την επιτυχή της εφαρμογή, δεν λαμβάνει ιδιαίτερα υπόψη τη χρονική διάσταση της εισόδου. Τα μοντέλα που έχουν χρησιμοποιηθεί για τη λήψη υπόψιν της χρονικής συνιστώσας της εισόδου χρησιμοποιούν κατά κύριο λόγο πολύπλοκες επαναλαμβανόμενες αρχιτεκτονικές (RNN) [7, 24, 26], οι οποίες ενδεχομένως να έχουν αρκετά μεγαλύτερη χωρητικότητα από την απαιτούμενη, με αποτέλεσμα την υπερπροσαρμογή (overfitting).

Οι λόγοι αυτοί συνετέλεσαν στην παραγωγή του μοντέλου που θα παρουσιάσουμε σε αυτήν την ενότητα. Προτείνουμε μία αρχιτεκτονική συνελικτικού αυτοκωδικοποιητή δύο διαστάσεων, ο οποίος λειτουργεί πάνω σε παράθυρα συσχέτισης των δειγμάτων, ώστε να εντοπίσει ανωμαλίες από μεταβολές στην συνολική εικόνα της δικτυακής κίνησης. Παρακάτω αναλύουμε την πλήρη δομή του μοντέλου μας.

Το πρώτο στάδιο επεξεργασίας των δειγμάτων εισόδου είναι η συμπίεση των χαρακτηριστικών τους με χρήση της μεθόδου ανάλυσης κύριων συνιστωσών (PCA). Με αυτόν τον τρόπο παράγουμε πιο πυκνές αναπαραστάσεις, αμβλύνοντας το πρόβλημα ότι πολλά χαρακτηριστικά, ιδιαίτερα αυτά που έχουν παραχθεί με κωδικοποίηση One-Hot κατηγορικών χαρακτηριστικών, δεν παρέχουν ιδιαίτερη πληροφορία. Ένα επιπρόσθετο πλεονέκτημα είναι ότι το πλήθος των νέων χαρακτηριστικών μπορεί να γίνει δύναμη του 2, το οποίο βοηθά στην ομαλή εφαρμογή ενός συμμετρικού συνελικτικού αυτοκωδικοποιητή, όπως θα δούμε και στο κεφάλαιο των λεπτομερειών υλοποίησης. Τέλος, η χρήση PCA παράγει διανύσματα τα οποία είναι συνολικά ορθογώνια μεταξύ τους. Αυτό είναι χρήσιμο, καθώς η έλευση ανωμαλιών αναμένεται να αλλοιώνει την έλλειψη συσχέτισης μεταξύ των νέων χαρακτηριστικών στο αντίστοιχο χρονικό παράθυρο, γεγονός που θα ανιχνευτεί από τον αυτοκωδικοποιητή.

Στο δεύτερο στάδιο επεξεργασίας των δεδομένων, υπολογίζουμε τον πίνακα συσχέτισης των δειγμάτων του παραθύρου. Χρησιμοποιούμε συσχέτιση Pearson, η οποία δίνεται από τον ακόλουθο τύπο:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

όπου $x_i, y_i, \bar{x}, \bar{y}$ οι τιμές της πρώτης και δεύτερης κατανομής και οι μέσες τιμές τους αντίστοιχα.

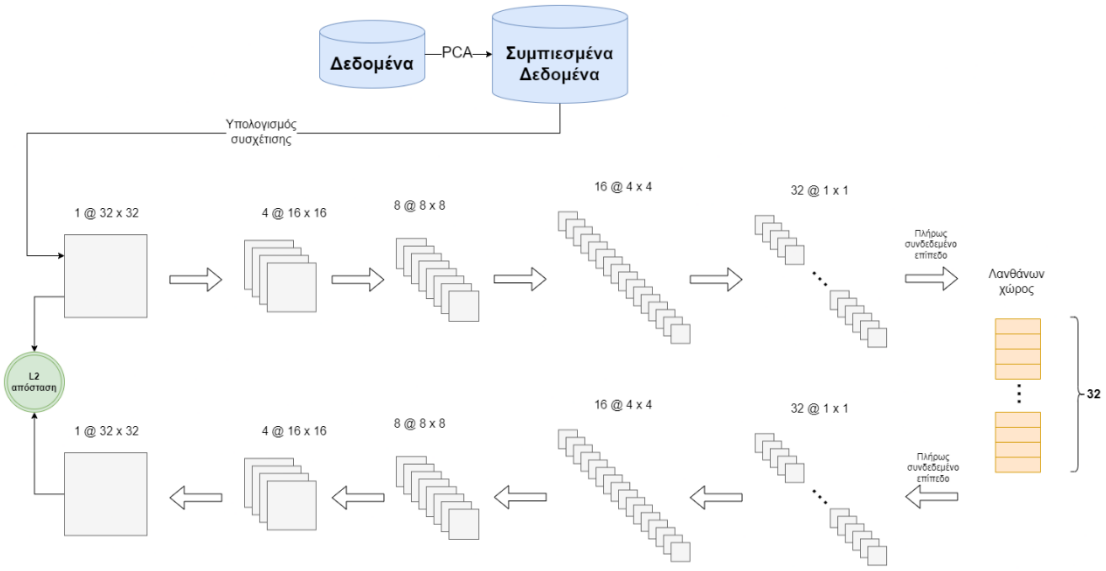
Στο τρίτο και τελευταίο στάδιο, ένας συνελικτικός αυτοκωδικοποιητής δύο διαστάσεων δέχεται ως είσοδο τον πίνακα συσχέτισης και προσπαθεί να τον αναπαραγάγει στην έξοδο. Ο CAE αποτελείται από τέσσερα συνελικτικά επίπεδα και ένα πλήρως συνδεδεμένο επίπεδο για την κωδικοποίηση, και από τα ίδια επίπεδα με αντίθετη σειρά στην αποκωδικοποίηση. Σημειώνουμε ότι κάθε δείγμα του συνόλου δεδομένων εκπαίδευσης και ελέγχου

αντιστοιχίζεται με τον πίνακα συσχέτισης του χρονικού παραθύρου των δειγμάτων που τελειώνει με το δείγμα υπό εξέταση.

Το μοντέλο προσπαθεί να ελαχιστοποιήσει την ακόλουθη αντικειμενική συνάρτηση:

$$V_{obj} = \mathbb{E}_{x \sim p_x}\left[\frac{\left\|X - D\left(E(X)\right)\right\|_2^2}{m^2}\right] + \theta \cdot \mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}}\left[\left(\frac{\left\|\tilde{X} - D\left(E(\tilde{X})\right)\right\|_2^2}{m^2}\right)^{-1}\right],$$

όπου X είναι ο $m \times m$ πίνακας συσχέτισης της εισόδου για τα ομαλά δείγματα, και $\tilde{X}$ ο πίνακας συσχέτισης της εισόδου για τα μη ομαλά δείγματα.



Σχήμα 36: Το pipeline του συνελικτικού αυτοκωδικοποιητή που προτείνουμε.

Το σκορ ανωμαλίας υπολογίζεται αντίστοιχα με τον απλό αποκωδικοποιητή, ως η μέση τετραγωνική απόσταση των χαρακτηριστικών της εισόδου με την ανακατασκευασμένη είσοδο. Με μαθηματικούς όρους, αν X είναι το ο πίνακας συσχέτισης που αντιστοιχεί στην είσοδο, τότε το σκορ ανωμαλίας δίνεται από τον τύπο:

$$A(X) = \frac{\left\|X - D\left(E(X)\right)\right\|_2^2}{m^2}.$$

5. Υλοποίηση Μοντέλων

5.1. Περιγραφή των συνόλων δεδομένων εκπαίδευσης

5.1.1. CIC UNSW-NB15

Το UNSW-NB15 που προτάθηκε από τους Nour Moustafa κ.ά. [14, 15, 16, 17, 18] το 2015 για την αντιμετώπιση του προβλήματος έλλειψης σύγχρονων και αντιπροσωπευτικών συνόλων δεδομένων στον τομέα της ανίχνευσης εισβολών σε δίκτυα. Αναλυτικότερα τα σημαντικότερα σύνολα δεδομένων πριν την έλευση του UNSW-NB15 ήταν τα KDDCUP99 [59] και NSL-KDD [60]. Το πρώτο πλήττεται από το πρόβλημα της ύπαρξης πολλαπλών διπλότυπων εγγραφών, που οδηγεί σε έντονη μεροληψία προς τις συχνές εγγραφές, ενώ το δεύτερο, παρά το γεγονός ότι δεν παρουσιάζει τα προηγούμενα προβλήματα, αδυνατεί να παράσχει μία ακριβή απεικόνιση των σύγχρονων επιθέσεων, οι οποίες αφήνουν αρκετά πιο ανεπαίσθητα ίχνη. Το UNSW-NB15 προέκυψε ως αποτέλεσμα των παραπάνω ελλείψεων.

Για την παραγωγή του, οι ερευνητές προσομοιώνουν ένα δίκτυο υπολογιστών χρησιμοποιώντας έναν IXIA παραγωγό δικτυακής κίνησης, διαμορφωμένο με τρεις εξυπηρετητές. Οι δύο εξυπηρετητές παράγουν ομαλή δικτυακή κίνηση, ενώ ο τρίτος παράγει κακόβουλη κίνηση, χρησιμοποιώντας πληροφορίες από τον ιστότοπο για τις συχνές τρωτότητες και εκθέσεις (Common Vulnerabilities and Exposures / CVE). Οι μετρήσεις λαμβάνονται σε δύο διαφορετικές μέρες, με 16 ώρες μετρήσεων την πρώτη μέρα και 15 ώρες τη δεύτερη. Στην πρώτη φάση προσομοίωσης παράγεται μία επίθεση ανά δευτερόλεπτο, ενώ στη δεύτερη φάση παράγονται δέκα επιθέσεις το δευτερόλεπτο. Συνολικά, 100GB καταγραφών προκύπτουν από τις προσομοιώσεις.

Μετά την παραγωγή των καταγραφών, τα εργαλεία Argus και Bro-IDS χρησιμοποιούνται για την εξαγωγή 49 χαρακτηριστικών (με 47 χαρακτηριστικά πληροφορίας και 2 χαρακτηριστικά-ετικέτες).

Να σημειωθεί ότι οι δοκιμές μας εκτελούνται σε ένα αντιπροσωπευτικό υποσύνολο του UNSW-NB15 που έχει παραχθεί από τους ίδιους τους ερευνητές για την εκπαίδευση μοντέλων μηχανικής μάθησης. Το σύνολο δεδομένων εκπαίδευσης περιέχει 175,341 εγγραφές και το σύνολο δεδομένων ελέγχου περιέχει 82,332 εγγραφές.

Στη συνέχεια παραθέτουμε αναλυτικά κάθε ένα από τα χαρακτηριστικά του συνόλου δεδομένων, οργανωμένα ανά κατηγορία:

Πίνακας 1: Χαρακτηριστικά Ροής του UNSW-NB15

#	Όνομα	Τύπος	Περιγραφή
1	*srcip*	N	IP διεύθυνση πηγής
2	*Sport*	I	Αριθμός θύρας πηγής
3	*dstip*	N	IP διεύθυνση προορισμού
4	*dsport*	I	Αριθμός θύρας προορισμού
5	*proto*	N	Πρωτόκολλο συναλλαγής

Πίνακας 2: Βασικά Χαρακτηριστικά του UNSW-NB15

#	Όνομα	Τύπος	Περιγραφή
6	*state*	N	Η κατάσταση και το εξαρτημένο από αυτήν πρωτόκολλο, π.χ. ACC, CLO, άλλο (-)
7	*dur*	F	Συνολική διάρκεια εγγραφής
8	*sbytes*	I	Bytes από πηγή προς προορισμό
9	*dbytes*	I	Bytes από προορισμό προς πηγή
10	*sttl*	I	Time to live από πηγή προς προορισμό
11	*dttl*	I	Time to live από προορισμό προς πηγή
12	*sloss*	I	Πακέτα πηγής που επαναμεταδόθηκαν ή απορρίφθηκαν
13	*dloss*	I	Πακέτα προορισμού που επαναμεταδόθηκαν ή απορρίφθηκαν
14	*service*	N	http, ftp, ssh, dns, …, άλλο (-)
15	*sload*	F	Bit το δευτερόλεπτο πηγής
16	*dload*	F	Bit το δευτερόλεπτο προορισμού
17	*spkts*	I	Πλήθος πακέτων πηγής -> προορισμού
18	*dpkts*	I	Πλήθος πακέτων προορισμού -> πηγής

Πίνακας 3: Χαρακτηριστικά Περιεχομένου του UNSW-NB15

#	Όνομα	Τύπος	Περιγραφή
19	*swin*	I	Διαφήμιση TCP παραθύρου πηγής
20	*dwin*	I	Διαφήμιση TCP παραθύρου προορισμού
21	*stcpb*	I	Αριθμός ακολουθίας (SN) TCP πηγής
22	*dtcpb*	I	Αριθμός ακολουθίας (SN) TCP προορισμού
23	*smeansz*	I	Μέσο μέγεθος πακέτου ροής που στέλνει η πηγή
24	*dmeansz*	I	Μέσο μέγεθος πακέτου ροής που στέλνει ο προορισμός
25	*trans_depth*	I	Το βάθος μέσα στη σύνδεση της συναλλαγής http αιτήματος / απάντησης
26	*res_bdy_len*	I	Το μέγεθος περιεχομένου των δεδομένων που μεταφέρονται από την http υπηρεσία του εξυπηρετητή

Πίνακας 4: Χρονικά Χαρακτηριστικά του UNSW-NB15

#	Όνομα	Τύπος	Περιγραφή
27	*sjit*	F	Jitter πηγής (mSec)
28	*djit*	F	Jitter προορισμού (mSec)
29	*stime*	T	Χρόνος έναρξης εγγραφής
30	*ltime*	T	Χρόνος τέλους εγγραφής
31	*sintpkt*	F	Χρόνος άφιξης μεταξύ πακέτων πηγής (mSec)
32	*dintpkt*	F	Χρόνος άφιξης μεταξύ πακέτων προορισμού (mSec)
33	*tcprtt*	F	Το άθροισμα των 'synack' και 'ackdat' του TCP
34	*synack*	F	Ο χρόνος μεταξύ των SYN και SYN_ACK πακέτων του TCP
35	*ackdat*	F	Ο χρόνος μεταξύ των SYN_ACK και των ACK πακέτων του TCP

Πίνακας 5: Επιπρόσθετα Χαρακτηριστικά του UNSW-NB15

#	Όνομα	Τύπος	Περιγραφή
Χαρακτηριστικά γενικού σκοπού			
36	*is_sm_ips_ports*	B	Αν οι διευθύνσεις IP και θύρες πηγής και προορισμού είναι ίσες τότε λαμβάνει 1, αλλιώς 0
37	*ct_state_ttl*	I	Αρ. για κάθε κατάσταση σύμφωνα με συγκεκριμένο εύρος τιμών για χρόνο ζωής πηγής/προορισμού
38	*ct_flw_http_mthd*	I	Αριθμός ροών που έχουν μεθόδους όπως GET και POST στην υπηρεσία HTTP
39	*is_ftp_login*	B	Αν αποκτήθηκε πρόσβαση στη σύνοδο ftp με όνομα χρήστη και κωδικό τότε 1, αλλιώς 0
40	*ct_ftp_cmd*	I	Αριθμός ροών που έχει εντολή στη σύνοδο ftp
Χαρακτηριστικά σύνδεσης			
41	*ct_srv_src*	I	Αρ. συνδέσεων που περιέχουν την ίδια υπηρεσία και διεύθυνση πηγής σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο.
42	*ct_srv_dst*	I	Αρ. συνδέσεων που περιέχουν την ίδια υπηρεσία και διεύθυνση προορισμού σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο.
43	*ct_dst_ltm*	I	Αρ. συνδέσεων με ίδια διεύθυνση προορισμού σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο

44	*ct_src_ ltm*	I	Αρ. συνδέσεων με ίδια διεύθυνση πηγής σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο
45	*ct_src_dport_ltm*	I	Αρ. συνδέσεων με ίδια διεύθυνση πηγής και θύρα προορισμού σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο
46	*ct_dst_sport_ltm*	I	Αρ. συνδέσεων με ίδια διεύθυνση προορισμού και θύρα πηγής σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο
47	*ct_dst_src_ltm*	I	Αρ. συνδέσεων με ίδια διεύθυνση πηγής και προορισμού σε 100 συνδέσεις σύμφωνα με τον τελευταίο χρόνο

Πίνακας 6: Χαρακτηριστικά με ετικέτα του UNSW-NB15

#	Όνομα	Τύπος	Περιγραφή
48	*attack_cat*	N	Το όνομα κάθε κατηγορίας επίθεσης. Υπάρχουν 9 κατηγορίες: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms
49	*Label*	B	0 για ομαλή κίνηση και 1 για επίθεση
Τύπος: N: κατηγορικό, I: ακέραιος, F: κινητής υποδιαστολής, T: χρονοσφραγίδα και B: δυαδικό			

Ακολουθεί πίνακας με την κατανομή των εγγραφών στις διαφορετικές κλάσεις.

Πίνακας 7: Κατανομή εγγραφών σε κλάσεις για το UNSW-NB15

Τύπος	Αρ. εγγραφών		Περιγραφή
	Σύνολο Εκπαίδ.	Σύνολο Αξιολόγ.	
Normal	56.000	37.000	Φυσιολογικά δεδομένα συναλλαγών
Fuzzers	18.184	6.062	Απόπειρα πλήξης ενός προγράμματος ή δικτύου στέλνοντας τυχαία δεδομένα
Analysis	2.000	677	Περιλαμβάνει διαφορετικές επιθέσεις: σάρωση θυρών, ανεπιθύμητη κίνηση, και διείσδυση μέσω αρχείων html
Backdoor	1.746	583	Μία τεχνική κατά την οποία ένας μηχανισμός ασφάλειας συστήματος παρακάμπτεται κρυφά για να αποκτηθεί πρόσβαση σε έναν υπολογιστή ή τα δεδομένα του
DoS	12.264	4.089	Μία κακόβουλη απόπειρα για να καταστεί ένας εξυπηρετητής ή δικτυακός πόρος μη διαθέσιμος στους χρήστες. Συνήθως γίνεται προσωρινά διακόπτοντας τις υπηρεσίες ενός host που συνδέεται στο διαδίκτυο
Exploits	33.393	11.132	Ο επιτιθέμενος γνωρίζει ένα κενό ασφαλείας στο λειτουργικό σύστημα ή λογισμικό ενός μηχανήματος και αξιοποιεί αυτή τη γνώση για να εκτελέσει ανεπιθύμητες ενέργειες
Generic	40.000	18.871	Η τεχνική αυτή δουλεύει ενάντια σε όλα τα κρυπτοσυστήματα block, χωρίς να λαμβάνεται υπόψη η δομή του κρυπτοσυστήματος
Reconnaissance	10.491	3.496	Περιέχει όλες τις περιπτώσεις που μπορούν να προσομοιώσουν επιθέσεις που συλλέγουν πληροφορίες
Shellcode	1.133	378	Ένα μικρό κομμάτι κώδικα που χρησιμοποιείται ως φορτίο κατά την εκμετάλλευση ενός κενού ασφαλείας στο λογισμικό
Worms	130	44	Ο επιτιθέμενος αντιγράφει τον εαυτό του σε άλλους υπολογιστές. Συχνά χρησιμοποιεί μία δικτυακή υποδομή για να μεταδοθεί, βασιζόμενος σε κενά ασφαλείας στους υπολογιστές-στόχους για την απόκτηση πρόσβασης
Σύνολο	175.341	82.332	

5.1.2. CSE-CIC-IDS2018

Το CSE-CIC-IDS2018 [13] είναι ένα συνεργατικό έργο μεταξύ του Ιδρύματος Ασφάλειας Επικοινωνιών (CSE) και του Καναδικού Ινστιτούτου Κυβερνοασφάλειας (CIC). Πρόκειται για μία ακόμα νεότερη προσπάθεια δημιουργίας ενός επίκαιρου και ανοιχτού συνόλου δεδομένων που περιέχει αντιπροσωπευτικά δείγματα δικτυακής κίνησης, σε έναν χώρο όπου οι επιθέσεις εξελίσσονται διαρκώς.

Για τη δημιουργία του συνόλου αυτού, οι ερευνητές δημιουργούν προφίλ χρηστών που περιέχουν αφηρημένες αναπαραστάσεις γεγονότων και συμπεριφορών που φαίνονται στο δίκτυο. Τα προφίλ αυτά χωρίζονται στα B-προφίλ (benign / καλόβουλα), που αντιστοιχούν σε φυσιολογική δικτυακή κίνηση, και στα M-προφίλ (malicious / κακόβουλος), που αντιστοιχούν σε κυβερνοεπιθέσεις. Ο συνδυασμός αυτών των προφίλ οδηγεί στη δημιουργία ενός διαφοροποιημένου συνόλου δεδομένων. Οι επιθέσεις που συμπεριλαμβάνονται στο CSE-CIC-IDS2018 είναι οι εξής: Brute-force (ωμής δύναμης), Hearbleed, Botnet, DoS (άρνηση υπηρεσιών), DDoS (κατανεμημένη άρνηση υπηρεσιών), Web attacks (επιθέσεις διαδικτύου) και infiltration (διείσδυσης). 50 υπολογιστές χρησιμοποιούνται για την παραγωγή κακόβουλης κίνησης, και 420 υπολογιστές και 30 εξυπηρετητές αποτελούν την υποδομή που δέχεται επίθεση. Οι ερευνητές εξάγουν 80 χαρακτηριστικά από την κίνηση που έχει καταγραφεί χρησιμοποιώντας το εργαλείο CICFlowMeter-V3. Παρακάτω παραθέτουμε πίνακα των επιθέσεων:

Πίνακας 8: Πληροφορίες για τις επιθέσεις που περιλαμβάνονται στο CSE-CIC-IDS2018

Επίθεση	Εργαλεία	Διάρκεια	Επιτιθέμενος	Θύμα
Επίθεση Bruteforce	FTP – Patator SSH – Patator	Μία μέρα	Kali linux	Ubuntu 16.4 (Εξυπηρετητής δικτύου)
Επίθεση DoS	Hulk, GoldenEye, Slowloris, Slowhttptest	Μία μέρα	Kali linux	Ubuntu 16.4 (Apache)
Επίθεση DoS	Heartleech	Μία μέρα	Kali linux	Ubuntu 16.4 (Εξυπηρετητής δικτύου)
Επίθεση Web	• Damn Vulnerable Web App (DVWA) • In-house selenium framework (XSS και Brute-force)	Δύο μέρες	Kali linux	Ubuntu 16.4 (Εξυπηρετητής δικτύου)
Infiltration	• Πρώτο επίπεδο: Κατέβασμα Dropbox σε μηχάνημα Windows	Δύο μέρες	Kali linux	Windows Vista και Macintosh

	• Δεύτερο επίπεδο: Nmap και σάρωση θυρών			
Επίθεση botnet	• Ares: απομακρυσμένο κέλυφος, ανέβασμα/κατέβασμα αρχείων, καταγραφή • Screenshots και key logging	Μία μέρα	Kali linux	Windows Vista, 7, 8.1, 10 (32-bit) και 10 (64-bit)
DDoS + Σάρωση θυρών	Low Orbit Ion Canon (LOIC) for UDP, TCP, ή αιτήματα HTTP	Δύο μέρες	Kali linux	Windows Vista, 7, 8.1, 10 (32-bit) και 10 (64-bit)

Τα χαρακτηριστικά του συνόλου δεδομένων αναλύονται στον ακόλουθο πίνακα:

Πίνακας 9: Τα χαρακτηριστικά του CSE-CIC-IDS2018

Όνομα χαρακτηριστικού	**Περιγραφή**
fl_dur	Διάρκεια ροής
tot_fw_pk	Συνολικά πακέτα στην πρόσθια κατεύθυνση
tot_bw_pk	Συνολικά πακέτα στην αντίθετη κατεύθυνση
tot_l_fw_pkt	Συνολικό μέγεθος πακέτων στην πρόσθια κατεύθυνση
fw_pkt_l_max	Μέγιστο μέγεθος πακέτου στην πρόσθια κατεύθυνση
fw_pkt_l_min	Ελάχιστο μέγεθος πακέτου στην πρόσθια κατεύθυνση
fw_pkt_l_avg	Μέσο μέγεθος πακέτου στην πρόσθια κατεύθυνση
fw_pkt_l_std	Τυπική απόκλιση μεγέθους πακέτου στην πρόσθια κατεύθυνση
bw_pkt_l_max	Μέγιστο μέγεθος πακέτου στην αντίθετη κατεύθυνση
bw_pkt_l_min	Ελάχιστο μέγεθος πακέτου στην αντίθετη κατεύθυνση
bw_pkt_l_avg	Μέσο μέγεθος πακέτου στην αντίθετη κατεύθυνση
bw_pkt_l_std	Τυπική απόκλιση μεγέθους πακέτου στην αντίθετη κατεύθυνση
fl_byt_s	Ρυθμός ροής byte (#bytes / δευτερόλεπτο)
fl_pkt_s	Ρυθμός ροής πακέτων (#πακέτα / δευτερόλεπτο)
fl_iat_avg	Μέσος χρόνος μεταξύ δύο ροών
fl_iat_std	Τυπική απόκλιση χρόνου μεταξύ δύο ροών
fl_iat_max	Μέγιστος χρόνος μεταξύ δύο ροών
fl_iat_min	Ελάχιστος χρόνος μεταξύ δύο ροών
fw_iat_tot	Συνολικός χρόνος μεταξύ δύο πακέτων με πρόσθια κατεύθυνση

fw_iat_avg	Μέσος χρόνος μεταξύ δύο πακέτων με πρόσθια κατεύθυνση
fw_iat_std	Τυπική απόκλιση χρόνου μεταξύ δύο πακέτων με πρόσθια κατεύθυνση
fw_iat_max	Μέγιστος χρόνος μεταξύ δύο πακέτων με πρόσθια κατεύθυνση
fw_iat_min	Ελάχιστος χρόνος μεταξύ δύο πακέτων με πρόσθια κατεύθυνση
bw_iat_tot	Συνολικός χρόνος μεταξύ δύο πακέτων με αντίθετη κατεύθυνση
bw_iat_avg	Μέσος χρόνος μεταξύ δύο πακέτων με αντίθετη κατεύθυνση
bw_iat_std	Τυπική απόκλιση χρόνου μεταξύ δύο πακέτων με αντίθετη κατεύθυνση
bw_iat_max	Μέγιστος χρόνος μεταξύ δύο πακέτων με αντίθετη κατεύθυνση
bw_iat_min	Ελάχιστος χρόνος μεταξύ δύο πακέτων με αντίθετη κατεύθυνση
fw_psh_flag	Αριθμός φορών που η σημαία PSH τέθηκε σε πακέτα με πρόσθια κατεύθυνση (0 για UDP)
bw_psh_flag	Αριθμός φορών που η σημαία PSH τέθηκε σε πακέτα με αντίθετη κατεύθυνση (0 για UDP)
fw_urg_flag	Αριθμός φορών που η σημαία URG τέθηκε σε πακέτα με πρόσθια κατεύθυνση (0 για UDP)
bw_urg_flag	Αριθμός φορών που η σημαία URG τέθηκε σε πακέτα με αντίθετη κατεύθυνση (0 για UDP)
fw_hdr_len	Συνολικά bytes που χρησιμοποιήθηκαν σε επικεφαλίδες στην πρόσθια κατεύθυνση
bw_hdr_len	Συνολικά bytes που χρησιμοποιήθηκαν σε επικεφαλίδες στην αντίθετη κατεύθυνση
fw_pkt_s	Αριθμός από πακέτα πρόσθιας κατεύθυνσης ανά δευτερόλεπτο
bw_pkt_s	Αριθμός από πακέτα αντίθετης κατεύθυνσης 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 άφιξης μεταξύ πακέτων
fin_cnt	Αριθμός πακέτων με FIN
syn_cnt	Αριθμός πακέτων με SYN
rst_cnt	Αριθμός πακέτων με RST
pst_cnt	Αριθμός πακέτων με PST
ack_cnt	Αριθμός πακέτων με ACK
urg_cnt	Αριθμός πακέτων με URG
cwe_cnt	Αριθμός πακέτων με CWE
ece_cnt	Αριθμός πακέτων με ECE
down_up_ratio	Λόγος κατεβάσματος / ανεβάσματος

pkt_size_avg	Μέσο μέγεθος πακέτου
fw_seg_avg	Μέσο μέγεθος που παρατηρήθηκε στην πρόσθια κατεύθυνση
bw_seg_avg	Μέσο μέγεθος που παρατηρήθηκε στην αντίθετη κατεύθυνση
fw_byt_blk_avg	Μέσος αριθμός του bytes bulk rate στην πρόσθια κατεύθυνση
fw_pkt_blk_avg	Μέσος αριθμός του packet bulk rate στην πρόσθια κατεύθυνση
fw_blk_rate_avg	Μέσος αριθμός του bulk rate στην πρόσθια κατεύθυνση
bw_byt_blk_avg	Μέσος αριθμός του bytes bulk rate στην αντίθετη κατεύθυνση
bw_pkt_blk_avg	Μέσος αριθμός του packet bulk rate στην αντίθετη κατεύθυνση
bw_blk_rate_avg	Μέσος αριθμός του bulk rate στην αντίθετη κατεύθυνση
subfl_fw_pk	Ο μέσος αριθμός πακέτων μίας υπο-ροής στην πρόσθια κατεύθυνση
subfl_fw_byt	Ο μέσος αριθμός bytes μίας υπο-ροής στην πρόσθια κατεύθυνση
subfl_bw_pk	Ο μέσος αριθμός πακέτων μίας υπο-ροής στην αντίθετη κατεύθυνση
subfl_bw_byt	Ο μέσος αριθμός bytes μίας υπο-ροής στην αντίθετη κατεύθυνση
fw_win_byt	# bytes που στάλθηκαν στο αρχικό παράθυρο στην πρόσθια κατεύθυνση
bw_win_byt	# bytes που στάλθηκαν στο αρχικό παράθυρο στην αντίθετη κατεύθυνση
fw_act_pkt	# πακέτων με τουλάχιστον 1 byte από δεδομένα TCP στην πρόσθια κατεύθυνση
fw_seg_min	Ελάχιστο μήκος τεμαχίου που παρατηρήθηκε στην πρόσθια κατεύθυνση
atv_avg	Μέσος χρόνος που η ροή ήταν ενεργή πριν γίνει ανενεργή
atv_std	Τυπική απόκλιση χρόνου που η ροή ήταν ενεργή πριν γίνει ανενεργή
atv_max	Μέγιστος χρόνος που η ροή ήταν ενεργή πριν γίνει ανενεργή
atv_min	Ελάχιστος χρόνος που η ροή ήταν ενεργή πριν γίνει ανενεργή
idl_avg	Μέσος χρόνος που η ροή ήταν ανενεργή πριν γίνει ενεργή
idl_std	Τυπική απόκλιση χρόνου που η ροή ήταν ανενεργή πριν γίνει ενεργή
idl_max	Μέγιστος χρόνος που η ροή ήταν ανενεργή πριν γίνει ενεργή
idl_min	Ελάχιστος χρόνος που η ροή ήταν ανενεργή πριν γίνει ενεργή

Σημειώνουμε ότι το πλήρες dataset είναι πάνω από 6GB και κατανεμημένο σε πολλά αρχεία. Συνενώνουμε τα αρχεία αυτά, επιλέγουμε με τυχαίο τρόπο το ένα δέκατο των συνολικών δειγμάτων και σχηματίζουμε το σύνολο δεδομένων στο οποίο θα εργαστούμε. Έπειτα, χωρίζουμε τυχαία αυτό το σύνολο σε σύνολο εκπαίδευσης και αξιολόγησης, με αναλογία 4 προς 1. Παραθέτουμε παρακάτω τον πίνακα με τις κατανομές δειγμάτων στις κλάσεις για τα dataset που προκύπτουν.

Πίνακας 10: Κατανομή εγγραφών σε κλάσεις για το υποσύνολο του CIC-IDS2018

Τύπος	Αριθμός εγγραφών	
	Σύνολο Εκπαίδευσης	Σύνολο Αξιολόγησης
Benign	1.078.936	270.169
DDOS attack-HOIC	54.335	13.716
DDoS attacks-LOIC-HTTP	46.051	11.438
DoS attacks-Hulk	37.064	9.296
Bot	22.877	5.622
FTP-BruteForce	15.421	3.869
SSH-Bruteforce	15.090	3.910
Infiltration	12.900	3.288
DoS attacks-SlowHTTPTest	11.119	2.737
DoS attacks-GoldenEye	3.268	854
DoS attacks-Slowloris	870	211
DDOS attack-LOIC-UDP	140	36
Brute Force-Web	36	12
Brute Force-XSS	19	6
SQL Injection	3	1
Σύνολο	1.298.129	325.165

5.2. Μετρικές αξιολόγησης

Μία μετρική αξιολόγησης περιγράφει τα αποτελέσματα με έναν μονοδιάστατο τρόπο, παρέχοντάς μας ένα «παράθυρο» στην πραγματική επίδοση του μοντέλου υπό εξέταση. Για παράδειγμα, η χρήση μόνο της μετρικής ακρίβειας (που θα αναλυθεί παρακάτω) δεν μας βοηθάει ιδιαίτερα, καθώς τα σύνολα δεδομένων ενδέχεται να παρουσιάζουν έντονη ανισότητα μεταξύ των κλάσεων, με αναλογίες της τάξης του 9 προς 1 ομαλά προς ανώμαλα δείγματα για την περίπτωση του CSE-CIC-IDS2018. Σε αυτήν την περίπτωση, ένα μοντέλο που πάντα αποφαίνεται ότι το δείγμα εισόδου είναι ομαλό παρουσιάζει 90% απόδοση, το οποίο είναι έντονα ανεπιθύμητο. Για την αποφυγή του παραπάνω, χρησιμοποιούμε μία ευρεία ποικιλία μετρικών απόδοσης.

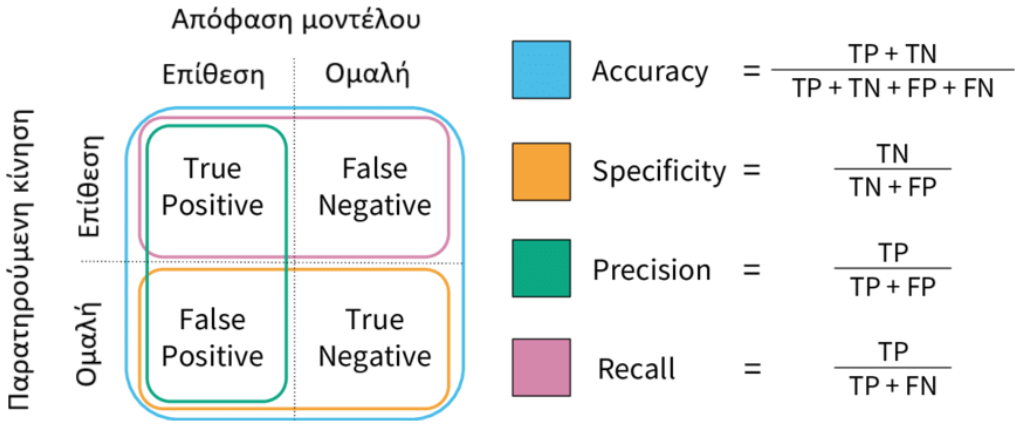
Στον τομέα του εντοπισμού ανωμαλιών σε δίκτυα η σύμβαση είναι ότι τα ανώμαλα δείγματα χαρακτηρίζονται «θετικά», ενώ τα ομαλά «αρνητικά». Συμβολίζουμε με TP τα true positives (δείγματα που το μοντέλο αποφαίνεται ορθά ότι είναι ανώμαλα), TN τα true negatives (το μοντέλο αποφαίνεται ορθά ότι είναι ομαλά), FP τα false positives (το μοντέλο αποφαίνεται λανθασμένα ότι είναι ανώμαλα), και FN τα false negatives (το μοντέλο αποφαίνεται λανθασμένα ότι είναι ομαλά).

- $Accuracy$: Η στατιστική πιθανότητα το μοντέλο να λάβει την σωστή απόφαση για ένα δείγμα. Αλγεβρικά έχουμε:
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
- $Precision$: Η στατιστική πιθανότητα το μοντέλο έχοντας αποφανθεί ότι ένα δείγμα είναι θετικό (ανώμαλο), να έχει αποφανθεί σωστά. Αλγεβρικά:
$$Precision = \frac{TP}{TP + FP}$$
- $Recall$: Η στατιστική πιθανότητα το μοντέλο έχοντας λάβει ένα ανώμαλο δείγμα στην είσοδό του, να το εντοπίσει. Αλγεβρικά:
$$Recall = \frac{TP}{TP + FN}$$
- $F1\ score$: Ο αρμονικός μέσος όρος των μετρικών Precision και Recall, σχεδιασμένος ώστε να συμφιλιώνει αυτές τις δύο μετρικές, παρέχοντας έναν μη αρνητικό αριθμό που μπορεί να χρησιμοποιείται για τη σύγκριση μοντέλων. Αλγεβρικά:
$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$
- $Specificity$: Η στατιστική πιθανότητα το μοντέλο, έχοντας λάβει ένα ομαλό δείγμα στην είσοδό του, να το εντοπίσει. Αλγεβρικά:
$$Specificity = \frac{TN}{TN + FP}$$
- $Area\ Under\ the\ Receiver\ Operator\ Characteristic\ (AUROC)$: Πρόκειται για μία μονοδιάστατη μετρική, η οποία όμως, σε αντίθεση με τις προηγούμενες, λαμβάνει υπόψη τη διακριτική ικανότητα του μοντέλου για όλες τις τιμές κατωφλίου σκορ

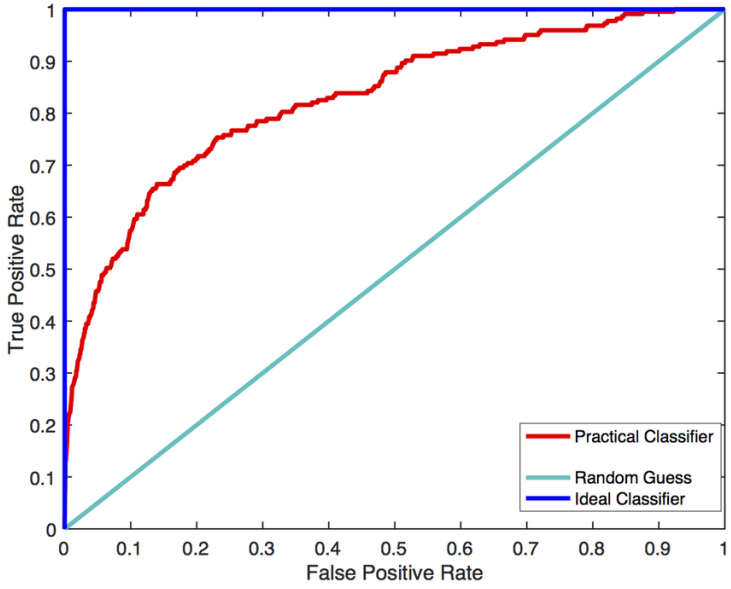
ανωμαλίας. Πιο συγκεκριμένα, ενώ οι προηγούμενες μετρικές λαμβάνουν ως είσοδο την απόφαση του μοντέλου (ομαλό / μη ομαλό δείγμα), η συγκεκριμένη λαμβάνει τα σκορ ανωμαλίας των δειγμάτων. Ο υπολογισμός της μετρικής γίνεται με τη σχεδίαση της καμπύλης ROC, που απαρτίζεται από τα σημεία $(ρυθμός\ False\ Positive, ρυθμός\ True\ positive) = \left(\frac{FP}{FP+TN}, \frac{TP}{TP+FN}\right)$ για όλα τα δυνατά κατώφλια ανωμαλίας. Το εμβαδό κάτω από την καμπύλη ROC είναι η τιμή της συγκεκριμένης μετρικής. Τιμές AUROC κοντά στο 0.5 υποδεικνύουν καμπύλη που μοιάζει με την ευθεία $y = x$ και αντιστοιχούν σε τυχαία ταξινόμηση, ενώ τιμές κοντά στο 1 υποδεικνύουν καμπύλη κοντά στις ευθείες $y = 0$ και $x = 1$, δηλαδή ιδανική ταξινόμηση. Τα προηγούμενα φαίνονται και στο Σχήμα 38.

- $Confusion\ matrix$: Ο πίνακας σύγχυσης διαφέρει από τις προηγούμενες μετρικές, καθώς δεν είναι ένας πραγματικός αριθμός, αλλά ένας πίνακας που περιέχει πλήρη εικόνα της διαχωριστικής ικανότητας του μοντέλου. Πιο συγκεκριμένα, πρόκειται για έναν πίνακα $num\_classes \times num\_classes$, εν προκειμένω 2×2, αφού έχουμε δυαδική ταξινόμηση, όπου οι σειρές αντιστοιχούν στις πραγματικές κλάσεις και οι στήλες στις κλάσεις που αναθέτει το μοντέλο στα δείγματα. Σε κάθε εσωτερικό κελί αναγράφεται ο αριθμός δειγμάτων της κλάσης που αντιστοιχεί στη σειρά του που ανατέθηκαν στην κλάση που αντιστοιχεί στη στήλη του.

Σχηματικά έχουμε τα παρακάτω:



Σχήμα 37: Ορισμός και σχηματική αναπαράσταση μετρικών αξιολόγησης (Τροποποιημένη έκδοση του [61])



Σχήμα 38: Η καμπύλη ROC. Πηγή [62]

5.3. Προεπεξεργασία των συνόλων δεδομένων

Όσον αφορά τα datasets, χρησιμοποιούμε το υποσύνολο του UNSWNB15 που παρουσιάζουν οι ερευνητές, και ένα τυχαία εκλεγόμενο υποσύνολο του CSE-CIC-IDS2018 στο ένα δέκατο του μεγέθους, όπως αναφέραμε και στις αντίστοιχες υποενότητες. Μόνη εξαίρεση αποτελεί το μοντέλο συνελικτικού αυτοκωδικοποιητή, το οποίο θα αναλύσουμε στη συνέχεια.

Για την προεπεξεργασία των συνόλων δεδομένων εφαρμόζουμε τα ακόλουθα:

Καταρχάς, αφαιρούμε οποιοδήποτε δείγμα έχει κενές εγγραφές. Τα συγκεκριμένα datasets δεν παρουσιάζουν έλλειψη χαρακτηριστικών, οπότε αυτό το στάδιο εφαρμόζεται μόνο για λόγους πληρότητας και επεκτασιμότητας. Έπειτα, υλοποιούμε την επιλογή για ταξινόμηση του συνόλου δεδομένων βάσει ενός χαρακτηριστικού, το οποίο αξιοποιεί ο συνελικτικός αυτοκωδικοποιητής. Στη συνέχεια, χαρακτηριστικά τα οποία δεν συνεισφέρουν πληροφορία, όπως το “id” και “attack_cat” (η κατηγορία της επίθεσης δεν μας απασχολεί σε αυτό το έργο, καθώς ασχολούμαστε αμιγώς με δυαδική ταξινόμηση) για το UNSW-NB15 και το “timestamp” για το CIC-IDS-2018, διαγράφονται, και τα χαρακτηριστικά ετικέτας διαχωρίζονται από τα υπόλοιπα. Ακολουθεί η κωδικοποίηση των κατηγορικών χαρακτηριστικών. Εφαρμόζουμε One-Hot κωδικοποίηση, κατά την οποία κάθε κατηγορικό χαρακτηριστικό επεκτείνεται σε τόσα χαρακτηριστικά όσα και οι διαφορετικές κατηγορίες, με κάθε χαρακτηριστικό να λαμβάνει την τιμή 1 όταν το δείγμα ανήκει στην αντίστοιχη κατηγορία, και 0 σε αντίθετη περίπτωση. Αυτό βέβαια καθίσταται προβληματικό όταν υπάρχει μεγάλος αριθμός διαφορετικών κατηγοριών, καθώς η διάσταση της εισόδου αυξάνεται υπερβολικά, πλήττοντας έντονα τη διαδικασία μάθησης. Για να αποφύγουμε αυτό το πρόβλημα, υλοποιούμε έναν μηχανισμό αντικατάστασης κάθε κατηγορίας με λιγότερες εμφανίσεις από ένα κατώφλι από τη γενική κατηγορία «άλλο» (“other”). Κρίνουμε ότι η επιλογή λογικών κατωφλίων δεν προκαλεί κάποια σημαντική απώλεια πληροφορίας, αφού άλλωστε τα μοντέλα βαθιάς μάθησης δεν είναι ευαίσθητα σε τιμές που εμφανίζονται με πολύ μικρή συχνότητα. Έπειτα από επισκόπηση των datasets, αυτή η τροποποίηση εφαρμόζεται μόνο στο UNSW-NB15, για το χαρακτηριστικό “proto” με κατώφλι τις 250 εμφανίσεις και το “state” με 2 εμφανίσεις. Μετά την κωδικοποίηση των κατηγορικών χαρακτηριστικών, κανονικοποιούμε τα αριθμητικά δεδομένα. Η κανονικοποίηση αποτελεί κρίσιμο στάδιο του σταδίου προεπεξεργασίας, καθώς επιτρέπει την ομαλή ρύθμιση των παραμέτρων του μοντέλου κατά την εκπαίδευση. Υπάρχουν πολλές επιλογές, με κυριότερες την min-max και την z-score. Επιλέγουμε την κανονικοποίηση min-max, η οποία βρίσκει το ελάχιστο και μέγιστο κάθε χαρακτηριστικού και κανονικοποιεί γραμμικά ώστε κάθε δείγμα εκπαίδευσης να βρίσκεται στο διάστημα $[0,1]$. Με μαθηματικό συμβολισμό, έχουμε το ακόλουθο:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}},$$

όπου $X, X_{\min}, X_{\max}, X'$ είναι η παλιά τιμή ενός χαρακτηριστικού, η ελάχιστη και μέγιστη τιμή του χαρακτηριστικού στο σύνολο εκπαίδευσης, και η νέα τιμή του χαρακτηριστικού αντίστοιχα.

Τα παραπάνω ισχύουν για τα πρώτα μοντέλα τα οποία παρουσιάζουμε. Η δομή του συνελικτικού αυτοκωδικοποιητή επιβάλλει μερικές αλλαγές στη μεθοδολογία. Πρώτη και σημαντικότερη είναι, όπως προαναφέραμε, η ταξινόμηση των εγγραφών με βάση τον χρόνο, ώστε να μπορεί να αξιοποιηθεί σωστά η χρονική διάσταση της εισόδου. Δεύτερη προσθήκη είναι η εφαρμογή του μετασχηματισμού PCA στα δεδομένα ως τελευταίο στάδιο προεπεξεργασίας, για την συμπίεση, αλλαγή διαστάσεων σε δύναμη του δύο και μεγαλύτερη ανεξαρτησία μεταξύ των χαρακτηριστικών, όπως αναλύσαμε στη σχετική ενότητα περιγραφής του μοντέλου. Επιπλέον, σημειώνουμε ότι ο υπολογισμός της συσχέτισης Pearson γίνεται με οκνηρό τρόπο, αμέσως πριν αυτή αξιοποιηθεί από το μοντέλο, καθώς η αποθήκευσή της για όλα τα δείγματα απαιτεί απαγορευτικά μεγάλο αποθηκευτικό χώρο. Αυτό το γεγονός καθιστά την εκπαίδευση του μοντέλου αργή, και μας αναγκάζει να περιοριστούμε σε ένα υποσύνολο του dataset για εκπαίδευση και έλεγχο. Για αυτόν τον λόγο επιλέγουμε στο UNSW-NB15 από το σύνολο εκπαίδευσης τα πρώτα 100.000 δείγματα για εκπαίδευση και τα 75.341 που μένουν για αξιολόγηση, και στο CSE-CIC-IDS2018 τα δεδομένα της Πέμπτης 01-03-2018, με τα πρώτα 100.000 δείγματα να χρησιμοποιούνται για εκπαίδευση και τα επόμενα 100.000 για αξιολόγηση. Κρίνουμε πως αυτά τα σύνολα δεδομένων είναι αρκετά συνεκτικά ώστε ένα μοντέλο μηχανικής μάθησης με την απαραίτητη χωρητικότητα να αναμένεται να έχει ικανοποιητική απόδοση.

5.4. Πειραματική διάταξη

Όλα τα πειράματα έχουν διεξαχθεί σε python με χρήση του προγραμματιστικού πλαισίου Pytorch [63]. Χρησιμοποιούμε την έκδοση *2.1.1+cu121* του Pytorch, σε υπολογιστή με λειτουργικό Windows 10 64-bit, 16 GB RAM και κάρτα γραφικών NVIDIA GTX 1660.

Ο σκοπός του έργου αυτού είναι η εξερεύνηση μίας καινούριας μεθόδου τροποποίησης ενός μοντέλου μη επιβλεπόμενης μηχανικής μάθησης για την ημιεπιβλεπόμενη μάθηση με αντιπαραδείγματα, και η απόδειξη της αξίας της. Η επίτευξη της βέλτιστης απόδοσης κάθε μοντέλου για την επίτευξη state-of-the-art δεν είναι στους στόχους μας. Ως εκ τούτου, οι τιμές των υπερπαραμέτρων ρυθμίζονται χειροκίνητα σε τιμές που εμφανίζουν ικανοποιητικά αποτελέσματα, και δεν απαιτείται η χρήση συνόλου επικύρωσης (validation set). Στις υποενότητες που ακολουθούν, αναλύουμε αρχιτεκτονικές λεπτομέρειες ανά μοντέλο που εξετάζουμε.

Παραθέτουμε τώρα μερικές γενικές λεπτομέρειες υλοποίησης, κοινές σε όλες τις περιπτώσεις.

Το πλήθος χαρακτηριστικών εισόδου είναι 67 για το UNSWNB15 και 78 για το CSE-CIC-IDS2018, μετά την προεπεξεργασία που περιγράψαμε στην προηγούμενη ενότητα. Και στις δύο περιπτώσεις, η διάσταση του λανθάνοντος χώρου επιλέγεται ίση με 32.

Τα βάρη αρχικοποιούνται σε όλες τις περιπτώσεις με εκλογή ανεξάρτητων δειγμάτων από γκαουσιανή κατανομή μέσου $\mu = 0$ και τυπικής απόκλισης $\sigma = 0.02$. Οι σταθεροί όροι αρχικοποιούνται στο 0.

Εκτελούμε εκπαίδευση για 20 εποχές με χρήση backpropagation [36] και του βελτιστοποιητή Adam [43]. Οι παράμετροι του Adam είναι: ρυθμός εκπαίδευσης 0.0001 και $\beta_1 = 0.5, \beta_2 = 0.999$. Κατά την εκπαίδευση, το μέγεθος παρτίδας των ομαλών δειγμάτων είναι 64, ενώ όταν χρησιμοποιούνται ανώμαλα αντιπαραδείγματα, αυτά έχουν μέγεθος παρτίδας 16. Κατά την αξιολόγηση χρησιμοποιούμε μέγεθος παρτίδας 256, καθώς αυτό οδηγεί σε καλύτερη αξιοποίηση των δυνατοτήτων της κάρτας γραφικών χωρίς να διαφέρει από άλλα μεγέθη. Το κατώφλι ανωμαλίας τίθεται αυτόματα, με τον εξής τρόπο: Τα σκορ ανωμαλίας του μοντέλου για τα δείγματα ελέγχου ταξινομούνται σε αύξουσα σειρά, και το κορυφαίο α% θεωρείται ανώμαλο, ενώ τα υπόλοιπα ομαλά. Το α είναι το ποσοστό των μη ομαλών δειγμάτων στο σύνολο ελέγχου και είναι ίσο με 55.06% στο UNSW-NB15 και 16.97% στο CIC-IDS-2018.

5.4.1. Αυτοκωδικοποιητής

Όπως προαναφέραμε κατά τη γενική περιγραφή των μοντέλων, αξιοποιούμε ασύμμετρη αρχιτεκτονική κωδικοποιητή – αποκωδικοποιητή, με τον αποκωδικοποιητή να ενισχύεται με ένα επιπλέον πλήρως συνδεδεμένο επίπεδο για αύξηση της εκφραστικής δύναμης. Σημειώνουμε ότι σε σύγκριση με άλλα έργα της βιβλιογραφίας, όπως στον παραλλακτικό αυτοκωδικοποιητή του [9], οι αρχιτεκτονικές των αυτοκωδικοποιητών έχουν πολύπλοκη εσωτερική δομή. Αποφεύγουμε συνειδητά αυτήν την επιλογή ώστε να αναδείξουμε την ικανότητα έντονης ενίσχυσης ακόμα και απλών αρχιτεκτονικών που επιτρέπει η τροποποίησή μας.

Πίνακας 11: Αρχιτεκτονική κωδικοποιητή (AE)

Κωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (Αρ. χαρακτηριστικών × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × διάσταση λανθάνοντα χώρου)

Πίνακας 12: Αρχιτεκτονική αποκωδικοποιητή (AE)

Αποκωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (διάσταση λανθάνοντα χώρου × 64)
ReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 128)
Ομαλοποίηση επιπέδου (128)
ReLU
Πλήρως συνδεδεμένο επίπεδο (128 × Αρ. χαρακτηριστικών)
Σιγμοειδής

Στον κωδικοποιητή χρησιμοποιούμε LeakyReLU [64] με συντελεστή κλίσης 0.2, ενώ στον αποκωδικοποιητή χρησιμοποιούμε απλή ReLU (η οποία χρησιμοποιήθηκε για πρώτη φορά στο [65]) στα ενδιάμεσα στρώματα και σιγμοειδή στο επίπεδο εξόδου. Η επιλογή της σιγμοειδούς γίνεται καθώς η εφαρμογή min-max κλιμάκωσης απεικονίζει όλα τα δεδομένα εκπαίδευσης στο διάστημα $[0,1]$. Το γεγονός ότι ένα δείγμα του συνόλου ελέγχου μπορεί να βρίσκεται εκτός αυτού του διαστήματος δεν μας απασχολεί, καθώς τότε θα αντιστοιχεί με μεγάλη πιθανότητα σε κακόβουλη κίνηση, της οποίας την κακή ανακατασκευή επιδιώκουμε. Αναφορικά με την υπερπαράμετρο θ, για την επιλογή της καταλληλότερης τιμής ελέγχουμε την απόδοση του μοντέλου για τιμές $[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1, 1, 10]$.

5.4.2. Παραλλακτικός αυτοκωδικοποιητής

Για τον παραλλακτικό αυτοκωδικοποιητή υιοθετούμε παραπλήσια αρχιτεκτονική με τον απλό αυτοκωδικοποιητή. Οι σχετικοί πίνακες παρατίθενται παρακάτω:

Πίνακας 13: Αρχιτεκτονική κωδικοποιητή (VAE)

Κωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (Αρ. χαρακτηριστικών × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 2 · διάσταση λανθάνοντα χώρου)

Πίνακας 14: Αρχιτεκτονική αποκωδικοποιητή (VAE)

Αποκωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (διάσταση λανθάνοντα χώρου × 64)
ReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 128)
Ομαλοποίηση επιπέδου (128)
ReLU
Πλήρως συνδεδεμένο επίπεδο (128 × Αρ. χαρακτηριστικών)
Σιγμοειδής

Όπου η συνάρτηση LeakyReLU έχει και εδώ συντελεστή κλίσης 0.2.

Στην πρώτη ρύθμιση, στην οποία χρησιμοποιούμε την παραδοσιακή απώλεια του β-VAE, το β τίθεται ίσο με 0.5, δίνοντας μεγαλύτερη έμφαση στην ανακατασκευή [47]. Η τιμή β είναι ίδια και στα δύο datasets και έχει επιλεχθεί μετά από δοκιμές, εμφανίζοντας ικανοποιητική απόδοση. Εστιάζουμε στην επίδραση της δικής μας, ημιεπιβλεπόμενης, προσθήκης, οπότε δεν κρίνουμε απαραίτητο να παραθέσουμε την πλήρη διαδικασία επιλογής της παραμέτρου β.

Στη δεύτερη ρύθμιση, έχουμε τιμή γ ίση με 10 και μέγιστη χωρητικότητα $C_{\max}$ ίση με 10, και για τα δύο σύνολα δεδομένων. Εντούτοις, λόγω της έντονης διαφοροποίησης στον αριθμό εγγραφών, και συνεπώς των αριθμών παρτίδων και επαναλήψεων στον βρόχο εκπαίδευσης, η υπερπαράμετρος των επαναλήψεων μέχρι την επίτευξη μέγιστης χωρητικότητας τίθεται ανάλογα με το σύνολο δεδομένων. Για το UNSW-NB15 θέτουμε την τιμή ίση με 15.000 επαναλήψεις και για το CIC-IDS2018 ίση με 10^5 επαναλήψεις.

Αναφορικά με την τιμή της υπερπαραμέτρου θ, εξετάζουμε μόνο την τιμή που εμφάνισε τις υψηλότερες επιδόσεις στην περίπτωση του απλού αυτοκωδικοποιητή, λόγω της υψηλής ομοιότητας των δύο αρχιτεκτονικών. Αυτή η τιμή είναι η $\theta = 0.001$, όπως θα αναλύσουμε στο κεφάλαιο των αποτελεσμάτων.

5.4.3. Προσαρμογή του μοντέλου GANomaly

Παραθέτουμε καταρχάς τη λεπτομερή αρχιτεκτονική δομή των μονάδων του μοντέλου GANomaly_variant που υλοποιήσαμε.

Πίνακας 15: Αρχιτεκτονική κωδικοποιητή γεννήτορα (GANomaly)

Κωδικοποιητής του Γεννήτορα
Πλήρως συνδεδεμένο επίπεδο (Αρ. χαρακτηριστικών × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × διάσταση λανθάνοντα χώρου)

Πίνακας 16: Αρχιτεκτονική αποκωδικοποιητή γεννήτορα (GANomaly)

Αποκωδικοποιητής του Γεννήτορα
Πλήρως συνδεδεμένο επίπεδο (διάσταση λανθάνοντα χώρου × 64)
ReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 128)
Ομαλοποίηση επιπέδου (128)
ReLU
Πλήρως συνδεδεμένο επίπεδο (128 × Αρ. χαρακτηριστικών)
Σιγμοειδής

Πίνακας 17: Αρχιτεκτονική κωδικοποιητή (GANomaly)

Κωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (Αρ. χαρακτηριστικών × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × διάσταση λανθάνοντα χώρου)

Πίνακας 18: Αρχιτεκτονική διαχωριστή (GANomaly)

Διαχωριστής
Πλήρως συνδεδεμένο επίπεδο (Αρ. χαρακτηριστικών × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 1)
Σιγμοειδής

Δομούμε τις παραγωγικές μονάδες του GANomaly (κωδικοποιητές και αποκωδικοποιητή) ακριβώς όπως στον αυτοκωδικοποιητή. Αυτό επιτρέπει τη σύγκριση των αρχιτεκτονικών ως προς τις μετρικές τους, όπως αναφέραμε σε προηγούμενη ενότητα.

Όπως προηγουμένως, στους κωδικοποιητές χρησιμοποιούμε LeakyReLU με συντελεστή κλίσης 0.2. Ο διαχωριστής έχει όμοια δομή με τους κωδικοποιητές, με διαφορά ότι το τελευταίο επίπεδο έχει μία έξοδο στο $[0,1]$, οπότε εφαρμόζουμε και μία σιγμοειδή συνάρτηση ενεργοποίησης.

Στη δημοσίευση του GANomaly [11] οι ερευνητές παρατηρούν βέλτιστη απόδοση με χρήση βαρών $w_{adv} = 1, w_{con} = 50, w_{enc} = 1$. Με δοκιμές παρατηρούμε ότι και στο τρέχον πρόβλημα οι συγκεκριμένες τιμές παράγουν ικανοποιητικά αποτελέσματα, οπότε τις υιοθετούμε. Στην ημιεπιβλεπόμενη ρύθμιση με αντιπαραδείγματα, η υπερπαράμετρος θ λαμβάνει τιμή 0.001, καθώς η δομή του αυτοκωδικοποιητή (γεννήτορα) είναι ίδια με αυτή του πρώτου μοντέλου που παρουσιάζουμε, και συνεπώς αναμένουμε η συμπεριφορά ως προς το θ να είναι παρόμοια.

5.4.4. Μοντέλο BiWGAN-GP

Το συγκεκριμένο μοντέλο πρόκειται για το μόνο μοντέλο που έχει υλοποιηθεί στοχευμένα για τον τομέα του εντοπισμού ανωμαλιών σε δίκτυα. Εντούτοις, δεν εντοπίσαμε υλοποίηση του μοντέλου από τους ερευνητές. Έτσι, ανακατασκευάζουμε το BiWGAN-GP ακολουθώντας πιστά τις περιγραφές των [12]. Στη συνέχεια, διαφοροποιούμαστε σε μερικές λεπτομέρειες της υλοποίησης, τις οποίες θα αναφέρουμε παρακάτω:

Πρώτη σημαντική διαφοροποίηση είναι η αντικατάσταση της ομαλοποίησης παρτίδας (Batch Normalization [56]) από την ομαλοποίηση επιπέδου (Layer Normalization [57]). Ο λόγος είναι ότι η αρχιτεκτονική χρησιμοποιεί απώλεια Wasserstein, η οποία σύμφωνα με τους ερευνητές που παρουσίασαν το Wasserstein GAN [53] δεν είναι συμβατή με τη χρήση ομαλοποίησης παρτίδας, αλλά μπορεί να ενισχυθεί αποτελεσματικά με τη χρήση ομαλοποίησης επιπέδου.

Δεύτερη διαφοροποίηση είναι η αφαίρεση του dropout, η οποία παρατηρήσαμε πειραματικά ότι δεν πλήττει την απόδοση.

Τρίτη διαφοροποίηση αποτελεί η αρχικοποίηση βαρών με ανεξάρτητη δειγματοληψία από κανονική κατανομή, αντί για Xavier αρχικοποίηση [66]. Κρίνουμε την επιλογή της Xavier αρχικοποίησης άστοχη, καθώς αυτή βοηθά επίπεδα με συναρτήσεις ενεργοποίησης όπως η σιγμοειδής [66], ενώ η πλειοψηφία των συναρτήσεων ενεργοποίησης στο συγκεκριμένο μοντέλο είναι στην οικογένεια ReLU.

Τέταρτη διαφοροποίηση είναι η χρήση 20 εποχών αντί για 200 που χρησιμοποιούν οι ερευνητές. Σύμφωνα με τους συγγραφείς [12] η σύγκλιση επιτυγχάνεται σχετικά κοντά στην περιοχή των 20 εποχών. Επιπλέον, ο στόχος μας είναι να παρατηρήσουμε την επίδραση που θα έχει η προσθήκη αντιπαραδειγμάτων σε αυτήν την αρχιτεκτονική, και όχι η μεγιστοποίηση της απόδοσής της. Θεωρούμε λοιπόν αποδεκτή τη μικρή απώλεια της απόδοσης που θα προκύψει από τη μείωση των εποχών εκπαίδευσης.

Αναφορικά με τις υπερπαραμέτρους σ (βάρος απώλειας κύκλου) και n_critic (η αναλογία ανανεώσεων βαρών του διαχωριστή και ταξινομητή προς του γεννήτορα και κωδικοποιητή, συμβολίζεται m στο [12]), αυτές τίθενται ίσες με 10 και 5 αντίστοιχα και στα δύο datasets.

Παραθέτουμε τώρα τη λεπτομερή αρχιτεκτονική δομή των μονάδων του μοντέλου.

Πίνακας 19: Αρχιτεκτονική κωδικοποιητή (BiWGAN-GP)

Κωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (Αρ. χαρακτηριστικών × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × διάσταση λανθάνοντα χώρου)

Πίνακας 20: Αρχιτεκτονική γεννήτορα (BiWGAN-GP)

Γεννήτορας
Πλήρως συνδεδεμένο επίπεδο (διάσταση λανθάνοντα χώρου × 64)
ReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 128)
Ομαλοποίηση επιπέδου (128)
ReLU
Πλήρως συνδεδεμένο επίπεδο (128 × Αρ. χαρακτηριστικών)
Σιγμοειδής

Πίνακας 21: Αρχιτεκτονική διαχωριστή (BiWGAN-GP)

Διαχωριστής
Πλήρως συνδεδεμένο επίπεδο ((Αρ. χαρακτηριστικών + διάσταση λανθάνοντα χώρου) × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 1)
Σιγμοειδής

Πίνακας 22: Αρχιτεκτονική ταξινομητή (BiWGAN-GP)

Ταξινομητής
Πλήρως συνδεδεμένο επίπεδο (διάσταση λανθάνοντα χώρου × 64)
Ομαλοποίηση επιπέδου (64)
LeakyReLU
Πλήρως συνδεδεμένο επίπεδο (64 × 1)
Σιγμοειδής

Η συνάρτηση LeakyReLU χρησιμοποιείται με συντελεστή κλίσης 0.2.

Στην ημιεπιβλεπόμενη ρύθμιση με αντιπαραδείγματα, δοκιμάζουμε αρχικά τιμή θ ίση με 0.001, όπως στα προηγούμενα μοντέλα. Εντούτοις, θα παρατηρήσουμε ότι στο UNSW-NB15 αυτή η ρύθμιση δεν βελτιώνει ικανοποιητικά, οπότε ελέγχουμε και για τα δύο σύνολα δεδομένων και την τιμή $\theta = 0.01$.

5.4.5. Μοντέλο συνελικτικού αυτοκωδικοποιητή

Η υλοποίηση του συνελικτικού αυτοκωδικοποιητή αποκλίνει μερικώς από αυτή των προηγούμενων μοντέλων. Καταρχάς, τα δεδομένα εισόδου ταξινομούνται και εισάγονται κατά αύξουσα σειρά ID για το UNSW-NB15 και timestamp για το CIC-IDS-2018, ώστε να διασφαλιστεί η σωστή κωδικοποίηση της χρονικής τους πληροφορίας. Έπειτα, το μοντέλο υπολογίζει τον πίνακα συσχέτισης κάθε δείγματος με τα προηγούμενά του και συλλέγει με αυτόν τον τρόπο μια παρτίδα 64 πινάκων εισόδου. Αυτή η παρτίδα περιέχει μεικτά ομαλά και ανώμαλα δείγματα σε μεταβλητές αναλογίες, σε αντίθεση με τα προηγούμενα μοντέλα, στα οποία χρησιμοποιούμε χωριστούς Pytorch Dataloaders για τα δύο είδη εγγραφών. Τέλος, το μοντέλο αναμενόμενα έχει διαφορετική εσωτερική αρχιτεκτονική από τα προηγούμενα, αφού περιλαμβάνει συνελικτικά μπλοκ, και δεν χρησιμοποιείται ομαλοποίηση. Η ομαλοποίηση αποφεύγεται, καθώς προσφέρει μόνο μικρές βελτιώσεις και, όπως θα σχολιάσουμε και στο τμήμα των αποτελεσμάτων, το μοντέλο αυτό αποδίδει ανεπαρκώς.

Παραθέτουμε τις λεπτομέρειες της αρχιτεκτονικής παρακάτω:

Πίνακας 23: Αρχιτεκτονική κωδικοποιητή (ConvAE)

Κωδικοποιητής
Συνέλιξη 2D ($\mathbf{1} \rightarrow \mathbf{4}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 2, συμπλήρωση μηδενικών 1)
Tanh
Συνέλιξη 2D ($\mathbf{4} \rightarrow \mathbf{8}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 2, συμπλήρωση μηδενικών 1)
Tanh
Συνέλιξη 2D ($\mathbf{8} \rightarrow \mathbf{16}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 2, συμπλήρωση μηδενικών 1)
Tanh
Συνέλιξη 2D (16→ $\mathbf{32}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 4, συμπλήρωση μηδενικών 1)
Tanh
Πλήρως συνδεδεμένο επίπεδο (32 × διάσταση λανθάνοντα χώρου)

Πίνακας 24: Αρχιτεκτονική αποκωδικοποιητή (ConvAE)

Αποκωδικοποιητής
Πλήρως συνδεδεμένο επίπεδο (διάσταση λανθάνοντα χώρου × 32)
Αντίστροφη Συνέλιξη 2D ($\mathbf{32} \rightarrow \mathbf{16}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 4, συμπλήρωση μηδενικών 0)
Tanh
Αντίστροφη Συνέλιξη 2D ($\mathbf{16} \rightarrow \mathbf{8}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 2, συμπλήρωση μηδενικών 1)
Tanh

Αντίστροφη Συνέλιξη 2D ($\mathbf{8} \rightarrow \mathbf{4}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 2, συμπλήρωση μηδενικών 1)
Tanh
Αντίστροφη Συνέλιξη 2D ($\mathbf{4} \rightarrow \mathbf{1}$ κανάλια, πυρήνας $\mathbf{4} \times \mathbf{4}$, βήμα 2, συμπλήρωση μηδενικών 1)
Tanh

Ως συνάρτηση ενεργοποίησης χρησιμοποιείται η Tanh, καθώς το πεδίο τιμών της είναι το $[-1,1]$, το οποίο ταυτίζεται με το πεδίο τιμών της συσχέτισης Pearson. Επίσης, σημειώνουμε ότι το παράθυρο συσχέτισης έχει μέγεθος 5 δείγματα.

Στην ημιεπιβλεπόμενη ρύθμιση με αντιπαραδείγματα, ελέγχουμε τις τιμές θ $[0.0001,0.001,0.01,0.1,1,10]$, για να διαπιστώσουμε εάν υπάρχει κάποια για την οποία το μοντέλο εμφανίζει βελτίωση.

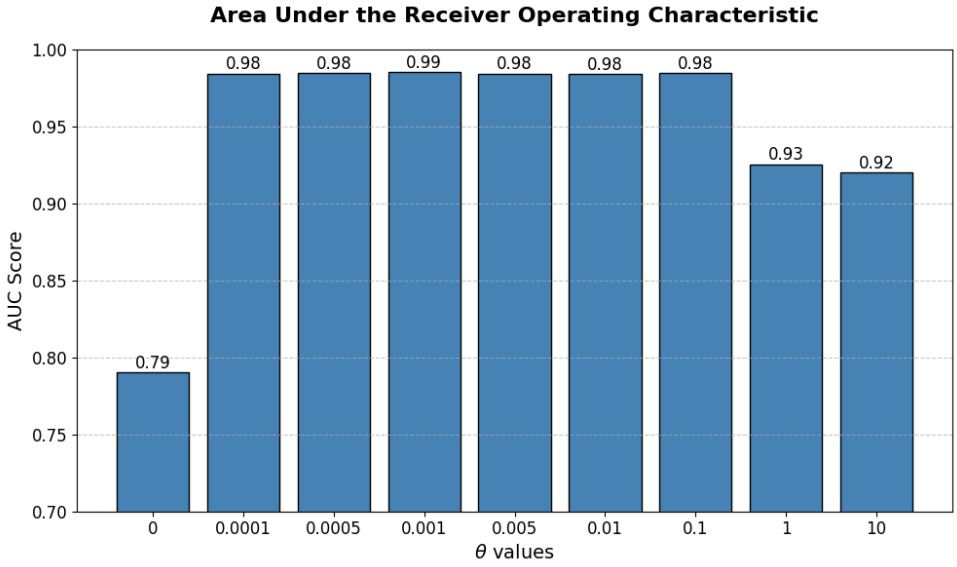
6. Αποτελέσματα

6.1. Απλός Αυτοκωδικοποιητής

Για όλες τις τιμές της υπερπαραμέτρου θ υπολογίζουμε τις μετρικές F1 και AUROC, όπου το F1 υπολογίζεται με τη μέθοδο που περιγράψαμε. Αφού επιλέξουμε την καταλληλότερη τιμή θ για αυτό το μοντέλο θα παρουσιάσουμε αναλυτικότερα τις πλήρεις τιμές μετρικών που αντιστοιχούν σε αυτήν, αλλά και στην περίπτωση εκπαίδευσης χωρίς αντιπαραδείγματα. Με αυτόν τον τρόπο, θα είμαστε σε θέση να παρατηρήσουμε απρόσκοπτα την επίδραση της προσθήκης αντιπαραδειγμάτων στην απόδοση του απλού αυτοκωδικοποιητή.

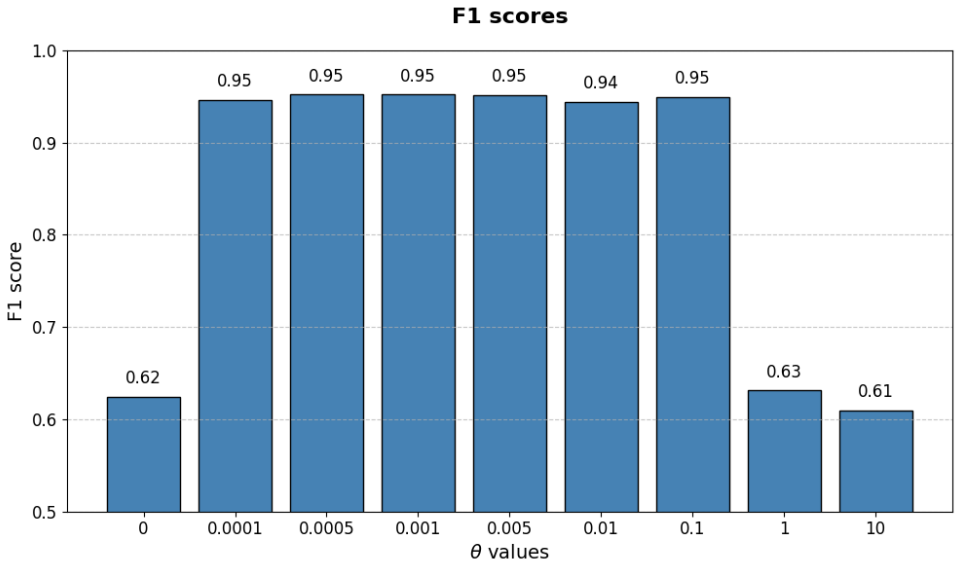
6.1.1. CIC-IDS-2018

Τα αποτελέσματα των πειραμάτων στο CIC-IDS-2018 παρατίθενται παρακάτω:



Σχήμα 39: Τιμές AUROC του AE στο CIC-IDS2018

Σημειώνουμε ότι η περίπτωση $\theta = 0$ αντιστοιχεί στη μη χρήση αντιπαραδειγμάτων.



Σχήμα 40: Τιμές F1 score του AE στο CIC-IDS2018

Καταρχάς, παρατηρούμε ότι ο απλός αυτοκωδικοποιητής δεν εμφανίζει ιδιαίτερα ικανοποιητικά αποτελέσματα όταν εκπαιδεύεται μόνο με χρήση ομαλών δειγμάτων. Αυτό είναι λογικό, καθώς η εσωτερική του δομή είναι αρκετά απλή, και έτσι η μάθηση των ομαλών δειγμάτων γίνεται με απλοϊκό τρόπο, αγνοώντας τις κρίσιμες λεπτομέρειες που διαφοροποιούν την καλόβουλη από την κακόβουλη δικτυακή κίνηση. Η προσθήκη ανώμαλων αντιπαραδειγμάτων βελτιώνει δραστικά την απόδοση του μοντέλου, με αύξηση του AUROC από 79% σε 99%, και του βέλτιστου F1 score από 62% σε 95% στην καλύτερη περίπτωση. Η μέθοδός μας, λοιπόν, φαίνεται να αποθαρρύνει με επιτυχία τη σωστή αναπαραγωγή των δειγμάτων επίθεσης, βοηθώντας το μοντέλο να τα διαχωρίζει ευκολότερα από τα δείγματα κανονικής κίνησης.

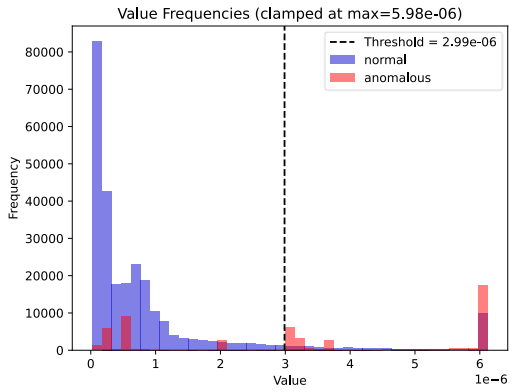
Συγκρίνοντας τις διαφορετικές τιμές $\theta \neq 0$ παρατηρούμε ότι η επιλογή τιμών από 0.0001 μέχρι και 0.1 δεν παρουσιάζει έντονη διαφοροποίηση, ενώ τιμές από 1 και μεγαλύτερες οδηγούν σε έντονη μείωση της απόδοσης. Αυτή η έλλειψη έντονης ευαισθησίας του μοντέλου στις τιμές της υπερπαραμέτρου θ παρουσιάζει ενδιαφέρον, όμως είναι λογική αν λάβει κανείς υπόψη τη δομή της όρου που αντιστοιχεί στα δείγματα επίθεσης. Πιο συγκεκριμένα, ο όρος αυτός αντιστοιχεί στον μέσο του αντίστροφου σφάλματος ανακατασκευής, και όχι τον αντίστροφο του μέσου σφάλματος ανακατασκευής. Αυτό σημαίνει πως αν σε μία παρτίδα 16 δειγμάτων ακόμα και ένα από αυτά ανακατασκευάζεται πολύ καλά, τότε ο όρος ποινικοποίησης για ολόκληρη την παρτίδα αποκτά πολύ μεγάλες τιμές, ασκώντας έντονη πίεση στο μοντέλο να αποφύγει αυτό το λάθος σε επίπεδο δείγματος. Επιπλέον, αν η ανακατασκευή είναι αρκετά ανεπαρκής τότε ο όρος αυτός λαμβάνει μικρές τιμές, επιτρέποντας στο μοντέλο να εστιάσει στη σωστή αναπαραγωγή των υπόλοιπων, ομαλών δειγμάτων. Αυτή η συμπεριφορά της συνάρτησης ποινικοποίησης χαρακτηρίζεται από έντονες μεταβολές στις τιμές, οι οποίες θα έχουν έντονο αποτέλεσμα στην εκπαίδευση του μοντέλου ακόμα και για πολύ μικρές τιμές της υπερπαραμέτρου θ. Όταν όμως η υπερπαράμετρος υπερβεί το κατώφλι $\theta = 1$, η απόδοση του μοντέλου πλήττεται. Αυτό

οφείλεται στην υπερβολική αύξηση του θ, το οποίο αναγκάζει το μοντέλο να εστιάζει τόσο έντονα στην κακή ανακατασκευή των ανώμαλων δειγμάτων, ώστε να μην μπορεί να μάθει αποτελεσματικά την κατανομή των ομαλών δειγμάτων.

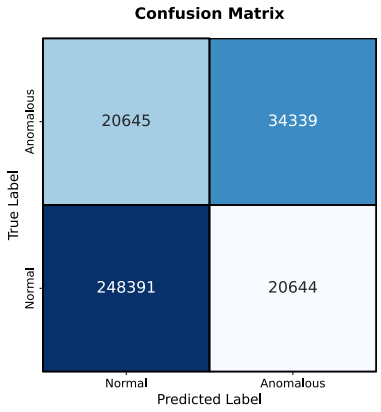
Λαμβάνοντας υπόψη τα παραπάνω, αποφασίζουμε από εδώ και στο εξής να επικεντρωθούμε στην τιμή $\theta = 0.001$ η οποία εμφανίζει ελαφρώς καλύτερες αποδόσεις από τις γειτονικές της τιμές. Παραθέτουμε τώρα αναλυτικότερες πληροφορίες για τιμές $\theta = 0$ και $\theta = 0.001$.

- $\theta = 0$:

  Παρουσιάζουμε πρώτα το διάγραμμα διαχωρισμού για τη διαμόρφωση αυτή του αυτοκωδικοποιητή. Το διάγραμμα αυτό περιέχει τα σκορ ανωμαλίας που ανατίθενται από το μοντέλο στα διαφορετικά δείγματα για την οπτικοποίηση της διακριτικής ικανότητάς του. Επίσης, παραθέτουμε τον πίνακα σύγχυσης.



Σχήμα 41: Διάγραμμα διαχωρισμού για τον AE στο CIC-IDS2018 με $\theta = 0$



Σχήμα 42: Πίνακας σύγχυσης για τον AE στο CIC-IDS2018 με $\theta = 0$

Να σημειωθεί ότι πολλά εκ των ανώμαλων δεδομένων έχουν μεγάλες τιμές σκορ ανωμαλίας, καθιστώντας την οπτικοποίησή τους δύσκολη. Για τον λόγο αυτό, το διάγραμμα φράσσεται στην διπλάσια τιμή του κατωφλίου. Οι τιμές σκορ ανωμαλίας μεγαλύτερες από την τιμή αυτή αντικαθίστανται από αυτήν και τοποθετούνται στον τελευταίο κάδο.

Επιπλέον, η τιμή των υπόλοιπων μετρικών είναι:

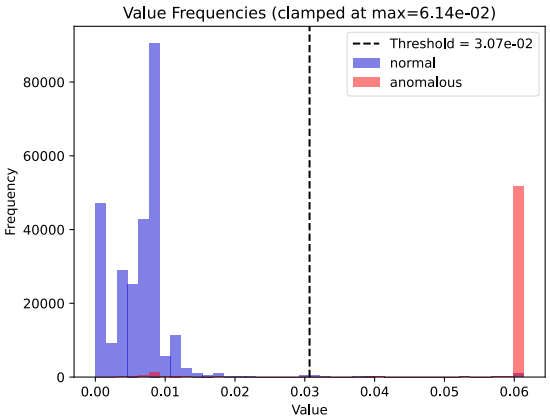
Πίνακας 25: Μετρικές του AE στο CIC-IDS2018 με $\theta = 0$

Accuracy	87.26%
Precision	62.45%
Recall	62.45%
Specificity	92.33%

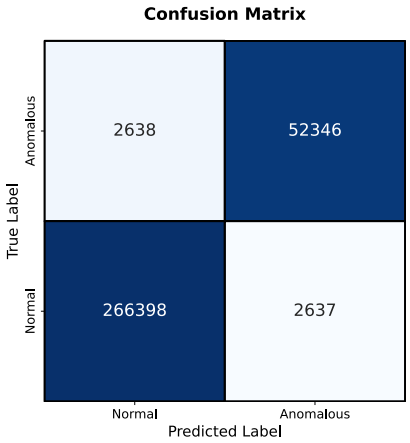
Παρατηρούμε από το διάγραμμα διαχωρισμού ότι το μοντέλο κατατάσσει σωστά τα περισσότερα ομαλά δείγματα, όμως πολλά ανώμαλα δείγματα αποκτούν χαμηλό σκορ ανωμαλίας. Έτσι εξηγείται η υψηλή τιμή του specificity και η χαμηλή τιμή recall.

- $\theta = 0.001$

Παραθέτουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 43: Διάγραμμα διαχωρισμού για τον AE στο CIC-IDS2018 με $\theta = 0.001$



Σχήμα 44: Πίνακας σύγχυσης για τον AE στο CIC-IDS2018 με $\theta = 0.001$

Επιπλέον, η τιμή των υπόλοιπων μετρικών είναι:

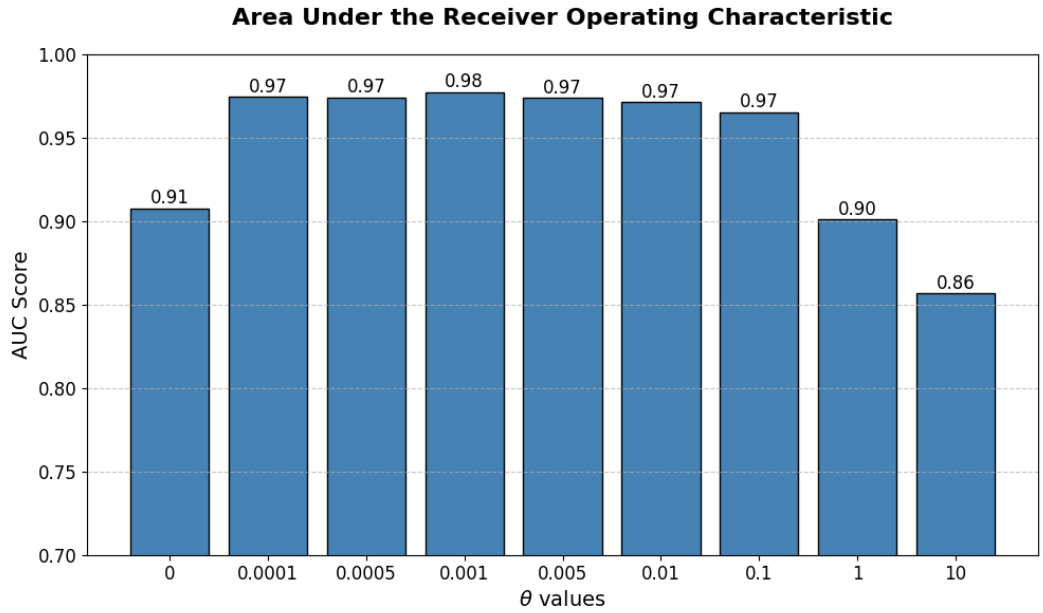
Πίνακας 26: Μετρικές του AE στο CIC-IDS2018 με $\theta = 0.001$

Accuracy	98.37%
Precision	95.20%
Recall	95.20%
Specificity	99.02%

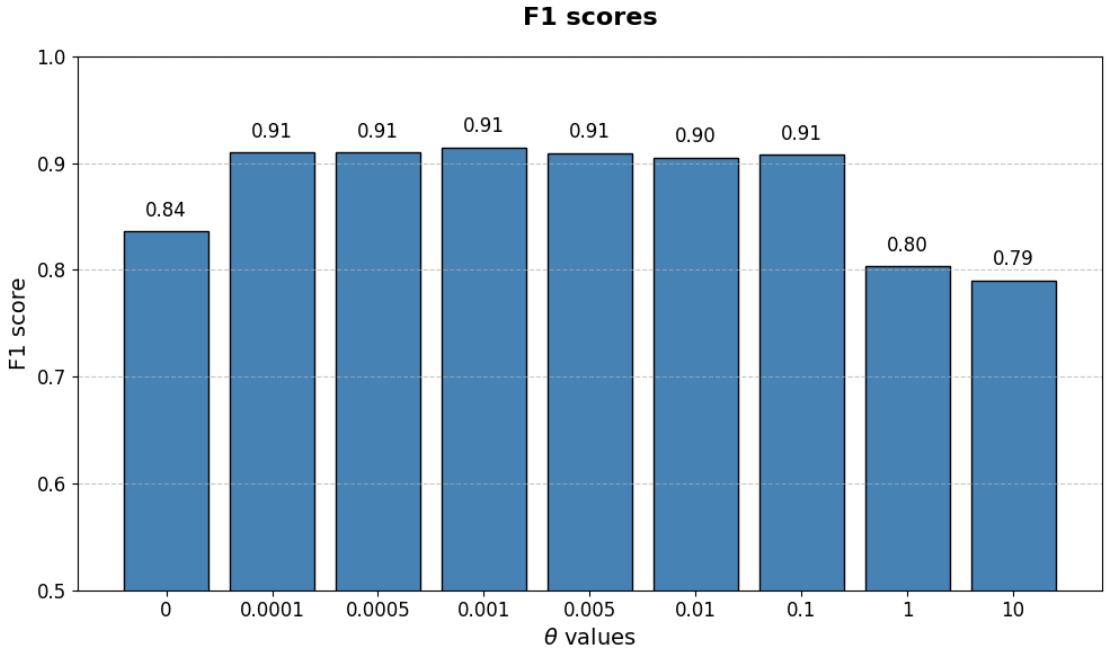
Από τα παραπάνω φαίνεται καθαρά η βελτίωση της απόδοσης. Καταρχάς, από το διάγραμμα διαχωρισμού βλέπουμε σχεδόν πλήρη διαχωρισμό των δειγμάτων, με την πλειοψηφία των δειγμάτων να απέχει αρκετά από το κατώφλι ανωμαλίας. Επιπλέον, βλέπουμε πώς όλες οι μετρικές έχουν πολύ υψηλές τιμές. Το 99% των ομαλών δειγμάτων χαρακτηρίζονται σωστά και το 95% των ανώμαλων δειγμάτων εντοπίζονται σωστά, ενώ όταν το μοντέλο αποφαίνεται πως ένα δείγμα είναι ανώμαλο, το κάνει σωστά με πιθανότητα 95%.

6.1.2. UNSW-NB15

Τα αποτελέσματα των πειραμάτων στο UNSW-NB15 παρατίθενται παρακάτω:



Σχήμα 45: Τιμές AUROC του AE στο UNSW-NB15



Σχήμα 46: Τιμές F1 score του AE στο UNSW-NB15

Καταρχάς, παρατηρούμε ότι για $\theta = 0$ ο απλός αυτοκωδικοποιητής εμφανίζει πολύ ικανοποιητικά αποτελέσματα, ιδιαίτερα αν αναλογιστεί κανείς την απλή δομή και εκπαίδευσή του, αφήνοντας όμως περιθώρια βελτίωσης. Η προσθήκη ανώμαλων αντιπαραδειγμάτων και εδώ βελτιώνει έντονα την απόδοση του μοντέλου, με αύξηση του

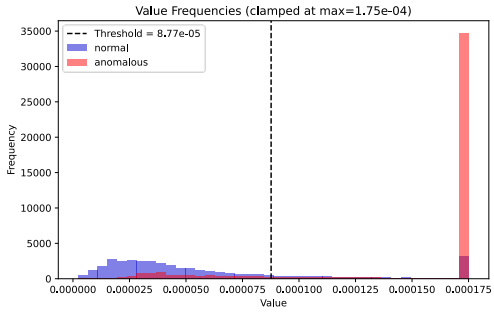
AUROC από 91% σε 98%, και του βέλτιστου F1 score από 84% σε 91%. Η μέθοδός μας, λοιπόν, αποδεικνύεται και εδώ αποτελεσματική.

Συγκρίνοντας τις διαφορετικές τιμές $\theta \neq 0$ παρατηρούμε και εδώ ότι η επιλογή τιμών από 0.0001 μέχρι και 0.1 δεν παρουσιάζει έντονη διαφοροποίηση, ενώ τιμές από 1 και μεγαλύτερες οδηγούν σε σημαντική μείωση της απόδοσης. Ο λόγος έχει ήδη αναλυθεί στην περίπτωση του CIC-IDS-2018.

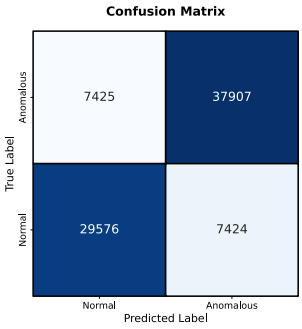
Λαμβάνοντας υπόψη τα παραπάνω, αποφασίζουμε από εδώ και στο εξής να επικεντρωθούμε στην τιμή $\theta = 0.001$ και σε αυτό το dataset, καθώς αυτή εμφανίζει ελαφρώς καλύτερες αποδόσεις από τις γειτονικές της τιμές. Παραθέτουμε τώρα αναλυτικότερες πληροφορίες για τιμές $\theta = 0$ και $\theta = 0.001$.

- $\theta = 0$:

Παρουσιάζουμε πρώτα το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 47: Διάγραμμα διαχωρισμού για τον AE στο UNSW-NB15 με $\theta = 0$



Σχήμα 48: Πίνακας σύγχυσης για τον AE στο UNSW-NB15 με $\theta = 0$

Επιπλέον, η τιμή των υπόλοιπων μετρικών είναι:

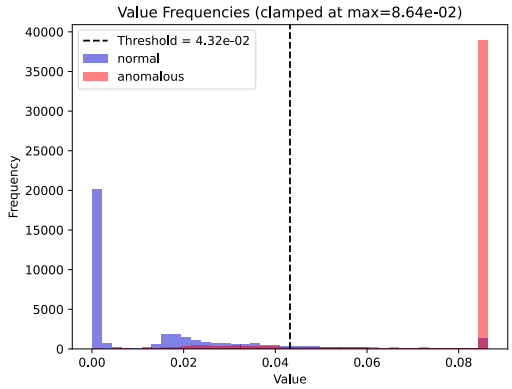
Πίνακας 27: Μετρικές του AE στο UNSW-NB15 με $\theta = 0$

Accuracy	81.96%
Precision	83.62%
Recall	83.62%
Specificity	79.94%

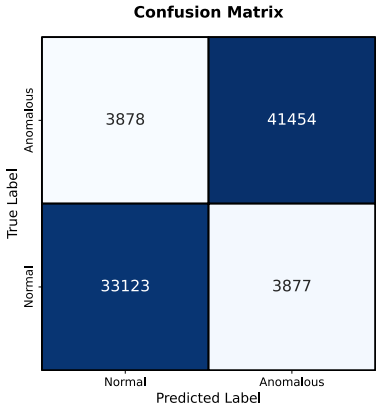
Τώρα που έχουμε στη διάθεσή μας τις πλήρεις μετρικές, μπορούμε να παρατηρήσουμε πως πράγματι το μοντέλο έχει ικανοποιητική διαχωριστική ικανότητα. Παρόλα αυτά, υπάρχουν πολλά δείγματα ομαλής κίνησης με μεγάλο σκορ ανωμαλίας, και δείγματα δικτυακών επιθέσεων με μικρό σκορ ανωμαλίας.

- $\theta = 0.001$:

Παρουσιάζουμε πρώτα το διάγραμμα διαχωρισμού:



Σχήμα 49: Διάγραμμα διαχωρισμού για τον AE στο UNSW-NB15 με $\theta = 0.001$



Σχήμα 50: Πίνακας σύγχυσης για τον AE στο UNSW-NB15 με $\theta = 0.001$

Επιπλέον, η τιμή των υπόλοιπων μετρικών είναι:

Πίνακας 28: Μετρικές του AE στο UNSW-NB15 με $\theta = 0.001$

Accuracy	90.58%
Precision	91.45%
Recall	91.45%
Specificity	89.52%

Όπως και στο CIC-IDS-2018 το διάγραμμα διαχωρισμού υποδεικνύει αρκετά βελτιωμένη διακριτική ικανότητα του μοντέλου, με την πλειοψηφία των δειγμάτων να ταξινομούνται στη σωστή κλάση και μάλιστα με μεγάλη σχετική απόσταση από το κατώφλι ανωμαλίας. Η βελτίωση αντικατοπτρίζεται και στις μετρικές, αφού οι τιμές όλων αυξάνονται έντονα. Το νέο μοντέλο πλέον μπορεί να εντοπίσει το 91.5% των ανωμαλιών, το 89.5% των ομαλών δειγμάτων, και αποφαίνεται ορθά ότι ένα δείγμα είναι ανώμαλο με πιθανότητα 91.5%.

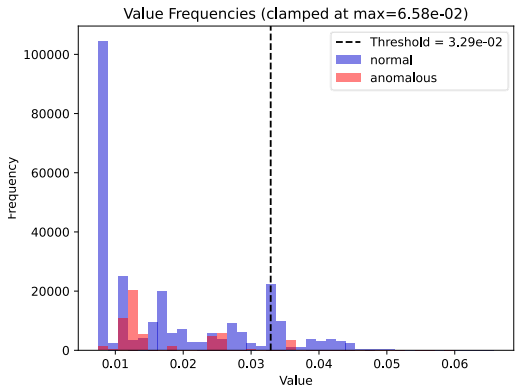
6.2. Παραλλακτικός Αυτοκωδικοποιητής

6.2.1. Με παραδοσιακή εκπαίδευση

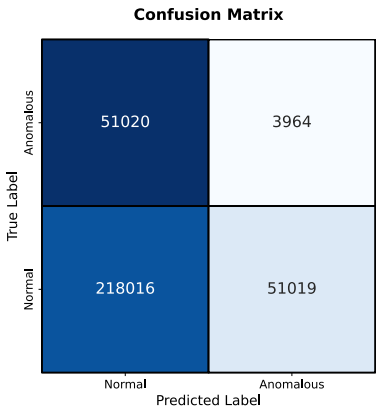
6.2.1.1. CIC-IDS-2018

- $\theta = 0$:

Παρουσιάζουμε τα διαγράμματα διαχωρισμού και πίνακα σύγχυσης:



Σχήμα 51: Διάγραμμα διαχωρισμού για τον απλό VAE στο CIC-IDS2018 για $\theta = 0$



Σχήμα 52: Πίνακας σύγχυσης για τον απλό VAE στο CIC-IDS2018 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

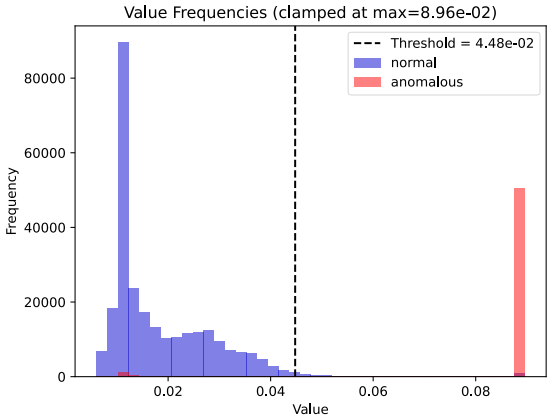
Πίνακας 29: Μετρικές του απλού VAE στο CIC-IDS2018 για $\theta = 0$

Accuracy	68.51%
Precision	7.21%
Recall	7.21%
F1 score	7.21%
Specificity	81.04%
AUROC	55.78%

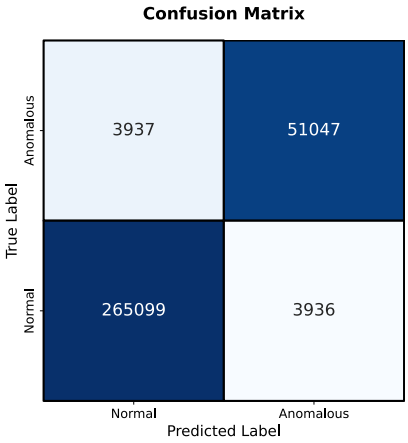
Από τα παραπάνω δεδομένα είναι εμφανές ότι το μοντέλο δεν διαθέτει πρακτικά καμία διακριτική ικανότητα για $\theta = 0$. Με την ιδανική τιμή κατωφλίου ανωμαλίας, μόλις 7% των ανώμαλων δειγμάτων εντοπίζονται και 7% των αποφάσεων ότι ένα δείγμα είναι ανώμαλο είναι σωστές. Επιπλέον, η μετρική AUROC είναι 55%, δηλαδή το μοντέλο είναι ελάχιστα καλύτερο από την τυχαία λήψη αποφάσεων.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 53: Διάγραμμα διαχωρισμού για τον απλό VAE στο CIC-IDS2018 για $\theta = 0.001$



Σχήμα 54: Πίνακας σύγχυσης για τον απλό VAE στο CIC-IDS2018 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 30: Μετρικές του απλού VAE στο CIC-IDS2018 για $\theta = 0.001$

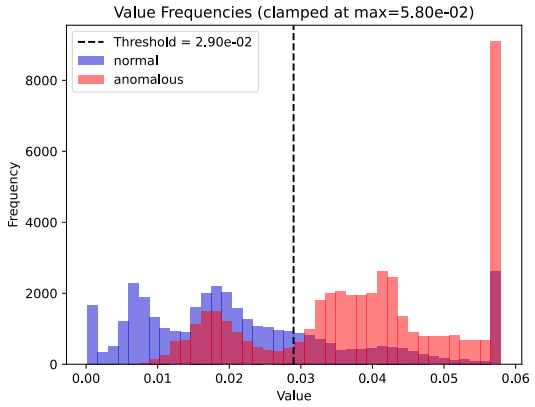
Accuracy	97.57%
Precision	92.84%
Recall	92.84%
F1 score	92.84%
Specificity	98.54%
AUROC	96.98%

Παρατηρούμε κατακόρυφη βελτίωση στα αποτελέσματα με την εισαγωγή αντιπαραδειγμάτων. Πλέον ο αυτοκωδικοποιητής μπορεί να διακρίνει αποτελεσματικά ομαλά και ανώμαλα δείγματα, με ρυθμό ανίχνευσης επιθέσεων ίσο με 93%, ρυθμό αναγνώρισης ομαλών δειγμάτων ίσο με 98.5% και 93% πιθανότητα να αποφανθεί σωστά, δεδομένου ότι έχει χαρακτηρίσει ένα δείγμα ανώμαλο.

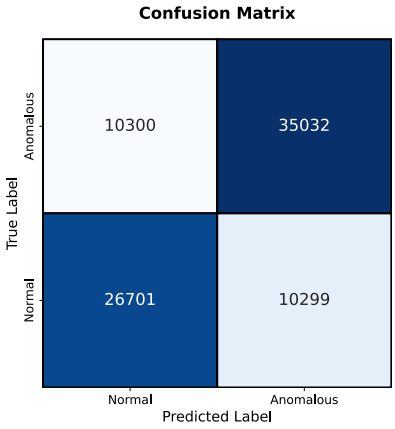
6.2.1.2. UNSW-NB15

- $\theta = 0$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 55: Διάγραμμα διαχωρισμού για τον απλό VAE στο UNSW-NB15 για $\theta = 0$



Σχήμα 56: Πίνακας σύγχυσης για τον απλό VAE στο UNSW-NB15 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

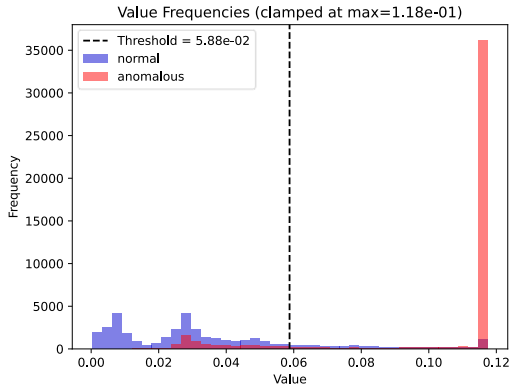
Πίνακας 31: Μετρικές του απλού VAE στο UNSW-NB15 για $\theta = 0$

Accuracy	74.98%
Precision	77.28%
Recall	77.28%
F1 score	77.28%
Specificity	72.16%
AUROC	78.22%

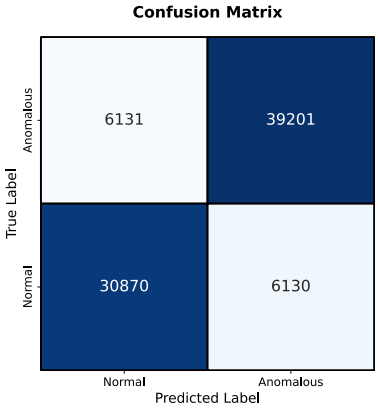
Παρατηρούμε ότι στο UNSW-NB15 ο παραλλακτικός αυτοκωδικοποιητής παρουσιάζει διακριτική ικανότητα, σε αντίθεση με το CIC-IDS-2018. Παρόλα αυτά, η διακριτική του ικανότητα είναι σχετικά ασθενής, με το 25% των δειγμάτων ελέγχου να ταξινομούνται σε λάθος κλάση.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 57: Διάγραμμα διαχωρισμού για τον απλό VAE στο UNSW-NB15 για $\theta = 0.001$



Σχήμα 58: Πίνακας σύγχυσης για τον απλό VAE στο UNSW-NB15 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 32: Μετρικές του απλού VAE στο UNSW-NB15 για $\theta = 0.001$

Accuracy	85.11%
Precision	86.48%
Recall	86.48%
F1 score	86.48%
Specificity	83.43%
AUROC	93.91%

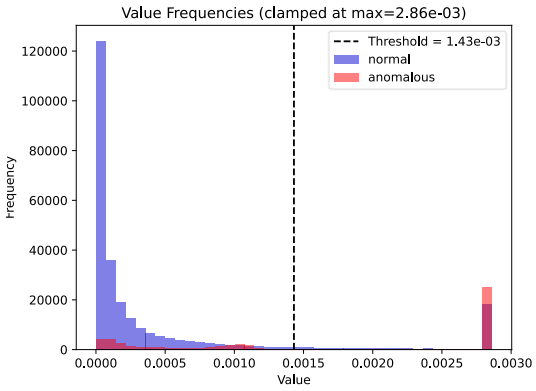
Παρατηρούμε έντονη βελτίωση στα αποτελέσματα με την εισαγωγή ανώμαλων αντιπαραδειγμάτων. Πιο συγκεκριμένα, οι τιμές όλων των μετρικών αυξάνονται κατά περίπου 10 ποσοστιαίες μονάδες, με εξαίρεση την AUROC η οποία αυξάνεται κατά 15 ποσοστιαίες μονάδες.

6.2.2. Με την τροποποιημένη μέθοδο εκπαίδευσης

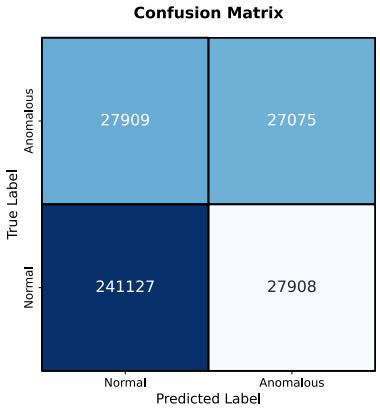
6.2.2.1. CIC-IDS-2018

- $\theta = 0$:

Παρουσιάζουμε τα διαγράμματα διαχωρισμού και πίνακα σύγχυσης:



Σχήμα 59: Διάγραμμα διαχωρισμού για τον τροποποιημένο VAE στο CIC-IDS2018 για $\theta = 0$



Σχήμα 60: Πίνακας σύγχυσης για τον τροποποιημένο VAE στο CIC-IDS2018 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

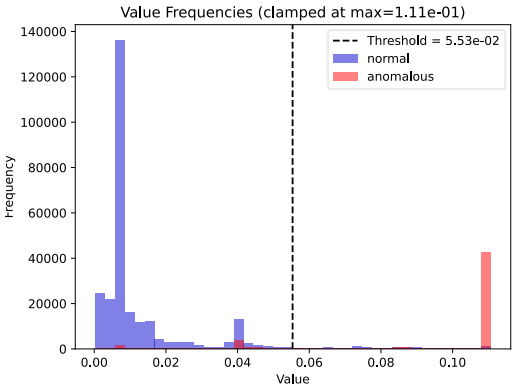
Πίνακας 33: Μετρικές του τροποποιημένου VAE στο CIC-IDS2018 για $\theta = 0$

Accuracy	82.77%
Precision	49.24%
Recall	49.24%
F1 score	49.24%
Specificity	89.63%
AUROC	82.40%

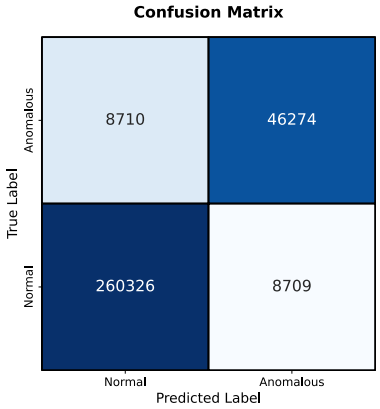
Από τα παραπάνω δεδομένα μπορούμε να συμπεράνουμε πως το μοντέλο εμφανίζει μία υποτυπώδη διακριτική ικανότητα, καθώς, παρά το γεγονός ότι ταξινομεί λανθασμένα το 50% των ανώμαλων δεδομένων, εντοπίζει σωστά το 89.5% των ομαλών δειγμάτων. Το παραπάνω σχόλιο φυσικά αναφέρεται συγκριτικά με τον απλό VAE, ο οποίος χωρίς αντιπαραδείγματα δεν εμφάνιζε καμία απολύτως διακριτική ικανότητα. Τα αποτελέσματα αυτά υπό αντικειμενική εξέταση δεν είναι ικανοποιητικά, και το μοντέλο δεν αποδίδει επαρκώς.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 61: Διάγραμμα διαχωρισμού για τον τροποποιημένο VAE στο CIC-IDS2018 για $\theta = 0.001$



Σχήμα 62: Πίνακας σύγχυσης για τον τροποποιημένο VAE στο CIC-IDS2018 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 34: Μετρικές του τροποποιημένου VAE στο CIC-IDS2018 για $\theta = 0.001$

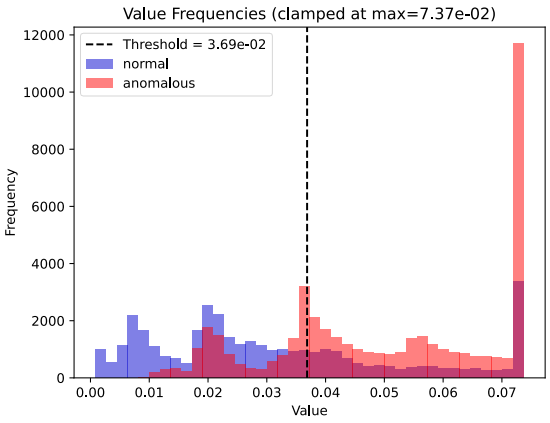
Accuracy	94.62%
Precision	84.16%
Recall	84.16%
F1 score	84.16%
Specificity	96.76%
AUROC	97.05%

Παρατηρούμε κατακόρυφη βελτίωση στα αποτελέσματα με την εισαγωγή αντιπαραδειγμάτων. Πλέον ο παραλλακτικός αυτοκωδικοποιητής μπορεί να διακρίνει αποτελεσματικά ομαλά και ανώμαλα δείγματα, με ρυθμό ανίχνευσης επιθέσεων ίσο με 84%, ρυθμό αναγνώρισης ομαλών δειγμάτων ίσο με 96.5% και 84% πιθανότητα να αποφανθεί σωστά, δεδομένου ότι έχει χαρακτηρίσει ένα δείγμα ανώμαλο. Παρατηρούμε ότι όλες οι μετρικές με εξαίρεση το AUROC είναι αρκετά υποδεέστερες από αυτές του απλού παραλλακτικού αποκωδικοποιητή. Εντούτοις, η μικρή αύξηση του AUROC υποδεικνύει ελαφρώς καλύτερη στιβαρότητα του μοντέλου στις αλλαγές του κατωφλίου ανωμαλίας. Αυτό επιβεβαιώνεται από το διάγραμμα διαχωρισμού, στο οποίο μπορούμε να παρατηρήσουμε ότι τα ομαλά δείγματα βρίσκονται κατά μέσο όρο πιο μακριά από τη διαχωριστική ευθεία.

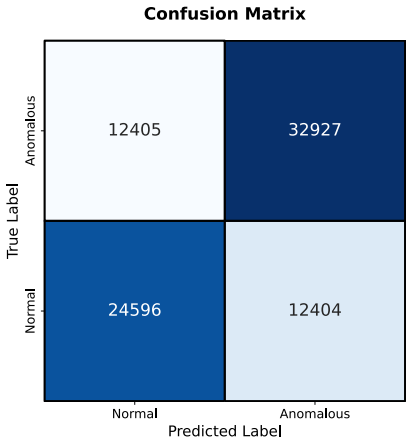
6.2.2.2. UNSW-NB15

- $\theta = 0$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 63: Διάγραμμα διαχωρισμού για τον τροποποιημένο VAE στο UNSW-NB15 για $\theta = 0$



Σχήμα 64: Πίνακας σύγχυσης για τον τροποποιημένο VAE στο UNSW-NB15 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

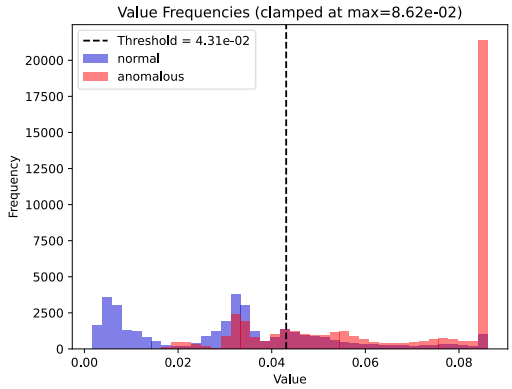
Πίνακας 35: Μετρικές του τροποποιημένου VAE στο UNSW-NB15 για $\theta = 0$

Accuracy	69.87%
Precision	72.64%
Recall	72.64%
F1 score	72.64%
Specificity	66.48%
AUROC	75.20%

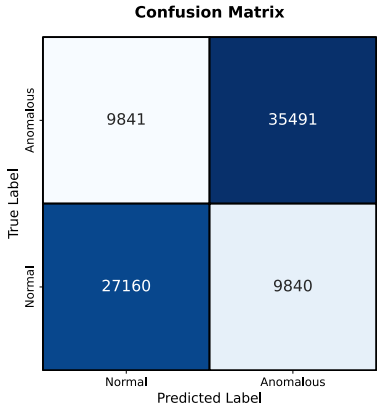
Παρατηρούμε ότι ο τροποποιημένος παραλλακτικός αυτοκωδικοποιητής εμφανίζει και εδώ διακριτική ικανότητα. Παρόλα αυτά, αυτή είναι σχετικά ασθενής, με το 30% των δειγμάτων ελέγχου να ταξινομούνται σε λάθος κλάση. Επίσης, παρατηρούμε ότι οι μετρικές Precision και Recall είναι ελαφρώς αυξημένες σε σχέση με τον απλό παραλλακτικό αυτοκωδικοποιητή, όμως οι μετρικές Specificity και AUROC είναι αρκετά μειωμένες. Συνολικά, η απόδοση είναι και εδώ υποδεέστερη.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 65: Διάγραμμα διαχωρισμού για τον τροποποιημένο VAE στο UNSW-NB15 για $\theta = 0.001$



Σχήμα 66: Πίνακας σύγχυσης για τον τροποποιημένο VAE στο UNSW-NB15 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 36: Μετρικές του τροποποιημένου VAE στο UNSW-NB15 για $\theta = 0.001$

Accuracy	76.10%
Precision	78.29%
Recall	78.29%
F1 score	78.29%
Specificity	73.41%
AUROC	86.14%

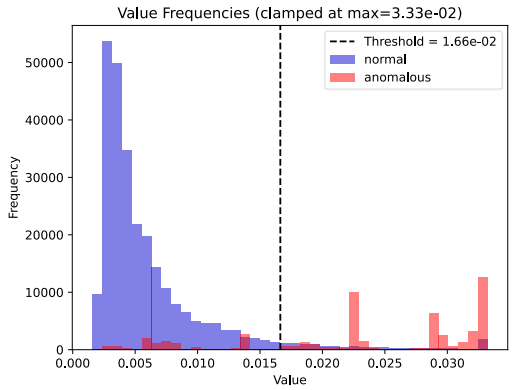
Παρατηρούμε ότι η εισαγωγή αντιπαραδειγμάτων με την μέθοδο που προτείνουμε έχει και εδώ ως αποτέλεσμα την έντονη αύξηση όλων των μετρικών, καθιστώντας το μοντέλο πιο στιβαρό στον εντοπισμό των λεπτών διαφορών μεταξύ ομαλών και ανώμαλων δειγμάτων. Εντούτοις, παρατηρούμε ότι, συγκριτικά με τον παραδοσιακό VAE με $\theta = 0.001$, η απόδοση είναι σημαντικά χειρότερη, με 8% λιγότερα δείγματα επίθεσης να εντοπίζονται, 8% λιγότερες αποφάσεις ότι ένα δείγμα αντιστοιχεί σε κακόβουλη κίνηση να είναι ορθές, και κατά 10 ποσοστιαίες μονάδες μειωμένη ικανότητα εντοπισμού ομαλών δειγμάτων.

6.3. Προσαρμογή του GANomaly

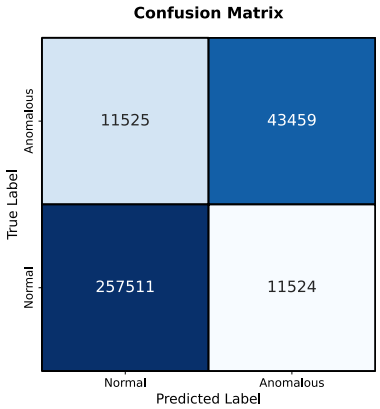
6.3.1. CIC-IDS-2018

- $\theta = 0$:

Παρουσιάζουμε τα διαγράμματα διαχωρισμού και πίνακα σύγχυσης:



Σχήμα 67: Διάγραμμα διαχωρισμού για το GANomaly_variant στο CIC-IDS2018 για $\theta = 0$



Σχήμα 68: Πίνακας σύγχυσης για το GANomaly_variant στο CIC-IDS2018 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

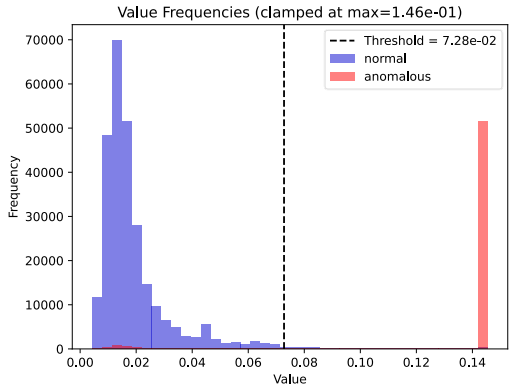
Πίνακας 37: Μετρικές του GANomaly_variant στο CIC-IDS2018 για $\theta = 0$

Accuracy	92.89%
Precision	79.04%
Recall	79.04%
F1 score	79.04%
Specificity	95.72%
AUROC	93.34%

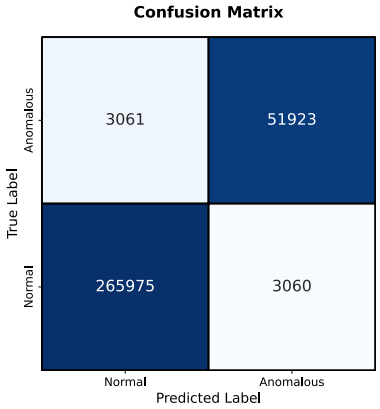
Τα παραπάνω αποτελέσματα υποδεικνύουν ισχυρή διακριτική ικανότητα του μοντέλου ακόμα και χωρίς την προσθήκη αντιπαραδειγμάτων. Μάλιστα σχεδόν το 93% των αποφάσεων είναι ορθές, με 79% ρυθμό εντοπισμό κακόβουλης κίνησης, 79% πιθανότητα επιτυχίας με δεδομένο ότι ένα δείγμα χαρακτηρίζεται ανώμαλο, και μόλις 4.5% των ομαλών δειγμάτων να ταξινομούνται λανθασμένα.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 69: Διάγραμμα διαχωρισμού για το GANomaly_variant στο CIC-IDS2018 για $\theta = 0.001$



Σχήμα 70: Πίνακας σύγχυσης για το GANomaly_variant στο CIC-IDS2018 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 38: Μετρικές του GANomaly_variant στο CIC-IDS2018 για $\theta = 0.001$

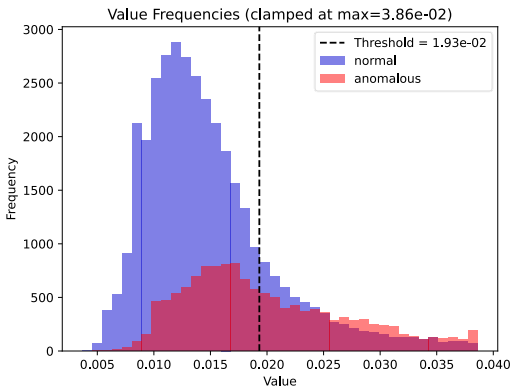
Accuracy	98.11%
Precision	94.43%
Recall	94.43%
F1 score	94.43%
Specificity	98.86%
AUROC	97.66%

Το μοντέλο εμφανώς ανταποκρίνεται πολύ θετικά στη χρήση αντιπαραδειγμάτων στην εκπαιδευτική διαδικασία, με αύξηση των μετρικών Precision, Recall, F1 score κατά 15 ποσοστιαίες μονάδες, αύξηση του Specificity κατά 3 ποσοστιαίες μονάδες, και σχεδόν υποτετραπλασιασμό των συνολικών δειγμάτων που ταξινομούνται σε λάθος κλάση. Η βελτίωση αυτή φαίνεται καθαρά και στο διάγραμμα διαχωρισμού, όπου παρατηρούμε σχεδόν πλήρη διαχωρισμό των δύο κλάσεων.

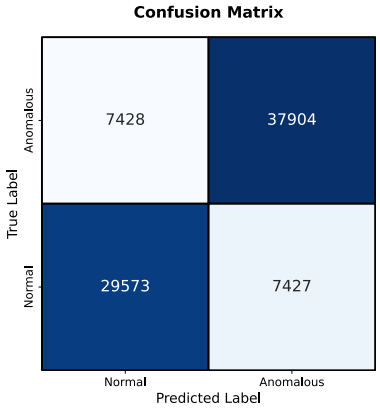
6.3.2. UNSW-NB15

- $\theta = 0$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 71: Διάγραμμα διαχωρισμού για το GANomaly_variant στο UNSW-NB15 για $\theta = 0$



Σχήμα 72: Πίνακας σύγχυσης για το GANomaly_variant στο UNSW-NB15 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

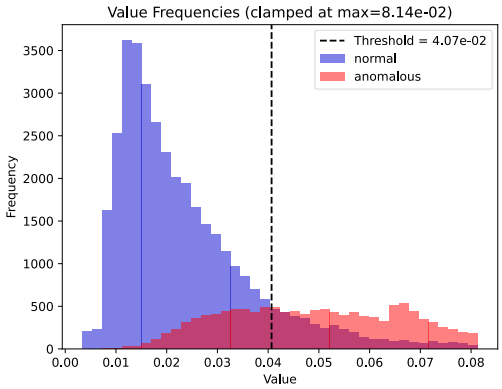
Πίνακας 39: Μετρικές του GANomaly_variant στο UNSW-NB15 για $\theta = 0$

Accuracy	81.96%
Precision	83.62%
Recall	83.61%
F1 score	83.62%
Specificity	79.93%
AUROC	90.20%

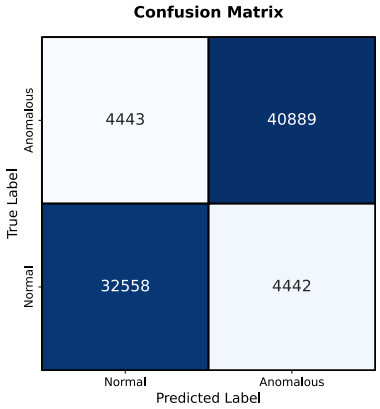
Παρατηρούμε ότι και στο UNSW-NB15 το τροποποιημένο μοντέλο GANomaly εμφανίζει ικανοποιητική διακριτική ικανότητα χωρίς τη προσθήκη αντιπαραδειγμάτων. Εντούτοις, όπως φαίνεται και από το διάγραμμα διαχωρισμού, υπάρχουν αρκετά περιθώρια βελτίωσης, καθώς οι δύο κατανομές των σκορ ανωμαλιών έχουν έντονη επικάλυψη.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 73: Διάγραμμα διαχωρισμού για το GANomaly_variant στο UNSW-NB15 για $\theta = 0.001$



Σχήμα 74: Πίνακας σύγχυσης για το GANomaly_variant στο UNSW-NB15 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 40: Μετρικές του GANomaly_variant στο UNSW-NB15 για $\theta = 0.001$

Accuracy	89.21%
Precision	90.20%
Recall	90.20%
F1 score	90.20%
Specificity	87.99%
AUROC	96.31%

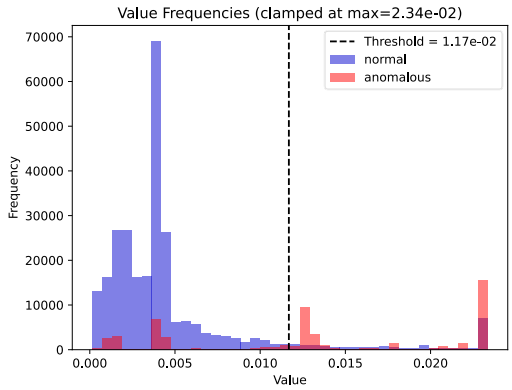
Παρατηρούμε ότι η εισαγωγή αντιπαραδειγμάτων με την μέθοδο που προτείνουμε έχει και εδώ ως αποτέλεσμα την έντονη αύξηση όλων των μετρικών, ενισχύοντας την ικανότητα του μοντέλου να διαχωρίζει τα ανώμαλα από τα ομαλά δείγματα. Η αύξηση είναι ίση με περίπου 6 ποσοστιαίες μονάδες για το AUROC, 6.5 ποσοστιαίες μονάδες για τις μετρικές Precision, Recall, F1 score, 7 για το Accuracy και 8 για το Specificity. Στα πλαίσια του συνόλου δεδομένων UNSW-NB15 πρόκειται για ικανοποιητική βελτίωση, παρά το γεγονός ότι οι τελικές μετρικές δεν είναι ιδανικές, όπως υποδεικνύεται και από την υπολογίσιμη επικάλυψη των δύο κατανομών στο διάγραμμα διαχωρισμού.

6.4. BiWGAN-GP

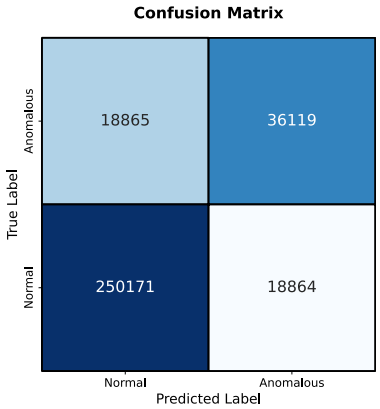
6.4.1. CIC-IDS-2018

- $\theta = 0$:

Παρουσιάζουμε τα διαγράμματα διαχωρισμού και πίνακα σύγχυσης:



Σχήμα 75: Διάγραμμα διαχωρισμού για το BiWGAN-GP στο CIC-IDS2018 για $\theta = 0$



Σχήμα 76: Πίνακας σύγχυσης για το BiWGAN-GP στο CIC-IDS2018 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 41: Μετρικές του BiWGAN-GP στο CIC-IDS2018 για $\theta = 0$

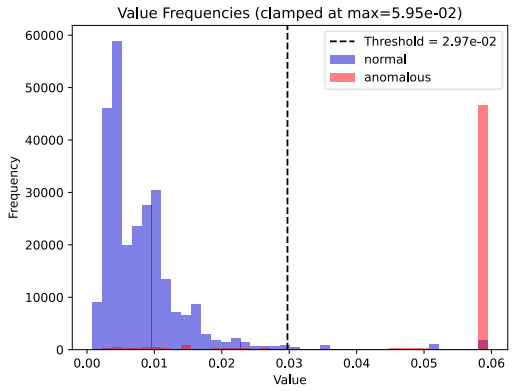
Accuracy	88.36%
Precision	65.69%
Recall	65.69%
F1 score	65.69%
Specificity	92.99%
AUROC	78.99%

Το μοντέλο BiWGAN-GP εμφανίζει αρκετά ασθενή διακριτική ικανότητα, με μόλις το 65.5% των ανώμαλων δειγμάτων να εντοπίζονται, ενώ το 34.5% των αποφάσεων ότι ένα δείγμα αντιστοιχεί σε κακόβουλη κίνηση είναι λανθασμένες. Το μοντέλο πάντως μπορεί να εντοπίσει αποτελεσματικά τα ομαλά δείγματα, με ρυθμό εντοπισμού ομαλών δειγμάτων 93%. Τα παραπάνω φαίνονται και στο διάγραμμα διαχωρισμού, όπου τα σκορ ανωμαλίας των ομαλών δειγμάτων είναι στην πλειοψηφία τους μικρότερα από το κατώφλι διαχωρισμού, όμως πολλά δείγματα που αντιστοιχούν σε κακόβουλη δικτυακή κίνηση έχουν μικρά σκορ ανωμαλίας. Οι ερευνητές δεν έχουν αξιολογήσει το μοντέλο σε αυτό το dataset, οπότε δεν υπάρχουν δεδομένα σύγκρισης. Έχοντας δοκιμάσει διαφορετικές τιμές για την υπερπαράμετρο σ, αυτά είναι τα καλύτερα αποτελέσματα που μπορέσαμε να πετύχουμε χωρίς την προσθήκη αντιπαραδειγμάτων. Δεν κρίνουμε απαραίτητο να παρουσιάσουμε διαγράμματα για όλες τις τιμές σ που δοκιμάσαμε, καθώς ο σκοπός αυτής της εργασίας είναι

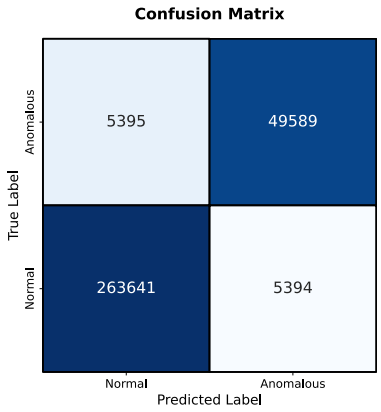
η παρατήρηση της επίδρασης της προσθήκης ανώμαλων αντιπαραδειγμάτων στη διαδικασία εκπαίδευσης με τον τρόπο που αναλύσαμε. Κρατώντας την τιμή σ σταθερή, αυτή η σύγκριση μπορεί να γίνει με ουσιώδη τρόπο, και συνεπώς ο στόχος μας δεν πλήττεται.

- $\theta = 0.001$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 77: Διάγραμμα διαχωρισμού για το BiWGAN-GP στο CIC-IDS2018 για $\theta = 0.001$



Σχήμα 78: Πίνακας σύγχυσης για το BiWGAN-GP στο CIC-IDS2018 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 42: Μετρικές του BiWGAN-GP στο CIC-IDS2018 για $\theta = 0.001$

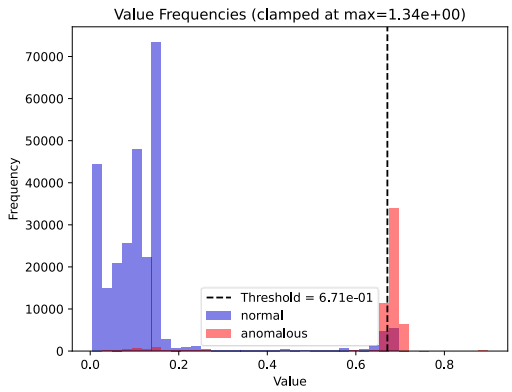
Accuracy	96.67%
Precision	90.19%
Recall	90.19%
F1 score	90.19%
Specificity	98.00%
AUROC	97.32%

Παρατηρούμε πως η χρήση αντιπαραδειγμάτων με τη μέθοδό μας οδηγεί σε κατακόρυφη αύξηση όλων των μετρικών. Πλέον, το μοντέλο συνολικά κατηγοριοποιεί λάθος ένα δείγμα με πιθανότητα μόνο 3.5%, με το 90% των δειγμάτων επίθεσης να ανιχνεύονται, και το 90% των αποφάσεων ότι ένα δείγμα είναι ανώμαλο να είναι ορθές. Όσον αφορά τα ομαλά δείγματα, το False Alarm Rate, η πιθανότητα δηλαδή ένα ομαλό δείγμα να ταξινομηθεί λάθος, είναι μόλις 2%. Το διάγραμμα διαχωρισμού ενισχύει την εικόνα που σχηματίζουν οι μετρικές, καθώς είναι φανερό πως οι δύο κατανομές είναι σαφώς διαχωρισμένες και απέχουν αρκετά από την διαχωριστική ευθεία.

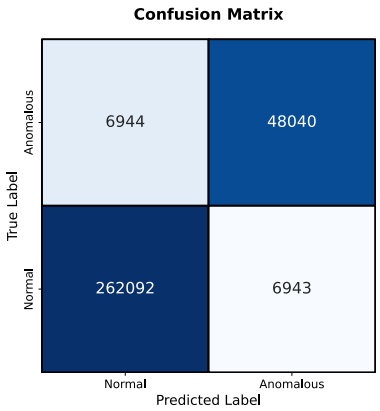
- $\theta = 0.01$:

Παρά τα θετικά αποτελέσματα της εφαρμογής του $\theta = 0.001$, ελέγχουμε και την περίπτωση αυτή, για λόγους συμμετρίας με το UNSW-NB15, του οποίου τα αποτελέσματα θα παραθέσουμε αμέσως μετά.

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 79: Διάγραμμα διαχωρισμού για το BiWGAN-GP στο CIC-IDS2018 για $\theta = 0.01$



Σχήμα 80: Πίνακας σύγχυσης για το BiWGAN-GP στο CIC-IDS2018 για $\theta = 0.01$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 43: Μετρικές του BiWGAN-GP στο CIC-IDS2018 για $\theta = 0.01$

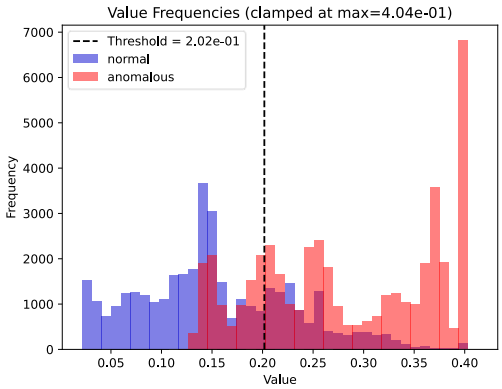
Accuracy	95.71%
Precision	87.37%
Recall	87.37%
F1 score	87.37%
Specificity	97.42%
AUROC	97.17%

Παρατηρούμε ότι η περεταίρω αύξηση του θ οδηγεί σε επιδείνωση της απόδοσης του μοντέλου. Πιο συγκεκριμένα, η νέα τιμή θ οδηγεί σε τόσο έντονη αποθάρρυνση των χαμηλών σφαλμάτων ανακατασκευής των ανώμαλων δειγμάτων, που το μοντέλο «υπεραντιδρά», ανακατασκευάζοντας κακώς και πολλά ομαλά δείγματα. Η προηγούμενη, λοιπόν, διαμόρφωση κρίνεται καταλληλότερη για το συγκεκριμένο σύνολο δεδομένων.

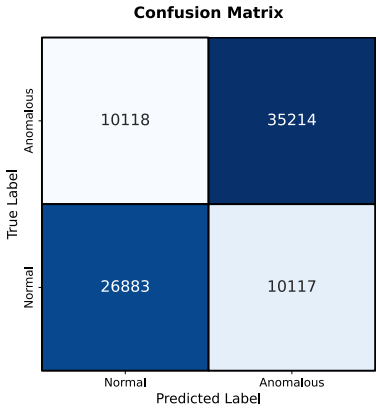
6.4.2. UNSW-NB15

- $\theta = 0$:

Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 81: Διάγραμμα διαχωρισμού για το BiWGAN-GP στο UNSW-NB15 για $\theta = 0$



Σχήμα 82: Πίνακας σύγχυσης για το BiWGAN-GP στο UNSW-NB15 για $\theta = 0$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 44: Μετρικές του BiWGAN-GP στο UNSW-NB15 για $\theta = 0$

Accuracy	75.42%
Precision	77.68%
Recall	77.68%
F1 score	77.68%
Specificity	72.66%
AUROC	84.68%

Θυμίζουμε ότι αυτό το μοντέλο είναι αναπαραγωγή του μοντέλου που προτείνουν οι Yao κ.ά. [12], και έχει εξεταστεί από αυτούς στο UNSW-NB15. Οι μετρικές απόδοσης που παρουσιάζουν οι ερευνητές είναι ελαφρώς υψηλότερες, και παρατίθενται παρακάτω:

Πίνακας 45: Μετρικές του BiWGAN-GP που αναφέρονται στο [12]

Accuracy	80.1%
Precision	81.9%
Recall	81.9%
F1 score	81.9%
Specificity	77.8%
AUROC	87.16%

όπου το specificity υπολογίζεται ως $1 - FAR$ $(false\ alarm\ rate)$.

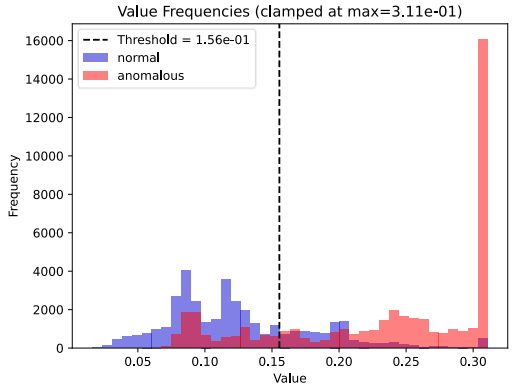
Αποδίδουμε τη διαφορά μεταξύ των μετρικών στη χρήση μόνο 20 εποχών εκπαίδευσης, ενώ οι ερευνητές εκπαιδεύουν για 200 εποχές. Οι παραπάνω επιλογή ήταν συνειδητή, καθώς δεν

μας απασχολεί η πλήρης αναπαραγωγή των αποτελεσμάτων των ερευνητών, ούτε η υπέρβαση του state-of-the-art, αλλά η επίδειξη της αξίας της μεθόδου μας για τη συμπερίληψη ανώμαλων αντιπαραδειγμάτων στην εκπαιδευτική διαδικασία.

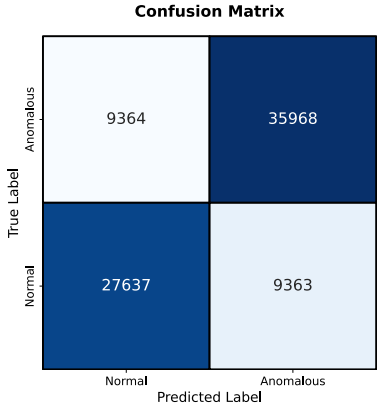
Αναφορικά με τα αποτελέσματά μας, αυτά κρίνονται ικανοποιητικά, δεδομένου ότι προκύπτουν με τη χρήση μόνο ομαλών δειγμάτων. Παρά ταύτα, όπως φαίνεται και στο διάγραμμα διαχωρισμού, υπάρχει έντονη επικάλυψη μεταξύ των κατανομών του σκορ ανωμαλίας των δύο κλάσεων.

- $\theta = 0.001$:

  Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 83: Διάγραμμα διαχωρισμού για το BiWGAN-GP στο UNSW-NB15 για $\theta = 0.001$



Σχήμα 84: Πίνακας σύγχυσης για το BiWGAN-GP στο UNSW-NB15 για $\theta = 0.001$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 46: Μετρικές του BiWGAN-GP στο UNSW-NB15 για $\theta = 0.001$

Accuracy	77.25%
Precision	79.35%
Recall	79.34%
F1 score	79.34%
Specificity	74.69%
AUROC	85.24%

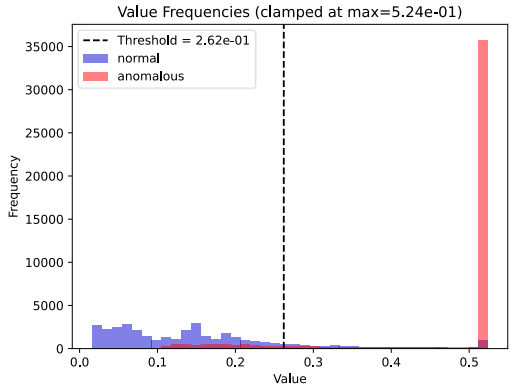
Παρατηρούμε πως αυτή τη φορά, η απόδοση δεν παρουσιάζει την έντονη βελτίωση που παρατηρήσαμε στις υπόλοιπες περιπτώσεις. Αντίθετα, οι μετρικές Accuracy, Precision, Recall και F1 score αυξήθηκαν κατά περίπου 2%, ενώ μάλιστα το Specificity και AUROC μειώθηκαν κατά 3% και 2% αντίστοιχα. Αποδίδουμε αυτή τη συμπεριφορά στον τρόπο εκπαίδευσης του BiWGAN-GP. Πιο συγκεκριμένα, σε αντίθεση με τον αυτοκωδικοποιητή, χρησιμοποιείται μία πληθώρα διαφορετικών όρων απώλειας, ζυγισμένων με διαφορετικά βάρη. Είναι πιθανό η

αντιθετική απώλεια και η απώλεια του κωδικοποιητή, μαζί με τους όρους ποινής κλίσεων, να επισκιάζουν την προσθήκη μας στην απώλεια συνέπειας κύκλου. Για να ελέγξουμε αυτήν την υπόθεση, δοκιμάζουμε την αύξηση του θ από 0.001 σε 0.01. Η νέα τιμή παρήγαγε καλά αποτελέσματα στην περίπτωση του αυτοκωδικοποιητή, καθιστώντας την καλό υποψήφιο και στην προκειμένη περίπτωση.

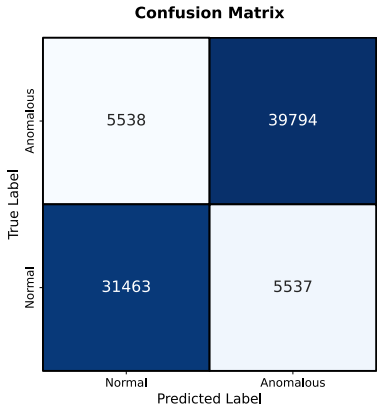
Τα παραπάνω σχόλια φυσικά ισχύουν στο πλαίσιο του συγκεκριμένου συνόλου δεδομένων, καθώς η συμπεριφορά του μοντέλου διαφέρει ανάλογα με το σύνολο που χρησιμοποιούμε. Αυτό άλλωστε παρατηρήσαμε προηγουμένως, όταν η χρήση $\theta = 0.001$ οδήγησε σε σημαντική βελτίωση της απόδοσης στο CIC-IDS-2018.

- $\theta = 0.01$:

  Παρουσιάζουμε το διάγραμμα διαχωρισμού και τον πίνακα σύγχυσης:



Σχήμα 85: Διάγραμμα διαχωρισμού για το BiWGAN-GP στο UNSW-NB15 για $\theta = 0.01$



Σχήμα 86: Πίνακας σύγχυσης για το BiWGAN-GP στο UNSW-NB15 για $\theta = 0.01$

Επιπλέον, οι τιμές των μετρικών είναι:

Πίνακας 47: Μετρικές του BiWGAN-GP στο UNSW-NB15 για $\theta = 0.01$

Accuracy	86.55%
Precision	87.79%
Recall	87.78%
F1 score	87.78%
Specificity	85.04%
AUROC	88.99%

Παρατηρούμε πως οι υποψίες μας επιβεβαιώνονται, αφού τώρα η απόδοση αυξάνεται όσο θα περιμέναμε, με αύξηση του AUROC κατά 4.5 ποσοστιαίες μονάδες, του Precision, Recall και F1 κατά 10 μονάδες, του Accuracy κατά 11 μονάδες, και του Specificity κατά 12.5 μονάδες. Η βελτίωση μπορεί να γίνει εμφανής και με απλή επισκόπηση του νέου διαγράμματος

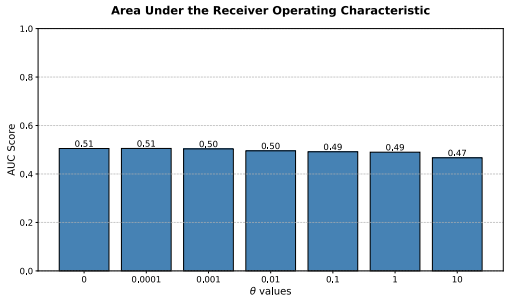
διαχωρισμού, στο οποίο οι κατανομές των σκορ ανωμαλίας των δύο κλάσεων είναι σαφώς διαχωρισμένες. Η προσθήκη, συνεπώς, των αντιπαραδειγμάτων κακόβουλης δικτυακής κίνησης με τη δική μας υλοποίηση συνεισφέρει στη δραστική βελτίωση της απόδοσης και αυτού του μοντέλου.

6.5. Συνελικτικός Αυτοκωδικοποιητής

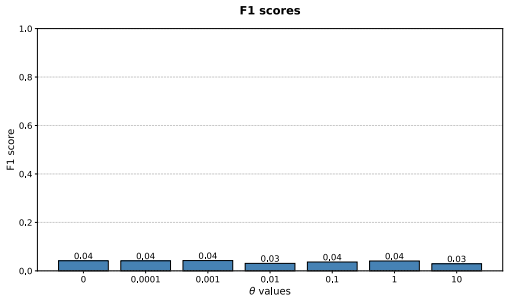
Σε αυτήν την ενότητα θα ακολουθήσουμε μία διαφορετική προσέγγιση παράθεσης των αποτελεσμάτων. Λαμβάνοντας υπόψιν το γεγονός ότι, όπως προαναφέραμε, η συγκεκριμένη αρχιτεκτονική δεν παρουσιάζει ικανοποιητική απόδοση σε καμία ρύθμιση, θα αποφύγουμε να παραθέσουμε αναλυτικά διαγράμματα διαχωρισμού και πίνακες με τις τιμές των μετρικών. Αντί αυτού, θα παρουσιάσουμε γραφήματα των μετρικών AUROC και F1 για όλες τις τιμές θ που εξετάσαμε και θα αναφέρουμε ορισμένους υποψήφιους λόγους για τους οποίους το μοντέλο αδυνατεί να εμφανίσει διακριτική ικανότητα.

6.5.1. CIC-IDS-2018

Έχουμε τα ακόλουθα διαγράμματα AUROC και F1 score συναρτήσει του θ.



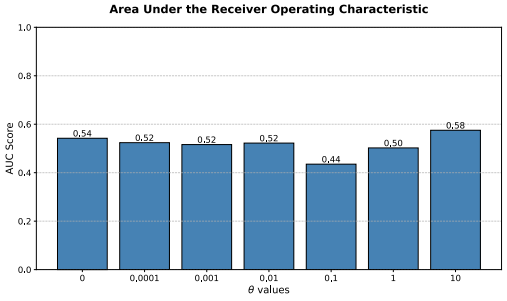
Σχήμα 87: Τιμές AUROC του ConvAE στο CIC-IDS2018



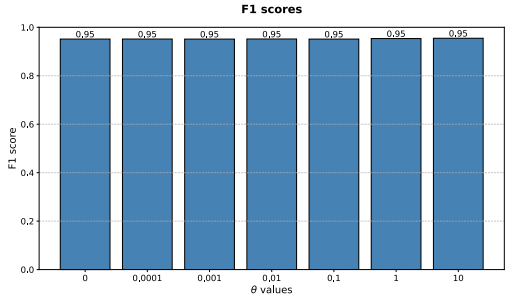
Σχήμα 88: Τιμές F1 score του ConvAE στο CIC-IDS2018

6.5.2. UNSW-NB15

Έχουμε τα ακόλουθα διαγράμματα AUROC και F1 score συναρτήσει του θ.



Σχήμα 89: Τιμές AUROC του ConvAE στο UNSW-NB15



Σχήμα 90: Τιμές F1 score του ConvAE στο UNSW-NB15

6.5.3. Σχολιασμός των αποτελεσμάτων

Από τα διαγράμματα που παρουσιάσαμε παραπάνω είναι φανερό πως το μοντέλο συνελικτικού αυτοκωδικοποιητή δεν διαθέτει απολύτως καμία διακριτική ικανότητα. Αναλυτικότερα, στο CIC-IDS2018, με το καλύτερο κατώφλι διαχωρισμού, έχει F1 score στην περιοχή 0.03 – 0.04, και η περιοχή κάτω από την καμπύλη ROC είναι πρακτικά 0.5 για όλες τις διαμορφώσεις, ενώ στο UNSW-NB15 οι τιμές αυτές είναι πρακτικά 0.95 και 0.5 για όλες τις διαμορφώσεις. Η πολύ υψηλή τιμή F1 score στο UNSW-NB15 δεν είναι θετική ένδειξη, καθώς οφείλεται απλώς στο γεγονός ότι στο συγκεκριμένο σύνολο δεδομένων αξιολόγησης η πλειοψηφία των δειγμάτων είναι ανώμαλα. Η πραγματική εικόνα δίνεται από τη μετρική AUROC που υποδεικνύει ισοδυναμία του μοντέλου με τυχαία ταξινόμηση, όπως είχαμε προαναφέρει και στη θεωρητική ενότητα. Σημειώνουμε ότι παρά την έντονη πόλωση των συνόλων αξιολόγησης προς μία κλάση, τα σύνολα εκπαίδευσης είναι ισορροπημένα, οπότε το μοντέλο έχει εκτεθεί σε επαρκή πληροφορία για την απόκτηση διακριτικής ικανότητας. Τα αποτελέσματα αυτά, λοιπόν, δεν γίνεται να παραχθούν από ένα μοντέλο με διακριτική ικανότητα.

Υπάρχουν αρκετοί υποψήφιοι λόγοι για τους οποίους η αρχιτεκτονική αυτή αποδεικνύεται ανεπαρκής για την επίλυση του προβλήματος του εντοπισμού δικτυακών ανωμαλιών.

Καταρχάς, η χρήση της συσχέτισης ενδέχεται προκαλεί μεγάλη απώλεια πληροφορίας. Αρκεί και μόνο να αναλογιστεί κανείς ότι η συσχέτιση Pearson απεικονίζει τις εισόδους σε δείγματα στο διάστημα $[-1,1]$ σε κάθε περίπτωση. Αυτό αναιρεί ένα σημαντικό πλεονέκτημα που έχει κάθε άλλη μέθοδος που εφαρμόσαμε, που προκύπτει από την κανονικοποίηση min-max του σταδίου της προεπεξεργασίας του συνόλου δεδομένων. Αυτό το πλεονέκτημα έγκειται στο γεγονός ότι τα μέγιστα και ελάχιστα που υπολογίζονται λαμβάνονται από το σύνολο εκπαίδευσης. Συνεπώς, στο σύνολο ελέγχου πολλά ανώμαλα δείγματα εμφανίζουν σε ορισμένα χαρακτηριστικά μεγαλύτερες τιμές από το μέγιστο που υπολογίστηκε στο σύνολο εκπαίδευσης. Τα ανώμαλα δείγματα αυτά θα έχουν μεγαλύτερες τιμές από τη μονάδα κατά την είσοδό τους στο μοντέλο, οδηγώντας οργανικά στην χειρότερη ανακατασκευή τους.

Επιπλέον, η χρήση του πίνακα συσχέτισης Pearson εστιάζει αποκλειστικά στη συσχέτιση μεταξύ διαφορετικών δειγμάτων, αγνοώντας ενδεχόμενες ανωμαλίες που μπορεί να παρουσιαστούν στα στοιχεία ενός μεμονωμένου ανώμαλου δείγματος. Αυτό οδηγεί σε μεγάλη απώλεια πολύτιμης πληροφορίας που μπορεί να χρησιμοποιηθεί κατά την λήψη της απόφασης για το είδος ενός δείγματος.

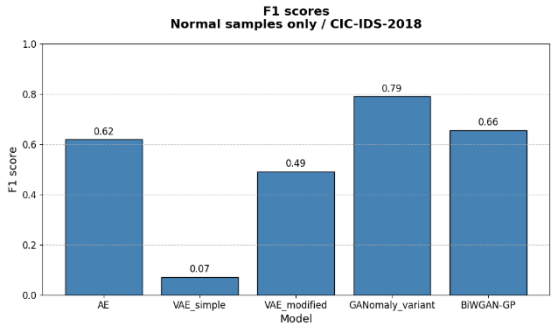
Τέλος, είναι πιθανό το μοντέλο συνελικτικού αυτοκωδικοποιητή να μην παρουσιάζει την απαραίτητη ευαισθησία σε αλλαγές κρίσιμων στοιχείων του πίνακα εισόδου. Εάν για παράδειγμα μία ανωμαλία εκφραστεί με έντονη αλλαγή σε ορισμένα μόνο στοιχεία του πίνακα εισόδου, τότε ενδέχεται αυτό να μην αποτελέσει αρκετά ισχυρό ερέθισμα ώστε ο αυτοκωδικοποιητής να αλλάξει σημαντικά την έξοδό του, με αποτέλεσμα η ανωμαλία να μην εντοπιστεί.

Να σημειώσουμε ότι αποκλείουμε το ενδεχόμενο η ανάλυση κύριων συνιστωσών (PCA) να φέρει σημαντικό μέρος της ευθύνης, αφαιρώντας την από το pipeline και ελέγχοντας για αλλαγή των αποτελεσμάτων. Τα αποτελέσματα δεν αλλάζουν, οπότε δεν κρίνουμε απαραίτητη την παράθεσή τους.

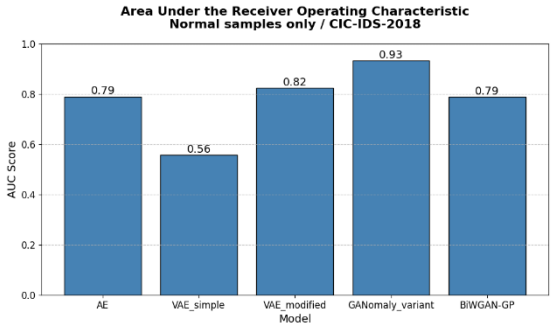
6.6. Συγκριτικός σχολιασμός αποτελεσμάτων

Συγκεντρώνουμε σε γραφήματα για όλα τα μοντέλα εκτός του συνελικτικού αυτοκωδικοποιητή τις δύο πιο σημαντικές μετρικές, F1 score και AUROC, και παρουσιάζουμε τα αποτελέσματα παρακάτω:

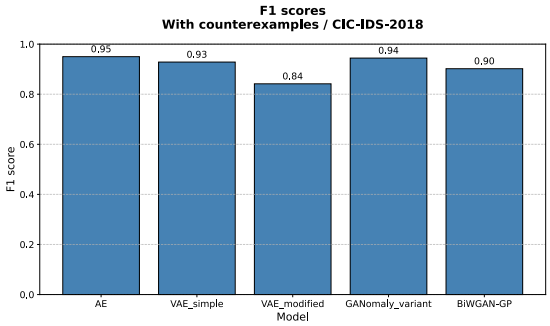
- CIC-IDS-2018



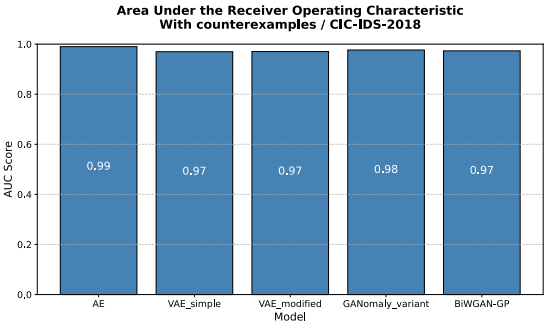
Σχήμα 91: Σύγκριση F1 score των μοντέλων στο CIC-IDS2018 για εκπαίδευση σε ομαλά δείγματα



Σχήμα 92: Σύγκριση AUROC των μοντέλων στο CIC-IDS2018 για εκπαίδευση σε ομαλά δείγματα

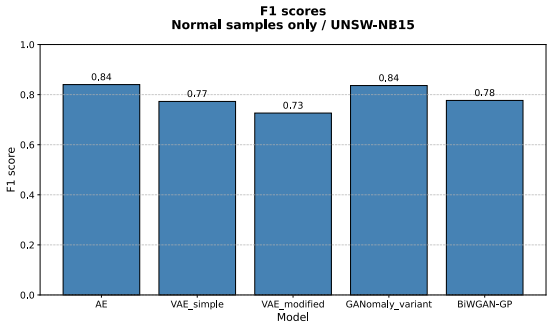


Σχήμα 93: Σύγκριση F1 score των μοντέλων στο CIC-IDS2018 για εκπαίδευση σε μεικτά δείγματα

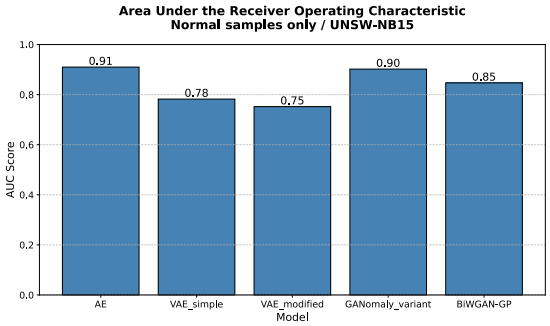


Σχήμα 94: Σύγκριση AUROC των μοντέλων στο CIC-IDS2018 για εκπαίδευση σε μεικτά δείγματα

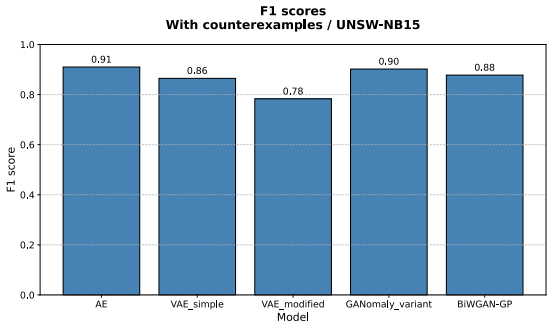
- UNSW-NB15



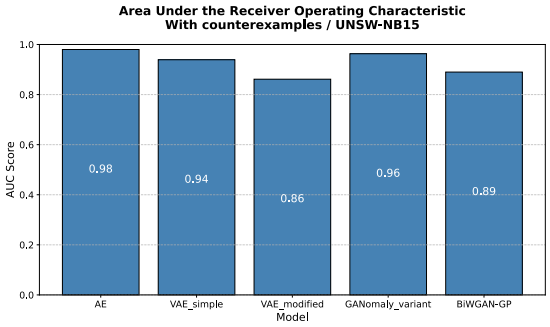
Σχήμα 95: Σύγκριση F1 score των μοντέλων στο UNSW-NB15 για εκπαίδευση σε ομαλά δείγματα



Σχήμα 96: Σύγκριση AUROC των μοντέλων στο UNSW-NB15 για εκπαίδευση σε ομαλά δείγματα



Σχήμα 97: Σύγκριση F1 score των μοντέλων στο UNSW-NB15 για εκπαίδευση σε μεικτά δείγματα



Σχήμα 98: Σύγκριση AUROC των μοντέλων στο UNSW-NB15 για εκπαίδευση σε μεικτά δείγματα

Εξαιρούμε από την ανάλυση που ακολουθεί την αρχιτεκτονική του συνελικτικού αυτοκωδικοποιητή, η οποία δεν πέτυχε ικανοποιητικά αποτελέσματα για τους λόγους που σχολιάσαμε στην αντίστοιχη ενότητα.

Ο παρόμοιος χειρισμός των μοντέλων και η διατήρηση των ίδιων υπερπαραμέτρων εκπαίδευσης, όπου αυτό ήταν δυνατό, μας επιτρέπει να συγκρίνουμε τις διαφορετικές προσεγγίσεις ως προς την απόδοσή τους στα δύο σύνολα δεδομένων.

Εστιάζοντας στην επίδοση των μοντέλων μετά την εκπαίδευση πάνω σε αμιγώς ομαλά δεδομένα, παρατηρούμε ότι η προσαρμογή του GANomaly πετυχαίνει την πιο στιβαρή συμπεριφορά, αφού έχει με διαφορά το καλύτερο F1 score και AUROC στο CIC-IDS-2018, ενώ στο UNSW-NB15 έχει καλύτερη απόδοση από όλα τα μοντέλα πέρα από τον αυτοκωδικοποιητή, οι μετρικές του οποίου είναι ελάχιστα καλύτερες. Ο απλός αυτοκωδικοποιητής εμφανίζει τις πιο υψηλές επιδόσεις με εκπαίδευση αμιγώς πάνω σε ομαλά δείγματα στο UNSW-NB15, όμως στο CIC-IDS-2018 είναι πρακτικά το τρίτο καλύτερο μοντέλο, μετά το BiWGAN-GP (αν ληφθούν υπόψιν συνδυαστικά τα F1 scores και AUROC). Το BiWGAN-GP εμφανίζει τη δεύτερη καλύτερη επίδοση στο CIC-IDS-2018 και την τρίτη καλύτερη στο UNSW-NB15. Αξίζει να θυμίσουμε σε αυτό το σημείο ότι το μοντέλο GANomaly σχεδιάστηκε αρχικά για εντοπισμό ανωμαλιών σε εικόνες [11] και προσαρμόστηκε από εμάς για την ανίχνευση δικτυακών εισβολών χωρίς ιδιαίτερη σχολαστικότητα στη ρύθμιση των

υπερπαραμέτρων, αφού ενδιαφερόμαστε κυρίως για την επίδραση της προσθήκης αντιπαραδειγμάτων. Παρά ταύτα, η δική μας υλοποίηση του τροποποιημένου GANomaly με εκπαίδευση πάνω σε αμιγώς ομαλά δεδομένα για μόλις 20 εποχές ξεπερνά με υπολογίσιμη διαφορά τις μετρικές του BiWGAN-GP που επιτυγχάνονται στο [12] σε 200 εποχές, ενώ στα δικά μας πειράματα οι διαφορές των δύο μοντέλων είναι συστηματικά έντονες σε όλες τις μετρικές. Τέλος, ο παραλλακτικός αυτοκωδικοποιητής εμφανίζει τη χειρότερη απόδοση (στο CIC-IDS-2018 έχει σχετικά υψηλή τιμή AUROC, αλλά πολύ χαμηλή τιμή F1 score), ενώ τα δεδομένα δεν αναδεικνύουν κάποια από τις δύο ρυθμίσεις του (απλή / τροποποιημένη συνάρτηση απώλειας) ως ανώτερη. Στο CIC-IDS-2018 κυριαρχεί η τροποποιημένη έκδοση του VAE, η οποία όμως εμφανίζει κατώτερη επίδοση στο UNSW-NB15.

Με την εισαγωγή των αντιπαραδειγμάτων στη διαδικασία εκπαίδευσης παρατηρούμε ότι η αξιοποίηση της νέας πληροφορίας γίνεται σε διαφορετικό βαθμό από κάθε μοντέλο. Τη μεγαλύτερη βελτίωση βλέπει ο απλός αυτοκωδικοποιητής, ο οποίος αναδεικνύεται πρώτος σε όλα τα datasets, με F1 score 95% και AUROC 99% στο CIC-IDS-2018, και 91% και 98% αντίστοιχα στο UNSW-NB15. Δεύτερη αρχιτεκτονική με μικρή διαφορά είναι το τροποποιημένο GANomaly, με τιμές F1 score και AUROC ίσες με 94.5% και 97.5% στο CIC-IDS-2018 και 90% και 96% στο UNSW-NB15. Ακολουθεί ο απλός παραλλακτικός αυτοκωδικοποιητής, το BiWGAN-GP και ο τροποποιημένος παραλλακτικός αποκωδικοποιητής. Από τα αποτελέσματα αυτά γίνεται σαφές ότι η εκπαίδευση μοντέλων με δομή αυτοκωδικοποιητή (είτε φανερά, όπως ο απλός και παραλλακτικός αυτοκωδικοποιητής, είτε κρυμμένα, όπως οι παράγωγες από τα BiGAN αρχιτεκτονικές) επωφελείται σε όλες τις περιπτώσεις από τη συμπερίληψη δειγμάτων ανωμαλιών ως αντιπαραδείγματα, των οποίων ο μέσος αντίστροφος του σφάλματος ανακατασκευής ενθαρρύνεται να ελαχιστοποιηθεί μαζί με το μέσο σφάλμα ανακατασκευής των ομαλών δειγμάτων.

7. Συμπεράσματα & Μελλοντικές κατευθύνσεις

Σε αυτό το έργο παρουσιάζουμε μία σχετικά παραμελημένη από τη βιβλιογραφία μέθοδο για την ενίσχυση της διαδικασίας εκπαίδευσης μοντέλων βαθιάς μάθησης σε ημιεπιβλεπόμενη ρύθμιση. Η μέθοδος αυτή εισήχθη για πρώτη φορά από τους Ruff κ.ά. [10] για την ενίσχυση της αντιληπτικότητας της αρχιτεκτονικής μοντέλου Deep SVDD στον τομέα της αναγνώρισης εικόνων, και λειτουργεί εισάγοντας έναν αντίστροφο όρο απώλειας για τα ανώμαλα δείγματα, ενθαρρύνοντας το μοντέλο όχι μόνο να μειώσει την απώλεια για τα ομαλά δείγματα, αλλά να την αυξήσει για τα ανώμαλα. Σύμφωνα με την έρευνά μας δεν υπάρχει εφαρμογή της μεθόδου αυτής στον τομέα του εντοπισμού ανωμαλιών σε δίκτυα. Η συνεισφορά μας συνίσταται στην κατάλληλη τροποποίηση αυτής της μεθόδου, ώστε να εφαρμοστεί σε πιο δημοφιλείς αρχιτεκτονικές στον τομέα της ανίχνευσης εισβολών σε δίκτυα, όπως οι αυτοκωδικοποιητές και τα παραγωγικά αντιπαραθετικά δίκτυα (GAN) τα οποία υλοποιούνται με χρήση δικτύου κωδικοποίησης (encoder). Στόχος μας είναι να ενθαρρύνουμε την κακή ανακατασκευή ανώμαλων δειγμάτων από τους αυτοκωδικοποιητές που βρίσκονται εντός κάθε αρχιτεκτονικής. Υλοποιούμε τέσσερις αρχιτεκτονικές: μία απλού αυτοκωδικοποιητή, μία παραλλακτικού αυτοκωδικοποιητή με δύο ρυθμίσεις απώλειας, μία τροποποίηση του δικτύου GANomaly [11], και μία αναπαραγωγή με μικρές αλλαγές του μοντέλου του [12], στο οποίο αναφερόμαστε ως BiWGAN-GP. Όλες οι αρχιτεκτονικές υλοποιούνται με παραπλήσια εσωτερική δομή (αριθμός επιπέδων και πλήθος μονάδων στα επίπεδα, διάσταση του λανθάνοντα χώρου), ώστε να είναι δυνατή η μεταξύ τους σύγκριση. Επιπλέον, εξετάζουμε τη μέθοδο των αντιπαραδειγμάτων σε μία νέα αρχιτεκτονική, η οποία βασίζεται στον υπολογισμό της συσχέτισης ενός παραθύρου δειγμάτων και τον εντοπισμό των δικτυακών επιθέσεων μέσω μίας αρχιτεκτονικής δισδιάστατου συνελικτικού αυτοκωδικοποιητή. Αυτή η αρχιτεκτονική αποδεικνύεται ανεπαρκής και αναφέρεται για λόγους πληρότητας.

Για την εκπαίδευση και έλεγχο χρησιμοποιούμε τα datasets CIC-IDS-2018 [13] και UNSW-NB15 [14, 15, 16, 17, 18], τα οποία αποτελούν αντιπροσωπευτικά σύνολα δεδομένων δικτυακής κίνησης. Διαπιστώνουμε ότι η εφαρμογή της προσθήκης μας οδηγεί σε δραστική βελτίωση των επιδόσεων σε όλες τις αρχιτεκτονικές. Ανώτερη επίδοση που παρατηρούμε είναι αυτή του απλού αυτοκωδικοποιητή, με F1 score 95% και AUROC 99% στο CIC-IDS-2018, και 91% και 98% αντίστοιχα στο UNSW-NB15. Η επίτευξη αυτών των επιδόσεων με τη σχετικά απλή δομή του αυτοκωδικοποιητή αναδεικνύει την αξία της μεθόδου που αναλύουμε, και αποδεικνύει ότι αυτή βοηθά αποτελεσματικά τα μοντέλα να εστιάζουν προσαρμοστικά στα ανώμαλα δείγματα που ανακατασκευάζονται σωστά, θεραπεύοντας τυχόν «τυφλά σημεία» που εμφανίζονται κατά την εκπαίδευση.

Υπογραμμίζουμε τώρα μερικές μελλοντικές κατευθύνσεις που πρέπει να ακολουθηθούν. Καταρχάς, δεδομένου ότι ο στόχος της διπλωματικής αυτής είναι η ανάδειξη της αξίας αυτής της μεθόδου ημιεπιβλεπόμενης μάθησης, η ανάλυσή μας χαρακτηρίζεται από ευρύτητα και όχι βάθος. Εστιάσαμε, δηλαδή, στην εξέταση πολλαπλών διαφορετικών αρχιτεκτονικών,

χωρίς όμως να βελτιστοποιήσουμε τις τιμές όλων των υπερπαραμέτρων τους. Προτείνουμε, συνεπώς, σε μελλοντικό έργο να εξερευνηθεί στοχευμένα μία τέτοια αρχιτεκτονική, με σχολαστική εκτέλεση finetuning και εκπαίδευση για περισσότερες εποχές με χρήση της προσθήκης που παρουσιάζουμε, για τον ανταγωνισμό του state-of-the-art. Επιπλέον, η χρήση του αντίστροφου σφάλματος ανακατασκευής για τους αυτοκωδικοποιητές είναι μία από τις μεθόδους που μπορούν να χρησιμοποιηθούν για τη μάθηση της κατανομής των ανώμαλων αντιπαραδειγμάτων με μη περιοριστικό τρόπο. Είναι επιτακτική ανάγκη να εξερευνηθούν και άλλοι τρόποι αποθάρρυνσης της σωστής ανακατασκευής δειγμάτων επίθεσης, ώστε να διαμορφωθεί μία ολοκληρωμένη εικόνα των μεθόδων που μπορεί κανείς να εφαρμόσει για την ενσωμάτωση αυτής της πολύτιμης επιπρόσθετης πληροφορίας στην εκπαιδευτική διαδικασία.

Βιβλιογραφία

[1] Y. Luo, Y. Ma και Z. Yang, «Multi-resolution auto-encoder for anomaly detection of retinal imaging,» *Physical and Engineering Sciences in Medicine,* τόμ. 47, pp. 517-529, 01 June 2024.

[2] D. Huang, D. Mu, L. Yang και X. Cai, «CoDetect: Financial Fraud Detection With Anomaly Feature Detection,» *IEEE Access,* τόμ. 6, pp. 19161-19174, 2018.

[3] V. Toufigh και I. Ranjbar, «Unsupervised deep learning framework for ultrasonic-based distributed damage detection in concrete: integration of a deep auto-encoder and Isolation Forest for anomaly detection,» *Structural Health Monitoring,* τόμ. 23, pp. 1313-1333, 2024.

[4] F. A. Bin Hamid Ali και Y. Y. Len, «Development of host based intrusion detection system for log files,» σε *2011 IEEE Symposium on Business, Engineering and Industrial Applications (ISBEIA)*, 2011.

[5] Y. Yang, K. Zheng, B. Wu, Y. Yang και X. Wang, «Network Intrusion Detection Based on Supervised Adversarial Variational Auto-Encoder With Regularization,» *IEEE Access,* τόμ. 8, pp. 42169-42184, 2020.

[6] P. Ioulianou, V. Vasilakis, I. Moscholios και M. Logothetis, «A signature-based intrusion detection system for the internet of things,» *Information and Communication Technology Form,* 2018.

[7] M. Said Elsayed, N.-A. Le-Khac, S. Dev και A. D. Jurcut, «Network Anomaly Detection Using LSTM Based Autoencoder,» σε *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, New York, NY, USA, 2020.

[8] Z. Yang, X. Liu, T. Li, D. Wu, J. Wang, Y. Zhao και H. Han, «A systematic literature review of methods and datasets for anomaly-based network intrusion detection,» *Computers & Security,* τόμ. 116, p. 102675, 2022.

[9] S. Zavrak και M. İskefiyeli, «Anomaly-Based Intrusion Detection From Network Flow Features Using Variational Autoencoder,» *IEEE Access,* τόμ. 8, pp. 108346-108358, 2020.

[10] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller και M. Kloft, *Deep Semi-Supervised Anomaly Detection,* 2020.

[11] S. Akcay, A. Atapour-Abarghouei και T. P. Breckon, «GANomaly: Semi-supervised Anomaly Detection via Adversarial Training,» σε *Computer Vision – ACCV 2018*, Cham, 2019.

[12] W. Yao, H. Shi και H. Zhao, «Scalable anomaly-based intrusion detection for secure Internet of Things using generative adversarial networks in fog environment,» *Journal of Network and Computer Applications,* τόμ. 214, p. 103622, 2023.

[13] Canadian Institute for Cybersecurity, *CSE-CIC-IDS2018 Dataset,* 2018.

[14] N. Moustafa και J. Slay, «UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),» σε *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015.

[15] N. Moustafa και J. Slay, «The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set,» *Information Security Journal: A Global Perspective,* τόμ. 25, p. 18–31, 2016.

[16] N. Moustafa, J. Slay και G. Creech, «Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation on Large-Scale Networks,» *IEEE Transactions on Big Data,* τόμ. 5, pp. 481-494, December 2019.

[17] N. Moustafa, G. Creech και J. Slay, «Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models,» σε *Data Analytics and Decision Support for Cybersecurity: Trends, Methodologies and Applications*, I. Palomares Carrascosa, H. K. Kalutarage και Y. Huang, Επιμ., Cham, Springer International Publishing, 2017, p. 127–156.

[18] M. Sarhan, S. Layeghy, N. Moustafa και M. Portmann, «NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems,» σε *Big Data Technologies and Applications*, Springer International Publishing, 2021, p. 117–135.

[19] G. Kakkavas, N. Fryganiotis, V. Karyotis και S. Papavassiliou, «Generative Deep Learning Techniques for Traffic Matrix Estimation From Link Load Measurements,» *IEEE Open Journal of the Communications Society,* 2024.

[20] G. Kakkavas, P. Maratos, V. Karyotis και S. Papavassiliou, «Traffic Matrix Estimation Using Invertible Neural Networks,» σε *32nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2024)*, 2024.

[21] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho και H. Chen, «Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection,» σε *International Conference on Learning Representations*, 2018.

[22] R.-Q. Chen, G.-H. Shi, W.-L. Zhao και C.-H. Liang, «A joint model for IT operation series prediction and anomaly detection,» *Neurocomputing,* τόμ. 448, pp. 130-139, 11 August 2021.

[23] I. Fosić, D. Žagar, K. Grgić και V. Križanović, «Anomaly detection in NetFlow network traffic using supervised machine learning algorithms,» *Journal of Industrial Information Integration,* τόμ. 33, p. 100466, 01 June 2023.

[24] C. Yin, Y. Zhu, J. Fei και X. He, «A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,» *IEEE Access,* τόμ. 5, pp. 21954-21961, 2017.

[25] A. T. Assy, Y. Mostafa, A. A. El-khaleq και M. Mashaly, «Anomaly-Based Intrusion Detection System using One-Dimensional Convolutional Neural Network,» *Procedia Computer Science,* τόμ. 220, pp. 78-85, 01 January 2023.

[26] R. Devendiran και A. V. Turukmane, «Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy,» *Expert Systems with Applications,* τόμ. 245, p. 123027, 01 July 2024.

[27] F. Farahnakian και J. Heikkonen, «A deep auto-encoder based approach for intrusion detection system,» σε *2018 20th International Conference on Advanced Communication Technology (ICACT)*, 2018.

[28] H. Choi, M. Kim, G. Lee και W. Kim, «Unsupervised learning approach for network intrusion detection system using autoencoders,» *The Journal of Supercomputing,* τόμ. 75, pp. 5597-5621, 01 September 2019.

[29] C. K. Ramu, T. S. Rao και E. U. S. Rao, «Attack classification in network intrusion detection system based on optimization strategy and deep learning methodology,» *Multimedia Tools and Applications,* τόμ. 83, pp. 75533-75555, 01 September 2024.

[30] B. Sharma, L. Sharma, C. Lal και S. Roy, «Anomaly based network intrusion detection for IoT attacks using deep learning technique,» *Computers and Electrical Engineering,* τόμ. 107, p. 108626, 01 April 2023.

[31] S. Huang και K. Lei, «IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks,» *Ad Hoc Networks,* τόμ. 105, p. 102177, 01 August 2020.

[32] M. Usama, M. Asim, S. Latif, J. Qadir και A. Al-Fuqaha, «Generative Adversarial Networks For Launching and Thwarting Adversarial Attacks on Network Intrusion Detection Systems,» σε *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 2019.

[33] G. Andresini, A. Appice, L. De Rose και D. Malerba, «GAN augmentation to deal with imbalance in imaging-based intrusion detection,» *Future Generation Computer Systems,* τόμ. 123, pp. 108-127, 01 October 2021.

[34] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, USA, 1996.

[35] M. Pérez-Enciso και L. M. Zingaretti, «A Guide on Deep Learning for Complex Trait Genomic Prediction,» *Genes,* τόμ. 10, 2019.

[36] D. E. Rumelhart, G. E. Hinton και R. J. Williams, «Learning representations by back-propagating errors,» *Nature,* τόμ. 323, pp. 533-536, 01 October 1986.

[37] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard και L. Jackel, «Handwritten Digit Recognition with a Back-Propagation Network,» σε *Advances in Neural Information Processing Systems*, 1989.

[38] Δ. Α. Μ. Κωνσταντίνος Διαμαντάρας, Μηχανική Μάθηση, Κλειδάριθμος, Ελλάδα, 2019.

[39] Y. Lecun, L. Bottou, Y. Bengio και P. Haffner, «Gradient-based learning applied to document recognition,» *Proceedings of the IEEE,* τόμ. 86, pp. 2278-2324, 1998.

[40] J. Zhai, S. Zhang, J. Chen και Q. He, «Autoencoder and Its Various Variants,» σε *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.

[41] Y. Lecun, PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models), Universite P. et M. Curie (Paris 6), 1987.

[42] D. Barman, A. Hasnat και R. Nag, «AN INTRODUCTION TO AUTOENCODERS».

[43] D. P. Kingma και J. Ba, *Adam: A Method for Stochastic Optimization,* 2017.

[44] H. Robbins και S. Monro, «A Stochastic Approximation Method,» *The Annals of Mathematical Statistics,* τόμ. 22, p. 400 – 407, 1951.

[45] C. C. Aggarwal, Νευρωνικά Δίκτυα και Βαθιά Μάθηση, Fountas, Ελλάδα, 2020.

[46] N. Hlaing, P. G. Morato, F. de Nolasco Santos, W. Weijtjens, C. Devriendt και P. Rigo, «Farm-wide virtual load monitoring for offshore wind structures via Bayesian neural networks,» *Structural Health Monitoring,* τόμ. 23, pp. 1641-1663, 2024.

[47] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed και A. Lerchner, «beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework,» σε *International Conference on Learning Representations*, 2017.

[48] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins και A. Lerchner, *Understanding disentangling in β-VAE,* 2018.

[49] D. Snover, C. W. Johnson, M. J. Bianco και P. Gerstoft, «Deep Clustering to Identify Sources of Urban Seismic Noise in Long Beach, California,» *Seismological Research Letters,* τόμ. 92, pp. 1011-1022, December 2020.

[50] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville και Y. Bengio, *Generative Adversarial Networks,* 2014.

[51] K. Remya Revi, K. R. Vidya και M. Wilscy, «Detection of Deepfake Images Created Using Generative Adversarial Networks: A Review,» σε *Second International Conference on Networks and Advances in Computational Technologies*, Cham, 2021.

[52] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth και G. Langs, «Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery,» σε *Information Processing in Medical Imaging*, Cham, 2017.

[53] M. Arjovsky, S. Chintala και L. Bottou, *Wasserstein GAN,* 2017.

[54] J. Donahue, P. Krähenbühl και T. Darrell, *Adversarial Feature Learning,* 2017.

[55] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin και A. Courville, *Improved Training of Wasserstein GANs,* 2017.

[56] S. Ioffe και C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,* 2015.

[57] J. L. Ba, J. R. Kiros και G. E. Hinton, *Layer Normalization,* 2016.

[58] H. Asgharzadeh, A. Ghaffari, M. Masdari και F. S. Gharehchopogh, «Anomaly-based intrusion detection system in the Internet of Things using a convolutional neural network and multi-objective enhanced Capuchin Search Algorithm,» *Journal of Parallel and Distributed Computing,* τόμ. 175, pp. 1-21, 2023.

[59] F. W. L. W. P. A. Stolfo και P. Chan, *KDD Cup 1999 Data,* 1999.

[60] M. Tavallaee, E. Bagheri, W. Lu και A. A. Ghorbani, «A detailed analysis of the KDD CUP 99 data set,» σε *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009.

[61] C. B. Anderson, «The CCB-ID approach to tree species mapping with airborne imaging spectroscopy,» *PeerJ,* τόμ. 6, p. e5666, October 2018.

[62] H. Chen, «Novel machine learning approaches for modeling variations in semiconductor manufacturing,» 2017.

[63] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai και S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library,* 2019.

[64] A. L. Maas, «Rectifier Nonlinearities Improve Neural Network Acoustic Models,» 2013.

[65] K. Fukushima, «Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements,» *IEEE Transactions on Systems Science and Cybernetics,* τόμ. 5, pp. 322-333, October 1969.

[66] X. Glorot και Y. Bengio, «Understanding the difficulty of training deep feedforward neural networks,» σε *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Chia Laguna Resort, Sardinia, Italy, 2010.