ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Ανίχνευση Κυβερνοεπιθέσεων με Τεχνικές Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης

Κορνηλίου Ευαγγελία

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος

Αναπληρωτής Καθηγητής

Λαμία 1/10/2025

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

CYBERATTACKS INTRUSION DETECTION USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Korniliou Evangelia

FINAL THESIS

ADVISOR

Dr Konstantinos Kolomvatsos, BSc MSc PhD
Associate Professor in Intelligent Systems for Pervasive Computing

Lamia 1/10/2025

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Η αυξανόμενη συχνότητα και πολυπλοκότητα των κυβερνοεπιθέσεων καθιστά επιτακτική την ανάγκη για σύγχρονες και αποδοτικές μεθόδους ανίχνευσης, ικανές να λειτουργούν αποτελεσματικά ακόμη και σε περιβάλλοντα αβεβαιότητας ή με ελλιπή δεδομένα. Στο πλαίσιο αυτό, η παρούσα πτυχιακή εργασία επικεντρώνεται στην ανάπτυξη και αξιολόγηση μοντέλων ανίχνευσης επιθέσεων με τη χρήση τεχνικών Μηχανικής Μάθησης (Machine Learning, ML) και Τεχνητής Νοημοσύνης (Artificial Intelligence, AI), με στόχο την έγκαιρη και ακριβή ανάλυση ροών δικτύου.

Για την εκπαίδευση και αξιολόγηση των μοντέλων αξιοποιήθηκαν τα ευρέως διαδεδομένα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018, τα οποία καθαρίστηκαν από διπλότυπα και σφάλματα και ομογενοποιήθηκαν ως προς τα χαρακτηριστικά (features) τους. Η μεθοδολογία στηρίχθηκε στον αλγόριθμο XGBoost, ο οποίος εφαρμόστηκε σε διαφορετικά σενάρια.

Επιπλέον, για την ενίσχυση του όγκου και της ποικιλίας των δεδομένων εφαρμόστηκαν τεχνικές Generative Adversarial Networks (GANs) και Variational Autoencoders (VAEs). Τα VAEs χρησιμοποιήθηκαν για εμπλουτισμό δεδομένων (Data Augmentation) ενώ τα GANs για μεταγενέστερη αξιολόγηση της γενίκευσης των μοντέλων σε αποκλειστικά συνθετικά δεδομένα.

Η εκπαίδευση των μοντέλων βασίστηκε σε επιβλεπόμενη μάθηση (supervised learning), ενώ η φάση πρόβλεψης σχεδιάστηκε να λειτουργεί σε μη επισημασμένα δεδομένα (unlabeled data), προσομοιώνοντας πραγματικά σενάρια ανίχνευσης όπου ο τύπος της ροής (κανονική ή κακόβουλη) δεν είναι εκ των προτέρων γνωστός.

Στα επόμενα κεφάλαια παρουσιάζεται το θεωρητικό υπόβαθρο, η μεθοδολογία που ακολουθήθηκε, η ανάλυση των αποτελεσμάτων καθώς και τα συμπεράσματα που προκύπτουν από την έρευνα.

ABSTRACT

The increasing frequency and complexity of cyberattacks require modern detection methods capable of operating under uncertain or incomplete data. This thesis focuses on the development and evaluation of attack detection techniques based on Machine Learning (ML) and Artificial Intelligence (AI), aiming for the effective analysis of network traffic flows. For the training and evaluation of the models, the widely used datasets CICIDS 2017 and CSE-CICIDS 2018 were utilized, cleaned, and homogenized regarding feature names and structure. The methodology was based on the XGBoost algorithm, which was applied in different experimental scenarios.

In addition, to increase the volume and diversity of the available data, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) were employed. VAEs were used as a tool for data augmentation and GANs for further evaluating the generalization capacity of the models on exclusively synthetic data.

The models were trained in a supervised manner, while the prediction phase was designed to operate without access to any labels, simulating real-world scenarios where detection must be performed on new or unknown network traffic.

The following chapters present the theoretical background, methodology, analysis of results, and the conclusions of this work.

Table of Contents

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

(Υποκεφάλαιο1.1) Ιστορικό και Πλαίσιο του Προβλήματος

Η συνεχώς αυξανόμενη εξάρτηση της κοινωνίας, των επιχειρήσεων και των οργανισμών από τα πληροφοριακά συστήματα και τα δίκτυα έχει οδηγήσει στο γεγονός της κυβερνοασφάλειας πλέον να αποτελεί ζήτημα ζωτικής σημασίας. Από την αρχή της ύπαρξης του διαδικτύου, όπου οι απειλές ήταν σχετικά απλές και σπάνιες, φτάσαμε σε ένα σημερινό περιβάλλον όπου οι κυβερνοεπιθέσεις έχουν εξελιχθεί τόσο σε πολυπλοκότητα όσο και σε συχνότητα [2].
Οι πρώτες καταγεγραμμένες κυβερνοεπιθέσεις εμφανίστηκαν στις δεκαετίες του 1980 και του 1990 [1] , αλλά η ραγδαία αύξηση της τεχνολογικής διασύνδεσης την τελευταία εικοσαετία έκανε το φαινόμενο παγκόσμιο και ιδιαίτερα κρίσιμο. Επιθέσεις όπως Malware, Ransomware, Phishing, Denial of Service (DoS) και Distributed Denial of Service (DDoS) έχουν επηρεάσει πολλούς οργανισμούς, προκαλώντας οικονομικές απώλειες, προβλήματα φήμης και διαρροές προσωπικών δεδομένων [3,4]. Η έντονη εξάπλωση έξυπνων συσκευών, η αλληλεπίδραση του διαδικτύου με την καθημερινότητα μας (για ενέργεια, μεταφορές, υγεία, οικονομία), καθώς και η ανάγκη για Cloud περιβάλλοντα έχουν αυξήσει σημαντικά το ψηφιακό αποτύπωμα των οργανισμών και κατά συνέπεια άνοιξαν πύλες εισόδου για επιτιθέμενους [6].
Σε αυτό το πλαίσιο, η ανάγκη για αποτελεσματική ανίχνευση και αντιμετώπιση επιθέσεων γίνεται όλο και πιο επιτακτική. Η κλασική προσέγγιση με στατικούς κανόνες (Signature-Based Detection) δεν αρκεί για να καλύψει τον συνεχώς μεταβαλλόμενο χαρακτήρα των σύγχρονων επιθέσεων [7]. Έτσι, η αναζήτηση και εφαρμογή νέων, ευφυών μεθόδων ασφάλειας βρίσκεται στο επίκεντρο της σύγχρονης επιστημονικής και επαγγελματικής δραστηριότητας στον τομέα της Πληροφορικής.

(Ενότητα 1.1.α Ιστορικό)

Δεκαετία 1970–1980
Η ιστορία της ασφάλειας των πληροφοριακών συστημάτων ξεκίνησε σχεδόν ταυτόχρονα με την ανάπτυξη των πρώτων δικτύων υπολογιστών. Τις δεκαετίες του 1970 και 1980, όταν τα συστήματα ήταν σχετικά απλά, οι απειλές περιορίζονταν κυρίως σε ακούσια σφάλματα ή ανθρώπινη αμέλεια.
Το πρώτο γνωστό παράδειγμα ιού ήταν το *Creeper* (*1971*), το οποίο, αν και ακίνδυνο, έθεσε τις βάσεις για τις μετέπειτα απειλές. [1]

Δεκαετία 1980
Με την εισαγωγή του Διαδικτύου και της μαζικής δικτύωσης άρχισαν να καταγράφονται και οι πρώτες σκόπιμες επιθέσεις. Ένα χαρακτηριστικό παράδειγμα ήταν το *Morris Worm (1988)*, μία από τις πρώτες μαζικές επιθέσεις που εξαπλώθηκαν στο Διαδίκτυο και προκάλεσε μεγάλη αναστάτωση αναδεικνύοντας τις ευπάθειες των δικτύων της εποχής. [1]

Δεκαετία 1990

Στη δεκαετία του 1990, οι hackers και τα πρώτα κακόβουλα λογισμικά έγιναν γνωστά στο ευρύ κοινό, με επιθέσεις που στόχευαν είτε στην απόκτηση πρόσβασης, είτε στην πρόκληση φθοράς ή στη διακίνηση ιών. [1,5]

Σταδιακά, η εισαγωγή ψηφιακών υποδομών στις επιχειρήσεις, η χρήση των προσωπικών υπολογιστών και του Διαδικτύου δημιούργησαν ένα νέο, πιο πολύπλοκο και ευάλωτο περιβάλλον όπου οι απειλές άρχισαν να αποκτούν σημαντική βαρύτητα και να προκαλούν σοβαρές ζημιές. [5]

(Ενότητα 1.1.β Εξέλιξη κυβερνοεπιθέσεων και σημασία ανίχνευσης)

Κατά τις τελευταίες δύο δεκαετίες, οι κυβερνοεπιθέσεις έχουν εξελιχθεί σε ιδιαίτερα πολύπλοκες και στοχευμένες επιθέσεις με σοβαρές συνέπειες. Η εμφάνιση οργανωμένων ομάδων κυβερνοεγκλήματος (Cybercrime Groups), η ανάπτυξη της βιομηχανίας Ransomware [4] και οι στοχευμένες επιθέσεις αποδεικνύουν την επικινδυνότητα που διατρέχει τη σύγχρονη ψηφιακή εποχή.

Νέα είδη επιθέσεων, όπως τα Distributed Denial of Service (DDoS), το Ransomware, το Phishing και τα Botnets, απαιτούν πολύ πιο αποτελεσματικές και αυτοματοποιημένες μεθόδους ανίχνευσης [3]. Ενώ η μαζική εξάπλωση των συσκευών IoT (Internet of Things) και η αύξηση του Cloud Computing έχουν αφήσει εκτεθειμένα σημεία στα συστήματα, τα οποία μπορούν να αποτελέσουν στόχο επιθέσεων [6].

Οι παραδοσιακές μέθοδοι ανίχνευσης αδυνατούν να καλύψουν τη διαρκώς μεταβαλλόμενη φύση των επιθέσεων. Οι επιτιθέμενοι προσαρμόζουν συνεχώς τις τεχνικές τους, καμουφλάροντας κακόβουλη δραστηριότητα ώστε να μοιάζει με κανονική χρήση του δικτύου [8].

Για όλους αυτούς τους λόγους, η έγκαιρη και αξιόπιστη ανίχνευση επιθέσεων έχει γίνει προτεραιότητα για οργανισμούς, κράτη και επιχειρήσεις. Η επιτυχής ανίχνευση επιτρέπει την έγκαιρη αποφυγή και ελαχιστοποίηση των συνεπειών, προστατεύοντας τα δεδομένα, τη λειτουργία των υπηρεσιών και, τελικά, την ίδια την κοινωνία.

(Υποκεφάλαιο 1.2) Ρόλος Μηχανικής Μάθησης στην ασφάλεια

Η μηχανική μάθηση έχει αναδειχθεί ως ένα από τα πιο αποτελεσματικά εργαλεία για την αντιμετώπιση των σύγχρονων απειλών στην ασφάλεια πληροφοριακών συστημάτων [9]. Σε αντίθεση με τις παραδοσιακές μεθόδους ανίχνευσης, που βασίζονται σε στατικές υπογραφές (ένα μοναδικό αποτύπωμα ενός κακόβουλου λογισμικού ή μιας απειλής) ή προκαθορισμένους κανόνες [7], τα συστήματα μηχανικής μάθησης μπορούν να αναλύουν μεγάλα σύνολα δεδομένων και να εντοπίζουν μοτίβα που υποδηλώνουν κακόβουλη δραστηριότητα.

Ένα από τα κύρια πλεονεκτήματα της μηχανικής μάθησης είναι η ικανότητά της να ανιχνεύει ακόμη και άγνωστες ή εξελιγμένες επιθέσεις, οι οποίες συχνά δεν έχουν σαφή χαρακτηριστικά ή προηγούμενα δεδομένα [9]. Μέσω της εκπαίδευσης σε πραγματικά δεδομένα δικτυακής κίνησης, τα μοντέλα μηχανικής μάθησης μπορούν να εντοπίζουν αποκλίσεις από δεδομένα φυσιολογικής συμπεριφοράς και να προσαρμόζονται σε νέες απειλές με παρόμοια μοτίβα.

Αυτές οι μέθοδοι προσφέρουν δυνατότητα αυτοματοποίησης, ταχύτητας στην επεξεργασία δεδομένων μεγάλων όγκων και ευελιξία στην αντιμετώπιση διαφορετικών τύπων επιθέσεων. Υπάρχουν δύο βασικές προσεγγίσεις στη χρήση μηχανικής μάθησης για την

ανίχνευση επιθέσεων, η επιβλεπόμενη (Supervised) και η μη επιβλεπόμενη (Unsupervised) μάθηση [6,9].

Στην επιβλεπόμενη μάθηση**,** τα μοντέλα εκπαιδεύονται με δεδομένα που έχουν ετικέτες (Labels), δηλαδή γνωρίζουμε αν κάθε παράδειγμα είναι επίθεση ή κανονική κίνηση [9]. Δημοφιλείς αλγόριθμοι αυτής της κατηγορίας είναι το Random Forest, το XGBoost [16], το Support Vector Machine (SVM) και τα Νευρωνικά Δίκτυα (Neural Networks).

Στη μη επιβλεπόμενη μάθηση, τα μοντέλα προσπαθούν να ανακαλύψουν ανωμαλίες ή νέους τύπους επιθέσεων χωρίς να υπάρχουν ετικέτες. Χρησιμοποιούνται αλγόριθμοι όπως το Isolation Forest, το k-Means Clustering και τεχνικές ανίχνευσης ανωμαλιών (Anomaly Detection).Στα σύγχρονα συστήματα, συνδυάζονται συχνά και οι δύο προσεγγίσεις για να επιτυγχάνεται τόσο η ανίχνευση γνωστών επιθέσεων όσο και η ανακάλυψη νέων, άγνωστων απειλών. [9]

Ένα παράδειγμα αλγορίθμου που έχει σημειώσει ιδιαίτερη επιτυχία στον τομέα είναι το XGBoost (Extreme Gradient Boosting), λόγω της υψηλής του ακρίβειας, της ανθεκτικότητας του σε θόρυβο δεδομένων και της ταχύτητας εκπαίδευσης [16] . Παράλληλα, παρέχει εργαλεία όπως το SHAP για ερμηνεία των αποτελεσμάτων και κατανόηση της συμβολής κάθε χαρακτηριστικού (Feature) στην τελική απόφαση του μοντέλου.

Η ενσωμάτωση τεχνικών μηχανικής μάθησης σε συστήματα ανίχνευσης εισβολών επιτρέπει τη γρήγορη επεξεργασία μεγάλου όγκου δεδομένων, την ανίχνευση περίπλοκων επιθέσεων σε πραγματικό χρόνο και την προσαρμογή των συστημάτων στις συνεχώς μεταβαλλόμενες απειλές.

ΚΕΦΑΛΑΙΟ 2 Περιγραφή & Ανάλυση των Δεδομένων

(Υποκεφάλαιο 2.1 CICIDS 2017/CSE-CICIDS 2018)

(Ενότητα 2.1.α Εισαγωγή)

Για να αξιοποιήσουν τα συστήματα ανίχνευσης επιθέσεων στο μέγιστο τις δυνατότητες της μηχανικής μάθησης (machine learning), είναι απαραίτητο να εκπαιδεύονται και να εφαρμόζονται σε κατάλληλα και αξιόπιστα σύνολα δεδομένων. Η επάρκεια, η ρεαλιστικότητα και η καθαρότητα των συνόλων δεδομένων (datasets) επιτρέπει στους αλγορίθμους να ανιχνεύουν τόσο γνωστές όσο και νέες μορφές επιθέσεων σε πραγματικές συνθήκες δικτύου. [9,12]
Για τον σκοπό αυτό, τα δημόσια διαθέσιμα σύνολα δεδομένων, όπως τα CICIDS 2017 και CSE-CICIDS 2018, έχουν καθιερωθεί ως ιδιαίτερα χρήσιμα εργαλεία στην επιστημονική κοινότητα. Τα συγκεκριμένα δεδομένα περιλαμβάνουν διάφορους τύπους επιθέσεων και μεγάλο όγκο κανονικής κίνησης, παρέχοντας τη δυνατότητα για ολοκληρωμένη μελέτη και αξιόπιστη σύγκριση τεχνικών ανίχνευσης εισβολών.

(Ενότητα 2.1.β Περιγραφή των Δεδομένων (CICIDS 2017 & CSE-CICIDS 2018))

Τα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018 δημιουργήθηκαν από το Canadian Institute for Cybersecurity μέσα από οργανωμένα πειράματα όπου ειδικοί αναπαρήγαγαν επιθέσεις σε εργαστηριακό περιβάλλον. Η καταγραφή της κίνησης πραγματοποιήθηκε μέσω του εργαλείου CICFlowMeter, το οποίο εξάγει τα δεδομένα σε προκαθορισμένη μορφή, περιλαμβάνοντας τόσο επιθετική όσο και κανονική δραστηριότητα δικτύου. [8]
Το σύνολο δεδομένων του 2017 περιλαμβάνει ροή διαδικτυακής κίνησης καταγεγραμμένη σε διάστημα πέντε ημερών, χωρισμένων σε πρωινές και απογευματινές ώρες, ενώ του 2018 σε διάστημα δέκα ημερών. Στα δεδομένα περιλαμβάνονται διάφορες επιθέσεις, όπως DoS, DDoS, Brute Force, Botnet, SQL injection, XSS, FTP Brute Force, SSH Brute Force, καθώς και άλλες. [10,11]
Τα δεδομένα αποτελούνται από εκατομμύρια εγγραφές, με κάθε εγγραφή να περιέχει 79 χαρακτηριστικά (features) που περιγράφουν λεπτομερώς τις ροές δικτύου (network flows), όπως αριθμός πακέτων, μέγεθος πακέτων, χρόνους μεταξύ πακέτων, σημαίες (flags), θύρες (ports) και άλλα στοιχεία. Αυτή η δομή παρέχει την δυνατότητα για βαθιά ανάλυση συμπεριφοράς ροών δικτύων κατά την διάρκεια επιθέσεων, συγκριτικά με αυτή φυσιολογικής κίνησης. [8]
Παρά την υψηλή ποιότητά τους, τα συγκεκριμένα σύνολά δεδομένων παρουσιάζουν ορισμένες προκλήσεις, όπως ανισορροπία ανάμεσα σε κατηγορίες (πολύ περισσότερα δείγματα κανονικής ροής), διπλότυπα δείγματα ή λανθασμένα αρχεία, καθώς και ορισμένους τύπους επιθέσεων με πολύ λίγα παραδείγματα. Για τον λόγο αυτό απαιτείται προσεκτική προεπεξεργασία (preprocessing) πριν από την εκπαίδευση των μοντέλων. [12]

(Ενότητα 2.1.γ Κριτική Αξιολόγηση των CICIDS 2017 & 2018 και Επιλογή Διορθωμένης Έκδοσης)

Παρότι τα CICIDS 2017 και CSE-CICIDS 2018 σύνολα δεδομένων θεωρούνται ευρέως αποδεκτά και χρησιμοποιούνται σε πολλές έρευνες ανίχνευσης επιθέσεων, έχει αναγνωριστεί ότι περιλαμβάνουν σημαντικά σφάλματα τα οποία επηρεάζουν την αξιοπιστία

των πειραμάτων. Σύμφωνα με τη μελέτη των Liu et al. (*2022*), έχουν εντοπιστεί προβλήματα κατά τη δημιουργία επιθέσεων, όπως λανθασμένες ετικέτες (labels), κυρίως στα δεδομένα με κανονική (benign) κίνηση, η οποία έχει χαρακτηριστεί λανθασμένα ως επιθετική και αντίστροφα. Επιπλέον, παρατηρήθηκε εσφαλμένη ανίχνευση πρωτοκόλλων για παράδειγμα, σε ροές TCP βρέθηκε ασυνέπεια λόγω ελλιπούς τερματισμού, με αποτέλεσμα να συγχέονται με ροές του πρωτοκόλλου UDP. [12]
Άλλα προβλήματα που εντοπίστηκαν αφορούσαν χρονικές ασυνέπειες λόγω κακής διαχείρισης των χρονικών σημάνσεων (timestamps), οι οποίες προκάλεσαν αλληλοεπικάλυψη μεταξύ επιθέσεων και φυσιολογικής δραστηριότητας. Τέλος, εντοπίστηκαν πολλά διπλότυπα και επαναλαμβανόμενα δεδομένα, τα οποία είχαν εξαχθεί εσφαλμένα από το εργαλείο CICFlowMeter. [12]
Τα παραπάνω σφάλματα μπορούν να οδηγήσουν σε ελλιπή και παραπλανητική εκπαίδευση και αξιολόγηση των μοντέλων, καθώς ενδέχεται αυτά να μάθουν με βάση δεδομένα που είναι είτε λανθασμένα είτε μη ρεαλιστικά. Για τον λόγο αυτό, επιλέχθηκε η χρήση της διορθωμένης έκδοσης των CICIDS 2017 και CSE-CICIDS 2018, όπως παρουσιάζεται από τους Liu et al. (*2022*). Σε αυτήν την έκδοση έχει εφαρμοστεί επανετικετοποίηση με βάση ανάλυσης περιεχομένου, έχουν αφαιρεθεί διπλότυπα, ενώ τα δεδομένα έχουν καθαριστεί και ευθυγραμμιστεί σε σωστή χρονολογική σειρά σύμφωνα με τα timestamps. [12]

(Υποκεφάλαιο 2.2 Καθαρισμός Δεδομένων)

Πάνω στα δεδομένα της διορθωμένης έκδοσης πραγματοποιήθηκε επιπλέον προεπεξεργασία (data preprocessing). Αρχικά, αφαιρέθηκαν πεδία που δεν συνεισέφεραν στην ανάλυση, όπως το Flow ID, η διεύθυνση πηγής (Src IP), η διεύθυνση προορισμού (Dst IP) και η χρονική σήμανση (Timestamp). Στη συνέχεια, απομακρύνθηκαν όλες οι γραμμές που περιείχαν τιμές NaN (Not a Number) ή Inf (Infinity).
Επιπλέον, πραγματοποιήθηκε ενοποίηση των ετικετών (labels) μεταξύ των σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018. Σε αρκετές περιπτώσεις η ίδια επίθεση εμφανιζόταν με διαφορετική ονομασία στα δύο σύνολα δεδομένων, παρότι είχε υλοποιηθεί με το ίδιο εργαλείο και ακολουθούσε το ίδιο μοτίβο. Για να αποφευχθούν τέτοιες ασυνέπειες και να εξασφαλιστεί ομοιομορφία στην κατηγοριοποίηση, οι επιθέσεις αυτές μετονομάστηκαν με κοινό όνομα τόσο στα αρχεία όσο και στο αντίστοιχο label.
Τέλος, εφαρμόστηκε τυχαία δειγματοληψία από τα δεδομένα κανονικής κίνησης (benign traffic) με στόχο να μειωθεί η πληθώρα benign δειγμάτων και να επιτευχθεί πιο ισορροπημένη κατανομή του συνόλου δεδομένων.

Οι κατηγορίες που χρησιμοποιήθηκαν μετά τη μετονομασία και ενοποίηση των ετικετών από τα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018 παρουσιάζονται στον παρακάτω πίνακα. Επιπλέον, σε ξεχωριστούς πίνακες αποτυπώνεται η διαδικασία καθαρισμού των δεδομένων ανά αρχείο.

CICIDS 2017	CSE-CICIDS 2018
BENIGN	BENIGN
DoS Hulk	DoS Hulk
DoS GoldenEye	DoS GoldenEye
DoS Slowloris	DoS Slowloris
DoS Slowhttptest	–

DDoS LOIC HTTP	DDoS LOIC HTTP
–	DDoS HOIC
–	DDoS LOIC UDP
Infiltration - Communication Victim Attacker	Infiltration - Communication Victim Attacker
Infiltration - NMAP Portscan	Infiltration - NMAP Portscan
–	Infiltration - Dropbox Download
Botnet Ares	Botnet Ares
Web Attack - Brute Force	Web Attack - Brute Force
Web Attack - SQL Injection	Web Attack - SQL Injection
Web Attack - XSS	Web Attack - XSS
Heartbleed	–
Portscan	–
SSH-BruteForce-Patator	SSH-BruteForce-Patator
FTP BruteForce Patator	–

Πίνακας Δεδομένων 2017 – Καθαρισμός Δεδομένων CICIDS 2017

Όνομα Αρχείου	Αρχικές Γραμμές	Τελικές Γραμμές	Στήλες που Αφαιρέθηκαν	Γραμμές με NaN/inf
DDoS LOIC HTTP.csv	95144	95144	Attempted Category, id, Flow ID, Src IP, Src Port, Dst IP, Timestamp	0
BENIGN.csv	1582566	1582561	ίδιες όπως παραπάνω	5
DoS Slowhttptest.csv	1740	1740	ίδιες όπως παραπάνω	0
DoS Hulk.csv	158468	158468	ίδιες όπως παραπάνω	0
DoS GoldenEye.csv	7567	7567	ίδιες όπως παραπάνω	0
DoS Slowloris.csv	3859	3859	ίδιες όπως παραπάνω	0
Portscan.csv	159066	159066	ίδιες όπως παραπάνω	0
SSH-BruteForce-Patator.csv	2961	2961	ίδιες όπως παραπάνω	0

Infiltration - Communication Victim Attacker.csv	32	32	ίδιες όπως παραπάνω	0
Heartbleed.csv	11	11	ίδιες όπως παραπάνω	0
Web Attack - Brute Force.csv	73	73	ίδιες όπως παραπάνω	0
FTP-BruteForce-Patator.csv	3972	3972	ίδιες όπως παραπάνω	0
Web Attack - SQL Injection.csv	13	13	ίδιες όπως παραπάνω	0
Web Attack - XSS.csv	18	18	ίδιες όπως παραπάνω	0
Infiltration - NMAP Portscan.csv	71767	71767	ίδιες όπως παραπάνω	0
Botnet Ares.csv	736	736	ίδιες όπως παραπάνω	0

Πίνακας Δεδομένων 2018 – Καθαρισμός Δεδομένων CSE-CICIDS 2018

Όνομα Αρχείου	Αρχικές Γραμμές	Τελικές Γραμμές	Στήλες που Αφαιρέθηκαν	Γραμμές με NaN/inf
DDoS LOIC HTTP.csv	289.328	289.328	Attempted Category, id, Flow ID, Src IP, Src Port, Dst IP, Timestamp	0
Infiltration - Dropbox Download.csv	85	85	ίδιες όπως παραπάνω	0
BENIGN.csv	5.000.000	4.999.999	ίδιες όπως παραπάνω	1
Botnet Ares.csv	142.921	142.921	ίδιες όπως παραπάνω	0
DoS Hulk.csv	1.803.160	1.803.160	ίδιες όπως παραπάνω	0

DoS GoldenEye.csv	22.560	22.560	ίδιες όπως παραπάνω	0
DoS Slowloris.csv	8.490	8.490	ίδιες όπως παραπάνω	0
DDoS HOIC.csv	1.082.293	1.082.293	ίδιες όπως παραπάνω	0
SSH-BruteForce-Patator.csv	94.197	94.197	ίδιες όπως παραπάνω	0
Infiltration - Communication Victim Attacker.csv	204	204	ίδιες όπως παραπάνω	0
Web Attack -Brute Force.csv	131	131	ίδιες όπως παραπάνω	0
Web Attack - SQL Injection.csv	39	39	ίδιες όπως παραπάνω	0
Web Attack - XSS.csv	113	113	ίδιες όπως παραπάνω	0
DDoS LOIC UDP.csv	2.527	2.527	ίδιες όπως παραπάνω	0
Infiltration - NMAP Portscan.csv	89.374	89.374	ίδιες όπως παραπάνω	0

Σημείωση: Ο όρος "ίδιες όπως παραπάνω" αναφέρεται στις στήλες που αφαιρέθηκαν, όπως καταγράφονται στην πρώτη γραμμή του πίνακα.

ΚΕΦΑΛΑΙΟ 3 Κυβερνοασφάλεια & Ανάλυση Επιθέσεων

(Υποκεφάλαιο 3.1 Κυβερνοασφάλεια και Συστήματα Ανίχνευσης Εισβολών (IDS))

(Ενότητα 3.1.α Κυβερνοασφάλεια)

Η κυβερνοασφάλεια (cybersecurity) αποτελεί το σύνολο των τεχνικών, διαδικασιών και μέτρων ασφαλείας που στοχεύουν στην προστασία των πληροφοριακών συστημάτων, των δικτύων και των δεδομένων από μη εξουσιοδοτημένη πρόσβαση, κακόβουλες ενέργειες ή καταστροφή προσωπικών δεδομένων. [5]

Υπάρχουν τρία θεμελιώδη στοιχεία που καθορίζουν την κυβερνοασφάλεια, γνωστά ως CIA Triad, τα οποία θέτουν τους βασικούς της στόχους: [5]

- Εμπιστευτικότητα (Confidentiality): Η προστασία των δεδομένων από μη εξουσιοδοτημένη πρόσβαση, ώστε μόνο οι εξουσιοδοτημένοι χρήστες να έχουν δυνατότητα πρόσβασης.
- Ακεραιότητα (Integrity): Η διασφάλιση ότι τα δεδομένα παραμένουν ακριβή και αναλλοίωτα κατά τη διάρκεια της αποθήκευσης, της επεξεργασίας και της μεταφοράς.
- Διαθεσιμότητα (Availability): Η εξασφάλιση ότι τα δεδομένα και τα συστήματα είναι διαθέσιμα κάθε στιγμή που χρειάζονται.

Η κυβερνοασφάλεια αναφέρεται στο πλαίσιο της πρόληψης, ανίχνευσης και απόκρισης σε περιστατικά κυβερνοεπιθέσεων ή παραβίασης ασφάλειας, τα οποία μπορεί να είναι σκόπιμα ή τυχαία (όπως π.χ. διαρροή πληροφοριών από ακούσιο ανθρώπινο λάθος). Στόχος της είναι όχι μόνο η άμεση αντιμετώπιση του προβλήματος, αλλά και η πλήρης αποκατάσταση των δεδομένων και της λειτουργίας του συστήματος μετά το περιστατικό. [5]

(Ενότητα 3.1.β Συστήματα Ανίχνευσης Εισβολών IDS)

Οι κυβερνοεγκληματίες έχουν εξελίξει τις μεθόδους τους σε τέτοιο βαθμό ώστε τα παραδοσιακά εργαλεία προστασίας, όπως τα antivirus και τα firewalls, να μην επαρκούν για την αποτελεσματική πρόληψη επιθέσεων [9]. Για τον λόγο αυτό αναπτύχθηκαν τα Συστήματα Ανίχνευσης Εισβολών (Intrusion Detection Systems – IDS), τα οποία βασίζονται σε προηγμένες μεθόδους παρακολούθησης και ανάλυσης ύποπτης διαδικτυακής κίνησης ή κακόβουλης δραστηριότητας σε συσκευές. Τα IDS δεν μπλοκάρουν ενεργά την κίνηση, αλλά χρησιμοποιούνται για την έγκαιρη ενημέρωση και προειδοποίηση των υπευθύνων ασφαλείας.

Τα IDS διακρίνονται σε δύο βασικές κατηγορίες: στα network-based IDS (NIDS), που παρακολουθούν την κίνηση σε επίπεδο δικτύου, και στα host-based IDS (HIDS), που ελέγχουν τη δραστηριότητα σε συγκεκριμένες συσκευές [9]. Με βάση την μέθοδο ανίχνευσης, διακρίνονται δύο κύριες μέθοδοι: η signature-based detection, που αναγνωρίζει επιθέσεις με βάση γνωστά μοτίβα (υπογραφές – signatures) [7], και η anomaly-based detection που εντοπίζει ασυνήθιστη ή αποκλίνουσα συμπεριφορά από το φυσιολογικό πρότυπο, συχνά με χρήση τεχνικών μηχανικής μάθησης (machine learning) [9].

Στην παρούσα εργασία υιοθετείται η προσέγγιση anomaly-based detection με χρήση flow-based χαρακτηριστικών, δηλαδή στατιστικών στοιχείων που προκύπτουν από τη ροή (flow)

της δικτυακής κίνησης [14]. Αντίθετα, η packet-based ανάλυση εξετάζει το περιεχόμενο κάθε πακέτου ξεχωριστά. Η μελέτη επικεντρώνεται στη flow-based μέθοδο, ενώ η packet-based παρουσιάζεται για λόγους πληρότητας και σύγκρισης [13].
Η packet-based ανάλυση εξετάζει το περιεχόμενο κάθε πακέτου ανεξάρτητα. Κάθε πακέτο διαθέτει το δικό του payload (το περιεχόμενο του πακέτου) και τα δικά του headers. Σε επίπεδο headers, παρακολουθούνται στοιχεία όπως το Ethernet header (διευθύνσεις MAC, τύπος πρωτοκόλλου), το IP header (διευθύνσεις IP, TTL, checksum, flags) και το TCP/UDP header (θύρες πηγής και προορισμού, checksum, offset) [13]. Στο payload μπορούν να εντοπιστούν μοτίβα όπως HTTP requests, προσπάθειες SQL injection, ή ακόμη και τμήματα κακόβουλου κώδικα στην περίπτωση malware traffic [14]. Συνεπώς, σε packet-based σύνολα δεδομένων κάθε γραμμή αντιστοιχεί σε ένα πακέτο, ενώ στη flow-based κάθε γραμμή αναφέρεται σε μία ροή δικτύου.
Αντίθετα, η flow-based ανάλυση δεν εξετάζει το περιεχόμενο των πακέτων αλλά εστιάζει σε στατιστικά χαρακτηριστικά που προκύπτουν από το σύνολο των πακέτων μιας ροής [14]. Τέτοια χαρακτηριστικά είναι αριθμητικές τιμές όπως ο μέσος όρος (mean), η τυπική απόκλιση (std), οι ελάχιστες και μέγιστες τιμές (min, max) και το πλήθος πακέτων (count) [14]. Η ροή ορίζεται από στοιχεία όπως διεύθυνση πηγής (SrcIP), διεύθυνση προορισμού (DstIP), θύρες πηγής και προορισμού (SrcPort, DstPort), το πρωτόκολλο (Protocol) και την χρονική σήμανση (Timestamp) [13].
Η flow-based προσέγγιση είναι ιδιαίτερα αποδοτική για την ανίχνευση τόσο γνωστών όσο και νέων επιθέσεων, ενώ συνιστάται για την επεξεργασία μεγάλου όγκου δεδομένων, όπως αυτά των σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018 που χρησιμοποιούνται στην παρούσα εργασία. Παράλληλα, συνδυάζεται ιδανικά με αλγορίθμους μηχανικής μάθησης, καθώς τα παραγόμενα στατιστικά χαρακτηριστικά μπορούν να χρησιμοποιηθούν ως είσοδοι στα μοντέλα.

(Υποκεφάλαιο 3.2 Ανάλυση επιθέσεων και χαρακτηριστικών)

(Ενότητα 3.2.α Χαρακτηριστικά Features)

Τα χαρακτηριστικά χωρίζονται, βάσει δικής μου επιλογής για λόγους καλύτερης κατανόησης, σε πέντε κύριες κατηγορίες, αξιοποιώντας πληροφορίες που προκύπτουν από τη σχετική βιβλιογραφία [8,13,14] ανάλογα με το είδος της πληροφορίας που περιγράφουν.
Πρώτη είναι τα flow-level (χαρακτηριστικά ροής), τα οποία αφορούν το συνολικό πλαίσιο της ροής, όπως τη διάρκεια, τον όγκο δεδομένων και τους ρυθμούς μετάδοσης, παρέχοντας μια συνολική εικόνα της επικοινωνίας.
Η δεύτερη κατηγορία είναι τα packet-level (χαρακτηριστικά πακέτων) που περιγράφουν στατιστικά στοιχεία σχετικά με τα μεγέθη πακέτων μέσα στη ροή (έχει καταγραφεί ο μέσος όρος, το μέγιστο, το ελάχιστο και η διακύμανση). Δίνουν πληροφορίες για την κίνηση και μπορούν να τονίσουν ανωμαλίες όπως ασυνήθιστα μεγάλα ή μικρά πακέτα.
Σε αντίθεση με τα παραπάνω χαρακτηριστικά που μετρούν όγκους και μεγέθη, η επόμενη κατηγορία είναι τα time-based (χρονικά χαρακτηριστικά), τα οποία επικεντρώνονται στις χρονικές σχέσεις μεταξύ πακέτων.
Ως τέταρτη κατηγορία έχουμε τα flags και τα πρωτόκολλα, που εμπεριέχουν κυρίως TCP flags (SYN, ACK, FIN, RST κ.λπ.) και καταγράφουν το είδος των σημάτων ελέγχου που χρησιμοποιούνται στις συνδέσεις.
Ακολουθούν τα bulk/subflow (χαρακτηριστικά υποροής), που περιγράφουν λεπτομερώς την κίνηση σε υποροές, μετρώντας πακέτα/bytes σε συγκεκριμένες κατευθύνσεις και ρυθμούς

bulk και είναι χρήσιμα για την αποτύπωση πιο μικρών διαφορών στη συμπεριφορά της κίνησης.
Τέλος, αν και δεν αποτελεί ξεχωριστή κατηγορία ,είναι η στήλη label (ετικέτα) που περιέχει μια αλφαριθμητική τιμή (BENIGN ή ετικέτα επίθεσης) και χρησιμοποιείται ως πραγματική αναφορά (ground truth) για την εκπαίδευση και αξιολόγηση των μοντέλων ανίχνευσης.

ΑΝΑΛΥΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΣΕ ΠΙΝΑΚΕΣ

Πίνακας 3.1 – Χαρακτηριστικά Δικτυακής Ροής (Flow-level Features)

Feature	Περιγραφή
Dst Port	Θύρα προορισμού της ροής.
Protocol	Πρωτόκολλο μεταφοράς (TCP, UDP, ICMP).
Flow Duration	Συνολική διάρκεια της ροής (ms).
Flow Bytes/s	Ρυθμός μετάδοσης δεδομένων σε bytes/sec.
Flow Packets/s	Ρυθμός μετάδοσης σε πακέτα/sec.
Total TCP Flow Time	Συνολικός χρόνος TCP σύνδεσης.

Πίνακας 3.2 – Στατιστικά Πακέτων (Packet-level Features)

Feature	Περιγραφή
Total Fwd Packets	Συνολικός αριθμός πακέτων εμπρός (source → destination).
Total Bwd Packets	Συνολικός αριθμός πακέτων πίσω (destination → source).
Total Length of Fwd Packets	Συνολικό μήκος πακέτων εμπρός (bytes).
Total Length of Bwd Packets	Συνολικό μήκος πακέτων πίσω (bytes).
Fwd Packet Length Max	Μέγιστο μέγεθος πακέτου εμπρός.
Fwd Packet Length Min	Ελάχιστο μέγεθος πακέτου εμπρός.
Fwd Packet Length Mean	Μέσο μέγεθος πακέτου εμπρός.
Fwd Packet Length Std	Τυπική απόκλιση μεγέθους πακέτων εμπρός.
Bwd Packet Length Max	Μέγιστο μέγεθος πακέτου πίσω.
Bwd Packet Length Min	Ελάχιστο μέγεθος πακέτου πίσω.
Bwd Packet Length Mean	Μέσο μέγεθος πακέτου πίσω.
Bwd Packet Length Std	Τυπική απόκλιση μεγέθους πακέτων πίσω.
Packet Length Min	Ελάχιστο μήκος πακέτου στη ροή.
Packet Length Max	Μέγιστο μήκος πακέτου στη ροή.
Packet Length Mean	Μέσο μήκος πακέτου στη ροή.
Packet Length Std	Τυπική απόκλιση μήκους πακέτου.
Packet Length Variance	Διακύμανση μεγέθους πακέτων.
Average Packet Size	Μέσο μέγεθος πακέτου (bytes).
Avg Fwd Segment Size	Μέσο μέγεθος TCP segment εμπρός.
Avg Bwd Segment Size	Μέσο μέγεθος TCP segment πίσω.

Πίνακας 3.3 – Χρονικά Χαρακτηριστικά (Time-based Features)

Feature	Περιγραφή
Flow IAT Mean	Μέσο χρονικό διάστημα μεταξύ πακέτων.
Flow IAT Std	Τυπική απόκλιση χρονικού διαστήματος μεταξύ πακέτων.
Flow IAT Max	Μέγιστο χρονικό διάστημα μεταξύ πακέτων.
Flow IAT Min	Ελάχιστο χρονικό διάστημα μεταξύ πακέτων.
Fwd IAT Total	Συνολικό χρονικό διάστημα μεταξύ πακέτων εμπρός.
Fwd IAT Mean	Μέσο χρονικό διάστημα μεταξύ πακέτων εμπρός.
Fwd IAT Std	Τυπική απόκλιση χρονικού διαστήματος μεταξύ πακέτων εμπρός.
Fwd IAT Max	Μέγιστο χρονικού διαστήματος μεταξύ πακέτων εμπρός.
Fwd IAT Min	Ελάχιστο χρονικού διαστήματος μεταξύ πακέτων εμπρός.
Bwd IAT Total	Συνολικός χρόνος χρονικού διαστήματος μεταξύ πακέτων πίσω.
Bwd IAT Mean	Μέσο χρονικού διαστήματος μεταξύ πακέτων πίσω.
Bwd IAT Std	Τυπική απόκλιση χρονικού διαστήματος μεταξύ πακέτων πίσω.
Bwd IAT Max	Μέγιστο χρονικού διαστήματος μεταξύ πακέτων πίσω.
Bwd IAT Min	Ελάχιστο χρονικού διαστήματος μεταξύ πακέτων πίσω.
Fwd Packets/s	Πακέτα ανά δευτερόλεπτο εμπρός.
Bwd Packets/s	Πακέτα ανά δευτερόλεπτο πίσω.
Active Mean	Μέσος χρόνος ενεργής φάσης επικοινωνίας.
Active Std	Τυπική απόκλιση ενεργών χρόνων.
Active Max	Μέγιστος χρόνος ενεργότητας.
Active Min	Ελάχιστος χρόνος ενεργότητας.
Idle Mean	Μέσος χρόνος αδράνειας.
Idle Std	Τυπική απόκλιση χρόνου αδράνειας.
Idle Max	Μέγιστος χρόνος αδράνειας.
Idle Min	Ελάχιστος χρόνος αδράνειας.

Πίνακας 3.4 – Χαρακτηριστικά Πρωτοκόλλου & TCP Flags

Feature	Περιγραφή
Fwd PSH Flags	Αριθμός TCP PSH flags εμπρός.
Bwd PSH Flags	Αριθμός TCP PSH flags πίσω.
Fwd URG Flags	Αριθμός TCP URG flags εμπρός.
Bwd URG Flags	Αριθμός TCP URG flags πίσω.

Fwd RST Flags	Αριθμός TCP RST flags εμπρός.
Bwd RST Flags	Αριθμός TCP RST flags πίσω.
Fwd Header Length	Μήκος TCP header εμπρός.
Bwd Header Length	Μήκος TCP header πίσω.
FIN Flag Count	Συνολικός αριθμός TCP FIN flags.
SYN Flag Count	Συνολικός αριθμός TCP SYN flags.
RST Flag Count	Συνολικός αριθμός TCP RST flags.
PSH Flag Count	Συνολικός αριθμός TCP PSH flags.
ACK Flag Count	Συνολικός αριθμός TCP ACK flags.
URG Flag Count	Συνολικός αριθμός TCP URG flags.
CWR Flag Count	Συνολικός αριθμός TCP CWR flags.
ECE Flag Count	Συνολικός αριθμός TCP ECE flags.
Down/Up Ratio	Αναλογία πακέτων download/upload.
FWD Init Win Bytes	Αρχικό TCP Window size εμπρός.
Bwd Init Win Bytes	Αρχικό TCP Window size πίσω.
Fwd Act Data Pkts	Αριθμός ενεργών data packets εμπρός.
Fwd Seg Size Min	Ελάχιστο TCP segment size εμπρός.
ICMP Code	ICMP κωδικός.
ICMP Type	ICMP τύπος (π.χ. Echo Request/Reply).

Πίνακας 3.5 – Bulk & Subflow Features

Feature	Περιγραφή
Fwd Bytes/Bulk Avg	Μέσος όρος bytes ανά bulk εμπρός.
Fwd Packet/Bulk Avg	Μέσος όρος πακέτων ανά bulk εμπρός.
Fwd Bulk Rate Avg	Μέσος ρυθμός bulk μεταφοράς εμπρός.
Bwd Bytes/Bulk Avg	Μέσος όρος bytes ανά bulk πίσω.
Bwd Packet/Bulk Avg	Μέσος όρος πακέτων ανά bulk πίσω.
Bwd Bulk Rate Avg	Μέσος ρυθμός bulk μεταφοράς πίσω.
Subflow Fwd Packets	Πακέτα εμπρός ανά subflow.
Subflow Fwd Bytes	Bytes εμπρός ανά subflow.
Subflow Bwd Packets	Πακέτα πίσω ανά subflow.
Subflow Bwd Bytes	Bytes πίσω ανά subflow.

Πίνακας 3.6 – Label

Feature	Περιγραφή
Label	Ετικέτα κατηγορίας (π.χ. BENIGN ή συγκεκριμένος τύπος επίθεσης).

(Ενότητα 3.2.β Ειδή Επιθέσεων)

Στην παρούσα ενότητα θα παρουσιαστούν οι κατηγορίες επιθέσεων που περιλαμβάνονται στα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018. Κάθε επίθεση έχει διαφορετικό τρόπο λειτουργίας, στόχο και πρωτόκολλο, ωστόσο όλες αφήνουν χαρακτηριστικά μοτίβα στην κίνηση του δικτύου. Για την μελέτη μας, κάθε επίθεση αναλύεται ξεχωριστά μέσα από πίνακες που περιλαμβάνουν, την κατηγορία στην οποία ανήκει η επίθεση, το βασικό πρωτόκολλο επικοινωνίας που χρησιμοποιείται, μια περιγραφή ως προς τι εκδηλώνει η κάθε επίθεση , ποια τεχνική χρησιμοποιεί και για ποιο στόχο. Για τις επιθέσεις που βρεθήκαν πληροφορίες συμπεριλαμβάνεται και το εργαλείο που έχει χρησιμοποιηθεί στο πλαίσιο εκτέλεσή της.

DDoS επίθεση

Η επίθεση DDoS (Distributed Denial of Service – κατανεμημένη άρνηση υπηρεσίας) στοχεύει σε τοποθεσίες web και διακομιστές, διακόπτοντας τις υπηρεσίες δικτύου σε μια προσπάθεια εξάντλησης των πόρων [28]. Κατά τη διάρκεια μιας επίθεσης DDoS, ένα σύνολο από bot ή botnet [48], κατακλύζουν τον στόχο με υπερβολικά μεγάλο όγκο αιτημάτων HTTP και δεδομένων.

Ουσιαστικά, πολλοί υπολογιστές επιτίθενται ταυτόχρονα σε έναν, με αποτέλεσμα η υπηρεσία να καθυστερεί ή να διακόπτεται για συγκεκριμένο χρονικό διάστημα. Ανάλογα με το εργαλείο και το πρωτόκολλο, οι DDoS επιθέσεις μπορούν να εμφανιστούν σε διάφορες μορφές, όπως HTTP floods, UDP floods, TCP floods [22]. Κάθε παραλλαγή παρουσιάζει δικά της χαρακτηριστικά μοτίβα κίνησης, τα οποία είναι εμφανή σε flow-based σύνολα δεδομένων και αξιοποιούνται για την ανίχνευσή τους [13-15].

Οι επιθέσεις DDoS στοχεύουν κάθε είδους κλάδο, καθώς και εταιρείες παγκοσμίως, ενώ συχνά μπορούν να διαρκέσουν ώρες ή ακόμα και ημέρες. Σε ορισμένες περιπτώσεις, οι εισβολείς εκμεταλλεύονται τη διάρκεια της επίθεσης για να διεισδύσουν σε βάσεις δεδομένων και να αποκτήσουν πρόσβαση σε ευαίσθητες πληροφορίες [28,48].

LOIC (Low Orbit Ion Cannon)

Το εργαλείο LOIC είναι ένα ανοιχτού κώδικα πρόγραμμα αρχικά σχεδιασμένο για stress testing*, το οποίο στη συνέχεια χρησιμοποιήθηκε ευρέως από την ομάδα Anonymous και άλλους χρηστές σε κυβερνοεπιθέσεις [19]. Το εργαλείο δημιουργεί έναν τεράστιο όγκο από HTTP GET/POST αιτήματα προς τον στόχο, χωρίς να αναμένει απάντηση [18]. Ο web server δέχεται υπερβολικό φορτίο, εξαντλώντας τους πόρους του (CPU, μνήμη, sockets) καθιστώντας την υπηρεσία μη διαθέσιμη στους νόμιμους χρήστες [22].

HOIC (High Orbit Ion Cannon)

Το HOIC είναι ένα εργαλείο ανοιχτού κώδικα που δημιουργήθηκε ως εναλλακτική/βελτίωση του LOIC [20]. Χρησιμοποιήθηκε ευρέως για πιο εξελιγμένες και αυτοματοποιημένες επιθέσεις, πλημμυρίζοντας τον στόχο με HTTP requests σε πολύ υψηλότερο ρυθμό [21]. Μπορεί να ρυθμιστεί ώστε να στοχεύει πολλούς διακομιστές ταυτόχρονα. Σε αντίθεση με το LOIC, συνδυάζει booster scripts (σκριπτάκια επιτάχυνσης) που παράγουν διαφοροποιημένα headers/παραμέτρους, με αποτέλεσμα το παραγόμενο traffic να είναι πιο δυσδιάκριτο

* *Είναι η διαδικασία κατά την οποία ένας διαχειριστής ή ερευνητής φορτώνει εσκεμμένα ένα σύστημα, δίκτυο ή εφαρμογή με πολύ υψηλή κίνηση/αιτήματα, ώστε να ελεγχθεί πώς ανταποκρίνεται σε ακραίες συνθήκες.*

καθώς εισάγεται τυχαία παραλλαγή στα αιτήματα [21]. Παρόλα αυτά, το τελικό αποτέλεσμα είναι ουσιαστικά το ίδιο με του LOIC στοχεύοντας την υπερφόρτωση του διακομιστή με τεράστιο όγκο αιτημάτων μέχρι η υπηρεσία να γίνει μη διαθέσιμη [22].

DDoS LOIC HTTP

Πεδίο	**Περιγραφή**
Επίθεση	DDoS LOIC HTTP
Κατηγορία	Distributed Denial of Service (DDoS)
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Αποστολή μαζικών HTTP GET/POST requests με το εργαλείο LOIC. Κάθε αίτημα μοιάζει με κανονικό HTTP request και τα αιτήματα στέλνονται σε υπερβολικό όγκο.
Στόχος	Υπερφόρτωση web server και εξάντληση CPU memory και πόρων.
Εργαλείο	LOIC (Low Orbit Ion Cannon)

DDoS HOIC

Πεδίο	**Περιγραφή**
Επίθεση	DDoS HOIC
Κατηγορία	Distributed Denial of Service (DDoS)
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Εξελιγμένο εργαλείο (HOIC) που δημιουργεί υψηλό αριθμό HTTP αιτημάτων, χρησιμοποιώντας randomization για αποφυγή caching και ευκολότερη παράκαμψη φίλτρων.
Στόχος	Αποδιοργάνωση υπηρεσιών ιστού με υπερβολική κίνηση.
Εργαλείο	HOIC (High Orbit Ion Cannon)

DDoS LOIC UDP

Πεδίο	**Περιγραφή**
Επίθεση	DDoS LOIC UDP
Κατηγορία	Distributed Denial of Service (DDoS)
Πρωτόκολλο	UDP (17)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Αποστολή τεράστιου όγκου UDP packets σε τυχαίες ή συγκεκριμένες θύρες με το LOIC. Δεν απαιτείται σύνδεση, άρα ο όγκος μπορεί να είναι πολύ υψηλός.
Στόχος	Υπερφόρτωση bandwidth και εξάντληση πόρων στον στόχο.
Εργαλείο	LOIC (Low Orbit Ion Cannon)

DoS επίθεση

Οι επιθέσεις DoS (Denial of Service) στοχεύουν στον αποσυντονισμό της λειτουργίας ενός συστήματος (server) ή μίας υπηρεσίας μέσω της υπερφόρτωσή του με συνεχή αιτημάτα [28]. Η βασική τους αρχή είναι η αποστολή μεγάλου όγκου πακέτων ή αιτημάτων (όπως HTTP, TCP, ICMP) από μόνο μια πηγή, με σκοπό την εξάντληση πόρων όπως CPU, μνήμη ή διαθέσιμα sockets [22]. Ως αποτέλεσμα, ο server καθίσταται αδύνατο να εξυπηρετήσει

νόμιμους χρήστες. Η βασική διαφορά από τις επιθέσεις DDoS (Distributed Denial of Service) είναι ότι στις DoS η επίθεση πραγματοποιείται από μία μόνο πηγή, ενώ στις DDoS η επίθεση γίνεται ταυτόχρονα από πολλαπλά συστήματα (botnets) [48].
Σε πειραματικό πλαίσιο, οι επιθέσεις DoS εκτελέστηκαν από ένα σύστημα Kali Linux το οποίο χρησιμοποιήθηκε ως πλατφόρμα εκτέλεσης εργαλείων DoS, με στόχο έναν Ubuntu 16.04 server που λειτουργούσε με Apache web server [8,10,11].

Slowhttptest: Το εργαλείο slowhttptest είναι διαθέσιμο στο Kali Linux και υποστηρίζει διάφορους τύπους slow HTTP επιθέσεων (π.χ. slowloris, slow Header, slow Body, slow Read). Χρησιμοποιείται για stress testing και παρέχει ρύθμιση για το είδος της slow HTTP επίθεσης που θα εκτελεστεί. Το εργαλείο προσομοιώνει πολλαπλές κακόβουλες συνδέσεις οι οποίες διατηρούνται ενεργές και καταναλώνουν πόρους του server, οδηγώντας σε εξάντληση sockets/μνήμης/επεξεργαστή και τελικά, σε μη διαθεσιμότητα της υπηρεσίας. [29]

Slowloris: Η Slowloris είναι μια επίθεση η οποία μπορεί να υλοποιηθεί με δυο μέθοδους είτε με παραμέτρους του slowhttptest [29] είτε με ένα ξεχωριστό script (slowloris.pl*) που εκτελεί αποκλειστικά αυτόν τον τύπο slow HTTP [23]. Το script διατηρεί πολλές ανοιχτές HTTP συνδέσεις αποστέλλοντας τα headers αργά, με σκοπό να αποτρέψει τον διακομιστή από το κλείσιμο των συνδέσεων. Το Canadian Institute for Cybersecurity (CIC/UNB) δεν διευκρινίζει εάν σε όλες τις δοκιμές χρησιμοποιήθηκε το slowhttptest με κατάλληλες παραμέτρους ή το ξεχωριστό script slowloris.pl.

Hulk: Το Hulk (HTTP Unbearable Load King) είναι ένα script γραμμένο σε Python που δημιουργεί τεράστιο αριθμό HTTP GET αιτήματα προς τον web server. Στόχος του είναι να εξαντλήσει τους πόρους του server (CPU, μνήμη, sockets) μέσω διαδοχικών αιτημάτων που χρησιμοποιούν διαφορετικά URLs. Σε αντίθεση με άλλα εργαλεία DoS, το Hulk παράγει παραπλανητικά HTTP requests με διαφοροποιημένα headers, URLs και user-agents, έτσι ώστε ο server να τα αντιμετωπίζει ως καινούρια. Με αυτόν τον τρόπο δεν μπορεί να αξιοποιήσει την cache (έτοιμες απαντήσεις που ελαφραίνουν τον φόρτο) και επεξεργάζεται ξεχωριστά κάθε αίτημα. Επιπλέον, η επίθεση μιμείται νόμιμη κίνηση, καθιστώντας την ανίχνευση από συστήματα άμυνας πιο δύσκολη. [24]

GoldenEye: Είναι είναι ένα script γραμμένο σε Python, αλλά διατίθεται και ως ενσωματωμένο εργαλείο HTTP DoS testing στο Kali Linux, χρησιμοποιούμενο για τον έλεγχο αδυναμίας ενός ιστότοπου σε επιθέσεις Denial of Service (DoS). Η λειτουργία του βασίζεται στο άνοιγμα πολλαπλών παράλληλων συνδέσεων προς μια συγκεκριμένη διεύθυνση URL, με στόχο την υπερφόρτωση ή την αποδιαθεσιμότητα του web server. Ως βασικό μηχανισμό επίθεσης χρησιμοποιεί τεχνικές στα headers, όπως HTTP Keep-Alive και no-cache, ώστε να καταναλώνει σταδιακά τους διαθέσιμους πόρους του server. [30]

DoS Slowhttptest

Πεδίο	**Περιγραφή**
Επίθεση	DoS Slowhttptest
Κατηγορία	Denial of Service (DoS)

* Στην ιστοσελίδα του Canadian Institute for Cybersecurity αναφέρεται πως για την slowloris στα CSE-CICIDS 2018 δεδομένα χρησιμοποιήθηκαν scripts γραμμένα σε Perl.

Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Εργαλείο που στέλνει HTTP headers πολύ αργά, διατηρώντας τις συνδέσεις ανοιχτές (no timeout) και δεσμεύοντας πόρους.
Στόχος	Εξάντληση συνδέσεων web server χωρίς την δυνατότητά για νέες συνδέσεις.
Εργαλείο	SlowHTTPTest

DoS Hulk

Πεδίο	**Περιγραφή**
Επίθεση	DoS Hulk
Κατηγορία	Denial of Service (DoS)
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Μαζικά HTTP requests με συνεχώς διαφορετικές παραμέτρους ώστε να υπερφορτωθούν οι cache μηχανισμοί.
Στόχος	Υπερφόρτωση web server με ψεύτικο traffic.
Εργαλείο	HULK script

DoS GoldenEye

Πεδίο	**Περιγραφή**
Επίθεση	DoS GoldenEye
Κατηγορία	Denial of Service (DoS)
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	HTTP DoS εργαλείο που στέλνει επαναλαμβανόμενα requests με τυχαία headers.
Στόχος	Κατάρρευση web υπηρεσίας λόγω μεγάλης πίεσης.
Εργαλείο	GoldenEye

DoS Slowloris

Πεδίο	**Περιγραφή**
Επίθεση	DoS Slowloris
Κατηγορία	Denial of Service (DoS)
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Κρατάει πολλές TCP συνδέσεις ανοιχτές στέλνοντας αργά HTTP headers χωρίς να τις κλείνει.
Στόχος	Δέσμευση όλων των διαθέσιμων TCP συνδέσεων του web server.
Εργαλείο	Slowloris script ή παράμετρός Slowhttptest εργαλείου

Brute Force επίθεση

Οι επιθέσεις Brute Force βασίζονται στην επαναλαμβανόμενη δοκιμή διαφορετικών συνδυασμών ονομάτων χρήστη και κωδικών πρόσβασης με σκοπό την παραβίαση ενός λογαριασμού [33]. Ο επιτιθέμενος δεν εκμεταλλεύεται κάποιο κενό ασφαλείας στο

λογισμικό, αλλά την αδυναμία του συστήματος να αντιμετωπίσει μαζικές, αυτοματοποιημένες προσπάθειες σύνδεσης [34]. Εάν ο κωδικός είναι αδύναμος ή το σύστημα δεν διαθέτει κατάλληλους μηχανισμούς προστασίας, η επιτυχία της επίθεσης είναι πολύ πιθανή.
Στο πλαίσιο των δεδομένων της παρούσας μελέτης, οι επιθέσεις brute force εκτελέστηκαν από σύστημα Kali Linux, με στόχο τον web server Ubuntu 16.04. Αυτή η ρύθμιση προσομοιώνει ρεαλιστικά σενάρια επίθεσης και επιτρέπει την αξιολόγηση μεθόδων ανίχνευσης και άμυνας [10,11].

SSH Brute Force: Η SSH brute force επίθεση αποτελεί μία από τις πιο συνηθισμένες μορφές brute force, καθώς στοχεύει το πρωτόκολλο SSH (Secure Shell), το οποίο χρησιμοποιείται για απομακρυσμένη πρόσβαση σε συστήματα [25]. Ο επιτιθέμενος προσπαθεί να μαντέψει τον σωστό συνδυασμό χρήστη και κωδικού (username/password) στέλνοντας διαδοχικές αιτήσεις σύνδεσης στον SSH server. Δοκιμάζει συστηματικά και αυτοματοποιημένα πολλούς συνδυασμούς, συνήθως με απλό script που πραγματοποιεί χιλιάδες προσπάθειες ανά δευτερόλεπτο, εκμεταλλευόμενο αδύναμους κωδικούς ή ελλιπή μέτρα προστασίας [26].
Στην παρούσα εργασία τα δεδομένα κίνησης για αυτήν την επίθεση συλλέχθηκαν με χρήση Kali Linux, αξιοποιώντας το εργαλείο Patator το οποίο διαθέτει ενσωματωμένη παράμετρο για SSH brute-force [32,35].

FTP Brute Force: Η FTP brute-force επίθεση στοχεύει το πρωτόκολλο FTP (File Transfer Protocol), το οποίο χρησιμοποιείται για τη μεταφορά αρχείων μεταξύ client και server. Η ταχύτητα και η αυτοματοποίηση με την οποία δοκιμάζονται τα στοιχεία πρόσβασης (credentials) μπορεί να οδηγήσει σε μαζικές προσπάθειες μέχρι να εντοπιστεί ο σωστός κωδικός [27]. Ο επιτιθέμενος χρησιμοποιεί εργαλεία όπως το Patator στο Kali Linux, τα οποία εκτελούν αυτοματοποιημένες, συνεχείς προσπάθειες εισόδου στον FTP server με διαφορετικούς συνδυασμούς username/password [32,35].

Διαφορές SSH και FTP: Οι κύριες διαφορές μεταξύ τους εντοπίζονται στα πρωτόκολλα και στις θύρες προορισμού, όπως φαίνεται παρακάτω. Αυτό καθορίζει και τις δυνατότητες που αποκτά ο επιτιθέμενος μετά την επιτυχημένη σύνδεση. Στην περίπτωση του SSH, ο εισβολέας αποκτά πλήρη έλεγχο του συστήματος και μπορεί να εκτελέσει εντολές ή να εγκαταστήσει κακόβουλο λογισμικό, επηρεάζοντας ολόκληρο το λειτουργικό σύστημα. Αντίθετα, στην περίπτωση του FTP, ο επιτιθέμενος αποκτά κυρίως πρόσβαση στα αρχεία του διακομιστή: μπορεί να κατεβάσει δεδομένα, να ανεβάσει κακόβουλο περιεχόμενο ή να τροποποιήσει/διαγράψει αρχεία. [35]

SSH-Brute Force-Patator

Πεδίο	**Περιγραφή**
Επίθεση	SSH Brute Force Patator
Κατηγορία	Brute Force Attack
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	22 SSH
Περιγραφή	Αυτοματοποιημένες προσπάθειες δοκιμής πολλών συνδυασμών username/password μέσω SSH για μη εξουσιοδοτημένη πρόσβαση.
Στόχος	Απόκτηση πρόσβασης σε SSH λογαριασμούς.
Εργαλείο	Patator

FTP-Brute Force-Patator

Πεδίο	Περιγραφή
Επίθεση	FTP Brute Force Patator
Κατηγορία	Brute Force Attack
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	21 FTP
Περιγραφή	Επαναλαμβανόμενες δοκιμές password στο FTP πρωτόκολλο με χρήση Patator μέχρι να βρεθεί σωστός συνδυασμός.
Στόχος	Απόκτηση πρόσβασης σε FTP λογαριασμούς/αρχεία.
Εργαλείο	Patator

Web Attacks

Οι Web attacks είναι επιθέσεις που στοχεύουν εφαρμογές web, όπως ιστοσελίδες, διεπαφές προγραμματισμού εφαρμογών (APIs), φόρμες εισόδου (login forms) και βάσεις δεδομένων μέσω του Internet. Αντί να στοχεύουν χαμηλά επίπεδα του στοίβας δικτύου (π.χ. TCP, UDP), επικεντρώνονται στο επίπεδο εφαρμογής HTTP/HTTPS. [34]

Στα δεδομένα της παρούσας μελέτης αναφέρεται* ότι χρησιμοποιήθηκε ένα εσωτερικό πλαίσιο εργασίας Selenium (in-house Selenium framework) για αυτοματοποιημένες επιθέσεις, καθώς και η ευάλωτη εφαρμογή Damn Vulnerable Web Application (DVWA), η οποία προορίζεται ειδικά για εκπαιδευτικά και ερευνητικά σενάρια ασφάλειας web. [36]

Web Attack – Brute Force (HTTP): Η Brute Force HTTP είναι επίθεση που στοχεύει web εφαρμογές μέσω login σελίδων [34]. Ο επιτιθέμενος δοκιμάζει μαζικά συνδυασμούς username και password με αυτοματοποιημένα εργαλεία [33]. Εάν ο server δεν διαθέτει κατάλληλους μηχανισμούς προστασίας, ο επιτιθέμενος ενδέχεται να αποκτήσει παράνομη πρόσβαση σε λογαριασμούς με έγκυρα στοιχεία πρόσβασης. Χαρακτηριστικά της ροής μετά από τέτοια επίθεση είναι πολλαπλά, επαναλαμβανόμενα HTTP POST/GET αιτήματα προς την ίδια σελίδα σύνδεσης, χαμηλός όγκος δεδομένων ανά αίτημα (καθώς στέλνονται μόνο τα στοιχεία σύνδεσης) και υψηλή συχνότητα αποτυχημένων απαντήσεων σύνδεσης [36].

Για την αυτοματοποιημένη δοκιμή στοιχείων πρόσβασης στην παρούσα μελέτη χρησιμοποιήθηκε ένα εσωτερικό πλαίσιο με Selenium (in-house Selenium framework), το οποίο εκτέλεσε τις προσπάθειες πάνω στην ευάλωτη εφαρμογή DVWA (Damn Vulnerable Web Application) [36,8].

Διαφορά HTTP από SSH/FTP Brute Force:

Τα SSH και FTP brute-force attacks στοχεύουν υπηρεσίες απομακρυσμένης πρόσβασης, ενώ το HTTP brute-force στοχεύει εφαρμογές web (κυρίως φόρμες σύνδεσης , login forms). [33,34]

Web Attack SQL Injection: Η SQL Injection (SQLi) είναι από τις πιο γνωστές και επικίνδυνες επιθέσεις σε web εφαρμογές. Ο επιτιθέμενος εκμεταλλεύεται πεδία εισαγωγής δεδομένων

* Η αναφορά βρίσκεται στην σελίδα του Canadian Institute for Cybersecurity για τα δεδομένα του CSE-CICIDS 2018 και στο σχετικό ερευνητικό άρθρο που αναφέρεται στη σελίδα των CICIDS 2017 δεδομένων.

(όπως φόρμες login, πλαίσια αναζήτησης ή παραμέτρους σε URL) [37], τα οποία στέλνουν πληροφορίες κατευθείαν στη βάση δεδομένων χωρίς να ελέγχονται σωστά. Με αυτόν τον τρόπο, εισάγεται κακόβουλος SQL κώδικας ώστε να αποκτηθεί πρόσβαση χωρίς τα σωστά στοιχεία, να εξαχθούν δεδομένα ή να τροποποιηθούν εγγραφές.
Η SQL Injection επίθεση στοχεύει DVWA εφαρμογή μέσω HTTP requests. Ο επιτιθέμενος (Kali) χρησιμοποιεί Selenium scripts για να στείλει κακόβουλα SQL payloads στον Ubuntu Web Server [8,36].

Web Attack XSS: Η επίθεση web XSS (Cross-Site Scripting) πραγματοποιείται με την εισαγωγή κακόβουλου script σε μια ιστοσελίδα που κανονικά θεωρείται ασφαλής και αξιόπιστη. Αυτό συμβαίνει όταν μια web εφαρμογή λαμβάνει δεδομένα από τον χρήστη και τα εμφανίζει στη σελίδα χωρίς να τα ελέγξει ή να τα κωδικοποιήσει σωστά. Το κακόβουλο script εκτελείται όχι στον server αλλά στον περιηγητή (browser) του θύματος. Ο browser δεν μπορεί να διακρίνει αν το script είναι ασφαλές και το εκτελεί καθώς φαίνεται να προέρχεται από την ίδια την ιστοσελίδα.[38]
Με αυτόν τον τρόπο ο επιτιθέμενος μπορεί να κλέψει cookies και session tokens ή άλλες ευαίσθητες πληροφορίες που αποθηκεύει ο browser για τον συγκεκριμένο ιστότοπο. Επίσης μπορεί να αλλοιώσει το περιεχόμενο που βλέπει ο χρήστης, για παράδειγμα να εμφανίσει ψεύτικες φόρμες, να προβάλλει κακόβουλα links ή να τροποποιήσει τη σελίδα με σκοπό phishing ή περαιτέρω εκμετάλλευση.[38]

Web Attack - Brute Force (HTTP)

Πεδίο	**Περιγραφή**
Επίθεση	Web Attack Brute Force
Κατηγορία	Brute Force (Web Login)
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Επαναλαμβανόμενα login attempts σε web εφαρμογές (π.χ. φόρμες login) με σκοπό την εύρεση σωστών credentials.
Στόχος	Παράκαμψη authentication μηχανισμών.
Εργαλείο	(βλ. Παραπάνω)

Web Attack - SQL Injection

Πεδίο	**Περιγραφή**
Επίθεση	Web Attack SQL Injection
Κατηγορία	Web Application Attack
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Εισαγωγή κακόβουλων SQL εντολών σε web forms,URLs για πρόσβαση σε βάση δεδομένων ή αλλοίωση δεδομένων.
Στόχος	Παράκαμψη ελέγχων, ανάκτηση ή αλλοίωση δεδομένων .
Εργαλείο	(βλ. Παραπάνω)

Web Attack – XSS

Πεδίο	**Περιγραφή**
Επίθεση	Web Attack XSS (Cross-Site Scripting)

Κατηγορία	Web Application Attack
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	80 (HTTP)
Περιγραφή	Εισαγωγή κακόβουλου κώδικα σε web εφαρμογές που εκτελείται στον browser των χρηστών.
Στόχος	Κλοπή cookies,session, redirection, αλλαγή μορφής ιστοσελίδας.
Εργαλείο	(βλ. Παραπάνω)

Portscan επίθεση

Η κατηγορία Portscan περιλαμβάνει επιθέσεις αναγνώρισης (reconnaissance attacks), κατά τις οποίες ο επιτιθέμενος επιχειρεί να εντοπίσει ποιες θύρες και υπηρεσίες είναι ενεργές στους στόχους του [45]. Ο επιτιθέμενος αποστέλλει στοχευμένα πακέτα (TCP, UDP ή ICMP) προς ένα εύρος θυρών [44], όταν μια θύρα είναι ανοικτή, επιστρέφει απάντηση, ενώ αν είναι κλειστή επιστρέφεται μήνυμα σφάλματος ή δεν υπάρχει απάντηση. Σε περίπτωση φιλτραρισμένης θύρας, η απάντηση αποκλείεται από firewall. Στόχος της σάρωσης θυρών είναι ο εντοπισμός ανοιχτών θυρών και υπηρεσιών ώστε να χαρτογραφηθούν πιθανοί δρόμοι εισχώρησης [44,45].

Στα δεδομένα της παρούσας μελέτης, η σάρωση πραγματοποιήθηκε με το εργαλείο Nmap χρησιμοποιώντας συνδυασμό παραμέτρων (switches) για την παραγωγή ποικίλων μοτίβων κίνησης [43]. Ο επιτιθέμενος στο σενάριο εμφανίζεται ως μηχάνημα Kali Linux, αν και σε κάποιες περιγραφές αναφέρεται και Windows 8.1 ως πηγή σάρωσης, πιθανότατα λόγω διαφορετικών δοκιμών ή τυπογραφικών ασυνεπειών. Το θύμα ήταν ένας web server με Ubuntu 16. Κατά τη διάρκεια των δοκιμών, οι κανόνες firewall αλλάζανε σε προκαθορισμένα χρονικά διαστήματα για να προσομοιωθούν διαφορετικά σενάρια δικτυακής κίνησης, με στόχο τον εμπλουτισμό του συνόλου δεδομένων με ποικίλες παραλλαγές σάρωσης [10,11,8].

Portscan

Πεδίο	**Περιγραφή**
Επίθεση	Portscan
Κατηγορία	Reconnaissance Attack
Πρωτόκολλο	TCP (6), UDP (17), ICMP (1)
Θύρα Προορισμού	Ποικίλουν.
Περιγραφή	Σάρωση θυρών του μηχανήματος με χρήση πολλών τύπων scan.
Στόχος	Εισχώρηση σε σύστημα ή μηχάνημα.
Εργαλείο	Nmap (με switches: -sS, -sT, -sF, -sX, -sN, -sP, -sV, -sU, -sO, -sA, -sW, -sR, -sL, -B)

Μερικά από τα switches που χρησιμοποιήθηκαν είναι:

SYN Scan : -sS , TCP Connect Scan : -sT , FIN Scan : -sF , XMAS Scan : -sX ,
NULL Scan : -sN , UDP Scan : -sU , Ping Scan : -sP , Version Detection : -sV

Infiltration επίθεση

Η Infiltration αναφέρεται σε επιθέσεις όπου ο επιτιθέμενος καταφέρνει να διεισδύσει σε ένα εσωτερικό δίκτυο, συνήθως ξεκινώντας από έναν χρήστη που άθελά του κατεβάζει ή εκτελεί ένα κακόβουλο αρχείο από το διαδίκτυο. Με το άνοιγμα ή την εκτέλεση του κακόβουλου προγράμματος, ο επιτιθέμενος αποκτά πρόσβαση σε εσωτερικούς πόρους του οργανισμού.

Συγκεκριμένα, μπορεί να εγκαταστήσει backdoors [39] ή να εξάγει δεδομένα χωρίς να γίνει αντιληπτός. Τα backdoors είναι κακόβουλα πρόγραμματα/agents που επιτρέπουν στον εισβολέα να επικοινωνεί με τον υπολογιστή και να μπαίνει αργότερα χωρίς να χρειάζεται ο χρήστης να κάνει login. Η επικοινωνία με τον επιτιθέμενο γίνεται συνήθως μέσω ενός Command-and-Control (C2) server [41]. Στις προσομοιώσεις που χρησιμοποιήθηκαν για τα datasets, χρησιμοποιήθηκε το Metasploit για να αναπαραστήσει τον επιτιθέμενο [40].

Infiltration Dropbox Download: Η επίθεση Infiltration ξεκινά με τη λήψη ενός κακόβουλου αρχείου από την υπηρεσία Dropbox, η οποία επιλέγεται επειδή θεωρείται ευρέως νόμιμη και αξιόπιστη. Επομένως, το κατέβασμα φαίνεται φυσιολογικό και δύσκολα ανιχνεύεται από τα συστήματα ασφαλείας [12]. Παρόλο που αυτό το σενάριο ενδέχεται να μην θεωρηθεί ως ξεχωριστός τύπος επίθεσης αλλά μάλλον ως μέθοδος εκμετάλλευσης αξιόπιστων υπηρεσιών cloud [42], η παραγόμενη δικτυακή κίνηση ενδέχεται να παρουσιάζει χαρακτηριστικά που τη διαφοροποιούν από τη φυσιολογική χρήση και να μπορεί να χρησιμοποιηθεί για ανίχνευση.

Infiltration NMAP Portscan*: Η κατηγορία Infiltration NMAP Portscan στα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018 αντιπροσωπεύει το δεύτερο στάδιο επίθεσης, όπου ο επιτιθέμενος έχει ήδη αποκτήσει πρόσβαση σε έναν υπολογιστή του δικτύου μέσω infiltration. Εκτελεί το εργαλείο αναγνώρισης Nmap, για σάρωση θυρών (port scans) σε ολόκληρο το δίκτυο του θύματος (victim network) προκειμένου να εντοπίσει πιθανές ευάλωτες υπηρεσίες και να προχωρήσει στα επόμενα βήματα της επίθεσης. [8,12,43]

Infiltration Communication Victim Attacker: Η κατηγορία Infiltration Communication Victim Attacker στα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018 αντιπροσωπεύει το τελικό στάδιο της επίθεσης Infiltration [8,12]. Έπειτα από την αρχική είσοδο του επιτιθέμενου σε έναν εσωτερικό υπολογιστή και την αναγνώριση του δικτύου μέσω Nmap Portscan [43], ο επιτιθέμενος καθιερώνει μια σταθερή επικοινωνία με το μολυσμένο θύμα. Η επικοινωνία αυτή μπορεί να είναι Command-and-Control (C2) όπου ο υπολογιστής θύμα στέλνει δεδομένα πίσω στον επιτιθέμενο, λαμβάνει νέες εντολές για επόμενες ενέργειες ή χρησιμοποιείται σαν ενδιάμεσος κόμβος για την εξάπλωση της επίθεσης [41].

Infiltration - Dropbox Download

Πεδίο	**Περιγραφή**
Επίθεση	Infiltration – Dropbox Download
Κατηγορία	Infiltration Attack
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	443

* Η κατηγορία επίθεσης Infiltration NMAP Portscan είναι ουσιαστικά ταυτόσημη με την κατηγορία Portscan. Οι διαφορές μεταξύ τους δεν αφορούν το πρότυπο της σάρωσης θυρών αλλά τον τρόπο ταξινόμησης στα δεδομένα. Οι δύο ετικέτες διαχωρίστηκαν αποκλειστικά για πειραματικούς σκοπούς: η μία αντιπροσωπεύει το πρωτότυπο δείγμα Portscan, ενώ η άλλη προέκυψε στο πλαίσιο περιστατικού Infiltration και επισημάνθηκε ξεχωριστά ώστε να μελετηθεί πώς η παρουσία εισβολής επηρεάζει τις ίδιες τεχνικές σάρωσης. Συνεπώς, από πλευράς συμπεριφοράς δικτύου και χαρακτηριστικών ροής, δεν υπάρχει ουσιαστική διαφορά.

Περιγραφή	Λήψη κακόβουλου λογισμικού
Στόχος	Εγκατάσταση κακόβουλου λογισμικού στο θύμα μέσω ενός φαινομενικά νόμιμου καναλιού επικοινωνίας.
Εργαλείο	Dropbox

Infiltration - NMAP Portscan

Πεδίο	**Περιγραφή**
Επίθεση	Infiltration – NMAP Portscan
Κατηγορία	Infiltration Attack
Πρωτόκολλο	ICMP (1), TCP (6), UDP (17)
Θύρα Προορισμού	Ποικίλουν
Περιγραφή	Σάρωση θυρών .
Στόχος	Εντοπισμός θυρών για μελλοντική επίθεση.
Εργαλείο	Nmap (Network Mapper)

Infiltration - Communication Victim Attacker

Πεδίο	**Περιγραφή**
Επίθεση	Infiltration – Communication Victim Attacker
Κατηγορία	Infiltration Attack
Πρωτόκολλο	TCP (6)
Θύρα Προορισμού	31337
Περιγραφή	Πρόκειται για φάση επικοινωνίας μεταξύ μολυσμένου θύματος και επιτιθέμενου. Η επικοινωνία αυτή χρησιμοποιείται για ανταλλαγή εντολών ή για εξαγωγή δεδομένων.
Στόχος	Κρυφή επικοινωνία του θύματος με τον επιτιθέμενο για έλεγχο του συστήματος.
Εργαλείο	Metasploit

Heartbleed

Η επίθεση Heartbleed προσομοιώθηκε με το εργαλείο Heartleech [31]. Το Heartleech ελέγχει αν ένα σύστημα είναι ευάλωτο αποστέλλοντας αιτήματα heartbeat προς τον server [46]. Στα πειραματικά σενάρια, ο server είχε εγκατεστημένη μια παλαιότερη και ευάλωτη έκδοση του OpenSSL σε περιβάλλον Ubuntu 12.04, η οποία περιείχε το γνωστό σφάλμα Heartbleed. Κατά την εκτέλεση της επίθεσης, το Heartleech προκάλεσε την επιστροφή τμημάτων μνήμης από τη διεργασία του web server. Λόγω του σφάλματος στον χειρισμό των heartbeat αιτημάτων από το OpenSSL, ο server ανταποκρινόταν επιστρέφοντας περισσότερα δεδομένα από τα προβλεπόμενα [8,47]. Ως αποτέλεσμα, αποκαλύπτονταν τμήματα μνήμης που μπορούσαν να περιέχουν ευαίσθητες πληροφορίες, όπως κωδικούς πρόσβασης, cookies ή κλειδιά κρυπτογράφησης [8].

Heartbleed

Πεδίο	**Περιγραφή**
Επίθεση	Heartbleed
Κατηγορία	Exploit (Application Layer)
Πρωτόκολλο	6 (TCP)
Θύρα Προορισμού	444 (HTTPS)

Περιγραφή	Ο επιτιθέμενος στέλνει heartbeat αιτήματα για να διαβάσει τη μνήμη του server.
Στόχος	Απόσπαση ευαίσθητων δεδομένων (passwords, κλειδιά, cookies) από μνήμη συστήματος.
Εργαλείο	Heartleech

Botnet επίθεση

Ένα botnet είναι ένα δίκτυο μολυσμένων υπολογιστών (zombies) που ελέγχονται κεντρικά από τον επιτιθέμενο (botmaster). Οι μολυσμένοι κόμβοι εκτελούν εντολές χωρίς γνώση των νόμιμων χρηστών και μπορούν να χρησιμοποιηθούν για μαζικές επιθέσεις, κλοπή δεδομένων ή εξάπλωση κακόβουλου λογισμικού. Η μόλυνση συνήθως εξαπλώνεται μέσω phishing emails, κακόβουλων downloads ή μολυσμένων ιστοσελίδων. Το μολυσμένο μηχάνημα συνδέεται με έναν server Command-and-Control (C2), μέσω του οποίου λαμβάνει εντολές και αποστέλλει δεδομένα, χωρίς την αντίληψη του θύματος [42]. Στόχος ενός botnet είναι να παρέχει στον επιτιθέμενο μαζικό και απομακρυσμένο έλεγχο συστημάτων για κακόβουλες δραστηριότητες όπως DDoS, κλοπή δεδομένων και εξάπλωση malware.

Στα σύνολα δεδομένων CICIDS2017 και CSE-CICIDS2018, αν και στην τελική περιγραφή αναφέρονται τόσο τα botnets Zeus όσο και Ares, οι ροές που περιλαμβάνονται στα δημοσιευμένα δεδομένα προέρχονται από το Ares Botnet [8]. Το Ares είναι ένα κακόβουλο λογισμικό ανοιχτού κώδικα γραμμένο σε Python που επιτρέπει στον επιτιθέμενο απομακρυσμένο έλεγχο μολυσμένων υπολογιστών [48]. Στις ενέργειες που καταγράφηκαν περιλαμβάνονται μεταφορτώσεις/λήψεις αρχείων, λήψη screenshots, keylogging και χρήση απομακρυσμένου shell (cmd.exe). Στα σενάρια του dataset, τα θύματα επικοινωνούσαν με τον C2 server, αποστέλλοντας δεδομένα και λαμβάνοντας εντολές· επιπλέον, κάθε 400 δευτερόλεπτα αποστέλλονταν αίτημα για λήψη screenshot [8,48].

Botnet Ares

Πεδίο	**Περιγραφή**
Επίθεση	Botnet Ares
Κατηγορία	Botnet
Πρωτόκολλο	6 (TCP)
Θύρα Προορισμού	8080
Περιγραφή	Δίκτυο μολυσμένων συστημάτων botnets που επικοινωνούν με κακόβουλο server.
Στόχος	Μαζική μόλυνση για κακόβουλες δραστηριότητες όπως DDoS, κ.α.
Εργαλείο	Ares Botnet

ΚΕΦΑΛΑΙΟ 4 Επεξεργασία & Εμπλουτισμός Δεδομένων

(Υποκεφάλαιο 4.1) Εφαρμογή GAN και VAE για τεχνητό εμπλουτισμό δεδομένων πινάκων.

Σε αυτό το Κεφάλαιο περιγράφονται οι διαδικασίες που εφαρμόστηκαν για την ενίσχυση του αρχικού συνόλου δεδομένων μέσω τεχνικών δημιουργίας συνθετικών δειγμάτων. Συγκεκριμένα, αξιοποιήθηκαν γενετικά μοντέλα βασισμένα σε διαφορετικές αρχιτεκτονικές, GANs (Generative Adversarial Networks), καθώς και VAEs (Variational Autoencoders). Σκοπός του εμπλουτισμού αυτού είναι η αντιμετώπιση της ανισορροπίας μεταξύ των κατηγοριών και η διερεύνηση της απόδοσης των μοντέλων σε περιβάλλον όπου η πρόβλεψη πραγματοποιείται αποκλειστικά σε συνθετικά δεδομένα.

(Ενότητα 4.1.α Εισαγωγή στα GANs και VAE)

Τα Generative Adversarial Networks (GANs) αποτελούν μια κατηγορία μοντέλων βαθιάς μάθησης που αναπτύχθηκαν από τον Ian Goodfellow το 2014 με σκοπό τη δημιουργία νέων, ρεαλιστικών δεδομένων [53]. Η βασική τους αρχή στηρίζεται στην αντιπαλότητα δύο νευρωνικών δικτύων: του γεννήτορα (generator) και του κριτή (discriminator). Ο γεννήτορας επιχειρεί να παράγει δείγματα τα οποία προσομοιάζουν τα πραγματικά δεδομένα, ενώ ο κριτής έχει ως στόχο να διακρίνει αν ένα δείγμα είναι αυθεντικό ή συνθετικό [50]. Η διαδικασία εκπαίδευσης μπορεί να θεωρηθεί ως μια συνεχή αντιπαράθεση μεταξύ των δύο δικτύων, όσο ο γεννήτορας βελτιώνεται στην παραγωγή δειγμάτων που προσομοιάζουν την πραγματικότητα, τόσο πιο απαιτητικό γίνεται το έργο του κριτή. Όταν επιτευχθεί ισορροπία, ο γεννήτορας έχει αποκτήσει την ικανότητα να παράγει δεδομένα τα οποία είναι ιδιαίτερα δύσκολο να διακριθούν από τα αυθεντικά. Χάρη σε αυτή τη δυναμική, τα GANs έχουν γνωρίσει ευρεία εφαρμογή σε πολλούς τομείς, όπως η δημιουργία εικόνων και μουσικής, αλλά και στην τεχνητή ενίσχυση δεδομένων (data augmentation) [55], όπως διερευνάται και στην παρούσα εργασία.

Τα Variational Autoencoders (VAEs) αποτελούν μια διαφορετική κατηγορία γενετικών μοντέλων βαθιάς μάθησης, τα οποία προτάθηκαν από τους Kingma και Welling το 2013 [54]. Στόχος τους είναι η δημιουργία νέων δειγμάτων που προσομοιάζουν τα αρχικά δεδομένα, αξιοποιώντας μια αρχιτεκτονική βασισμένη σε δύο βασικά τμήματα: τον Encoder και τον Decoder [49]. Ο Encoder συμπιέζει τα δεδομένα εισόδου σε μια χαμηλότερης διάστασης αναπαράσταση, γνωστή ως latent space, η οποία συνήθως προσεγγίζει μια κατανομή Gaussian. Στη συνέχεια, ο Decoder λαμβάνει δείγματα από τον latent space και επιχειρεί να τα ανακατασκευάσει σε μορφή παρόμοια με τα αρχικά δεδομένα, επιτυγχάνοντας έτσι την εκμάθηση μιας συνεχούς και κανονικοποιημένης αναπαράστασης [54]. Τα VAEs μπορούν να αξιοποιηθούν όχι μόνο για τη δημιουργία συνθετικών δεδομένων, όπως και τα GANs, αλλά και για εργασίες όπως η συμπίεση δεδομένων, η μείωση διάστασης και η ανίχνευση ανωμαλιών υπό συγκεκριμένες προϋποθέσεις [54].

(Ενότητα 4.1.β Εφαρμογή GAN για Εμπλουτισμό Δεδομένων)

Στο πλαίσιο της παρούσας εργασίας αξιοποιήθηκαν δύο γεννήτριες συνθετικών δεδομένων που παρέχονται από τη βιβλιοθήκη SDV (Synthetic Data Vault): ο CTGANSynthesizer και ο TVAESynthesizer [56,57]. Ο TVAE (Tabular Variational Autoencoder) αποτελεί ένα

παραμετρικό μοντέλο βασισμένο στην αρχιτεκτονική των Variational Autoencoders, ειδικά προσαρμοσμένο για την παραγωγή συνθετικών δεδομένων πινάκων [57,58], όπως εκείνων που χρησιμοποιούνται στην ανάλυση ροών δικτύου. Αντίθετα, ο CTGAN (Conditional Tabular GAN) στηρίζεται στην αρχιτεκτονική των Generative Adversarial Networks και έχει σχεδιαστεί ειδικά για τη δημιουργία συνθετικών δειγμάτων σε μορφή πινάκων. [51,52].

Στην παρούσα εργασία, ο TVAE χρησιμοποιήθηκε για την παραγωγή επιπλέον δειγμάτων (υπερδειγματοληψία) σε κατηγορίες επιθέσεων με περιορισμένο αριθμό αρχικών παρατηρήσεων, με σκοπό την ενίσχυση του συνόλου εκπαίδευσης και την επίτευξη μεγαλύτερης ισορροπίας μεταξύ των κατηγοριών [57]. Από την άλλη πλευρά, ο CTGAN αξιοποιήθηκε για την παραγωγή συνθετικών δεδομένων που χρησιμοποιήθηκαν αποκλειστικά στο στάδιο της πρόβλεψης, με στόχο την αξιολόγηση της ικανότητας του μοντέλου να ανταποκρίνεται σε δεδομένα τα οποία παρουσιάζουν ελάχιστες αποκλίσεις από τα πραγματικά [52]. Ο διαχωρισμός αυτός βασίστηκε τόσο στα διαφορετικά αποτελέσματα που παρείχαν οι δύο γεννήτριες όσο και στις διαφοροποιήσεις στον τρόπο λειτουργίας τους. [58,61,62].

Ο TVAE θεωρείται συχνά πιο αξιόπιστο μοντέλο για τη δημιουργία συνθετικών δεδομένων σε σύγκριση με τον CTGAN, καθώς δεν αποθηκεύει ακριβή αντίγραφα των χαρακτηριστικών, αλλά μαθαίνει την υποκείμενη κατανομή των δεδομένων, αποθηκεύοντας τη μέση τιμή (mean) και τη διασπορά (variance) [57,58]. Με τον τρόπο αυτό, κάθε δεδομένο αναπαρίσταται στον latent space ως κατανομή και όχι ως μεμονωμένο σημείο. Επιπλέον, ο TVAE χρησιμοποιεί συνάρτηση απώλειας η οποία αποτελείται από δύο όρους: την reconstruction loss, που εξασφαλίζει ότι τα παραγόμενα δείγματα προσεγγίζουν με πιστότητα τα αρχικά, και την KL divergence, η οποία επιβάλλει στον latent space να ακολουθεί μια προκαθορισμένη ομαλή κατανομή [58]. Η κανονικοποίηση αυτή οδηγεί σε καλύτερη γενίκευση του μοντέλου και συμβάλλει στην αποφυγή υπερπροσαρμογής , κάτι που δεν επιτυγχάνεται στον ίδιο βαθμό με τον CTGAN. Ως αποτέλεσμα, ο TVAE υπερέχει σε σταθερότητα και σε στατιστική ομοιότητα των κατανομών που παράγει σε σχέση με τον CTGAN [58].

Ο CTGAN παράγει συνθετικά δείγματα εφαρμόζοντας mode-specific normalization, το οποίο βασίζεται σε Variational Gaussian Mixture (VGM) model με Bayesιανή προσέγγιση [51]. Η τεχνική αυτή επιτρέπει τον εντοπισμό των επιμέρους κορυφών της κατανομής κάθε αριθμητικού χαρακτηριστικού και την κανονικοποίηση των τιμών με τρόπο που διατηρεί και τις μικρότερες κορυφές, αποφεύγοντας έτσι την αλλοίωση πολύπλοκων ή ασύμμετρων κατανομών [52].Η λειτουργία του μοντέλου στηρίζεται σε έναν Generator, ο οποίος εκπαιδεύεται να παράγει δεδομένα που να ξεγελούν τον Discriminator, σύμφωνα με την αρχή των GANs [51].

Ωστόσο, ένα από τα βασικά μειονεκτήματα του CTGAN είναι ότι, σε περιπτώσεις περιορισμένου όγκου δεδομένων ή υπερβολικού αριθμού εποχών εκπαίδευσης, ο Generator ενδέχεται να απομνημονεύσει δείγματα του αρχικού συνόλου αντί να μάθει την υποκείμενη κατανομή [63,64]. Αυτό μπορεί να οδηγήσει στην παραγωγή συνθετικών δειγμάτων που μοιάζουν υπερβολικά με συγκεκριμένα παραδείγματα του training set, μειώνοντας την ικανότητα γενίκευσης του μοντέλου [64].

Παρότι το φαινόμενο αυτό θα μπορούσε να θεωρηθεί ως πιθανό μειονέκτημα, στην παρούσα εργασία δεν τίθεται ζήτημα, καθώς έχουν ενσωματωθεί έλεγχοι αξιοπιστίας για τον εντοπισμό ακριβών αντιγράφων μεταξύ των αρχικών και των συνθετικών συνόλων δεδομένων [59].

Η αξιολόγηση της ποιότητας των παραγόμενων δειγμάτων πραγματοποιήθηκε με τη χρήση μετρικών όπως το KS test (Kolmogorov–Smirnov), η Ευκλείδεια απόσταση, η διαφορά μέσων τιμών, η RMSD (Root Mean Square Deviation) και το ποσοστό επαναλήψεων (duplicate ratio) [60]. Αφού επιβεβαιώθηκε η στατιστική ομοιότητα και η ποικιλία των συνθετικών δεδομένων σε σχέση με τα πραγματικά, τα δείγματα αυτά ενσωματώθηκαν στο σύνολο εκπαίδευσης και χρησιμοποιήθηκαν για τη βελτίωση της απόδοσης των μοντέλων.

Σημαντικές ορολογίες κεφαλαίου:

Kolmogorov–Smirnov (KS test): Ο έλεγχος Kolmogorov–Smirnov (KS test) αποτελεί μια στατιστική μέθοδο για τη σύγκριση δύο κατανομών, με σκοπό την εκτίμηση του βαθμού ομοιότητάς τους [65,66]. Στην παρούσα εργασία εφαρμόστηκε η εκδοχή two-sample KS test, η οποία υπολογίζει τη μέγιστη απόσταση μεταξύ των συναρτήσεων κατανομής (CDF) δύο δειγμάτων για κάθε χαρακτηριστικό. Τιμές του δείκτη κοντά στο 0 υποδηλώνουν υψηλή ομοιότητα, ενώ τιμές κοντά στο 1 δηλώνουν σημαντική απόκλιση [67]. Για λόγους καλύτερης ερμηνείας, χρησιμοποιήθηκε η τιμή 1 – KS score, ώστε όσο μεγαλύτερη είναι η τελική τιμή τόσο μεγαλύτερη να είναι η ομοιότητα των δύο συνόλων.
Επιπλέον, για κάθε σύνολο δεδομένων υπολογίστηκε ο μέσος όρος των επιμέρους τιμών ανά χαρακτηριστικό, αλλά και μια συνολική μέση τιμή, ώστε να παρέχεται μια συνοπτική εικόνα της συνολικής ομοιότητας [65]. Ο συγκεκριμένος έλεγχος επιλέχθηκε λόγω της ικανότητάς του να ανιχνεύει ακόμη και μικρές διαφορές στην κατανομή των δεδομένων, γεγονός που τον καθιστά ιδιαίτερα κατάλληλο για την αξιολόγηση της ποιότητας της διαδικασίας τεχνητού εμπλουτισμού του συνόλου δεδομένων [66].

Euclidean distance: Η Ευκλείδεια απόσταση (Euclidean distance) αποτελεί ένα από τα βασικότερα μέτρα απόστασης σε πολυδιάστατους χώρους και χρησιμοποιείται για την εκτίμηση της ομοιότητας ή της απόκλισης μεταξύ δύο σημείων [68]. Ορίζεται ως η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των διαφορών των αντίστοιχων συντεταγμένων τους [69].
Στην παρούσα εργασία, τα πραγματικά και τα συνθετικά δεδομένα κανονικοποιήθηκαν αρχικά μέσω MinMaxScaler. Για κάθε χαρακτηριστικό υπολογίστηκε η απόλυτη διαφορά των μέσων τιμών (μονοδιάστατο ισοδύναμο της Ευκλείδειας απόστασης). Για συνολική εκτίμηση σε όλα τα χαρακτηριστικά, υπολογίστηκε η Ευκλείδεια απόσταση μεταξύ των διανυσμάτων μέσων τιμών των δύο συνόλων. Μικρότερες τιμές και των δύο μετρικών υποδηλώνουν μεγαλύτερη ομοιότητα μεταξύ πραγματικών και συνθετικών δεδομένων. Επιπλέον, ως συνοπτικό δείκτη αναφοράς υπολογίστηκε και ο μέσος όρος των επιμέρους απολύτων διαφορών ανά χαρακτηριστικό.

RMSD: Ο Root Mean Squared Difference (RMSD) αποτελεί μέτρο που εκτιμά τη μέση τετραγωνική απόκλιση μεταξύ αντίστοιχων τιμών δύο συνόλων δεδομένων [70]. Στην παρούσα εργασία, μετά την κανονικοποίηση των δεδομένων με MinMaxScaler, υπολογίστηκε αρχικά το Mean Squared Difference (MSD) συγκρίνοντας τις μέσες τιμές και τις διακυμάνσεις κάθε χαρακτηριστικού μεταξύ πραγματικών και συνθετικών δεδομένων [71]. Συγκεκριμένα, για κάθε χαρακτηριστικό υπολογίστηκε η διαφορά των μέσων τιμών, η οποία υψώθηκε στο τετράγωνο και προστέθηκε στο άθροισμα των διακυμάνσεων των δύο συνόλων. Η τετραγωνική ρίζα της ποσότητας αυτής έδωσε την τιμή του RMSD για το εκάστοτε χαρακτηριστικό [70].

Η διαδικασία εφαρμόστηκε σε όλα τα χαρακτηριστικά, ώστε να εκτιμηθεί η απόκλιση μεταξύ πραγματικών και συνθετικών δεδομένων ανά χαρακτηριστικό, και στη συνέχεια υπολογίστηκε ο μέσος όρος όλων των RMSD τιμών για μια συνολική εκτίμηση. Μικρότερες τιμές RMSD υποδηλώνουν υψηλή ομοιότητα, ενώ μεγαλύτερες τιμές αναδεικνύουν σημαντικότερη απόκλιση. [70,71]

Duplicate ratio: Το ποσοστό επαναλήψεων (duplicate ratio) χρησιμοποιήθηκε ως μέτρο για την εκτίμηση της ποικιλομορφίας των συνθετικών δεδομένων. Για τον υπολογισμό του αφαιρέθηκε αρχικά η στήλη ετικέτας (label) τόσο από τα πραγματικά όσο και από τα συνθετικά δεδομένα.
Στον εξωτερικό έλεγχο (external duplicate ratio) αφαιρέθηκαν τα διπλότυπα από το σύνολο των πραγματικών δεδομένων και πραγματοποιήθηκε σύζευξη (inner join) με το σύνολο των συνθετικών, ώστε να εντοπιστούν οι κοινές εγγραφές που εμφανίζονται και στα δύο σύνολα. Ο λόγος του αριθμού αυτών των κοινών εγγραφών προς το συνολικό πλήθος των συνθετικών δειγμάτων αποτέλεσε την τελική τιμή του δείκτη.
Στον εσωτερικό έλεγχο (internal duplicate ratio) υπολογίστηκε ο αριθμός των διπλοτύπων μέσα στο ίδιο το συνθετικό σύνολο δεδομένων. Ο λόγος του αριθμού αυτών των επαναλαμβανόμενων εγγραφών προς το συνολικό πλήθος των συνθετικών δειγμάτων αντιστοιχεί στην τιμή του δείκτη.
Χαμηλές τιμές στον εξωτερικό δείκτη υποδηλώνουν ότι τα συνθετικά δείγματα δεν αποτελούν απλά αντίγραφα των πραγματικών και συνεπώς το μοντέλο δεν υπέπεσε σε υπερπροσαρμογή (overfitting). Αντίστοιχα, χαμηλές τιμές στον εσωτερικό δείκτη υποδηλώνουν υψηλή ποικιλομορφία εντός του ίδιου του συνθετικού συνόλου.

MinMaxScale: Η κανονικοποίηση με MinMaxScaler αποτελεί μέθοδο μετασχηματισμού δεδομένων κατά την οποία οι τιμές κάθε χαρακτηριστικού (feature) αναπροσαρμόζονται σε μία κοινή κλίμακα, συνήθως στο διάστημα [0,1]. Ο μετασχηματισμός βασίζεται στις ελάχιστες και μέγιστες τιμές του εκάστοτε χαρακτηριστικού, με αποτέλεσμα η ελάχιστη τιμή να αντιστοιχεί στο 0 και η μέγιστη στο 1, ενώ όλες οι ενδιάμεσες τιμές κλιμακώνονται αναλογικά. [72]
Στην παρούσα εργασία, ο MinMaxScaler εφαρμόστηκε τόσο στα πραγματικά όσο και στα συνθετικά δεδομένα, ώστε να βρίσκονται στην ίδια κλίμακα και να είναι άμεσα συγκρίσιμα. Η διαδικασία αυτή ήταν απαραίτητη για τον ακριβή υπολογισμό μετρικών όπως η Euclidean distance και το RMSD, οι οποίες επηρεάζονται από την κλίμακα των δεδομένων.

CDF (Cumulative Distribution Function): Η Σωρευτική Συνάρτηση Κατανομής (Cumulative Distribution Function – CDF) εκφράζει την πιθανότητα μια τυχαία μεταβλητή X να λάβει τιμή μικρότερη ή ίση από μια συγκεκριμένη τιμή x, δηλαδή $F(x) = P(X \leq x)$ [66]. Οι τιμές της CDF κυμαίνονται από 0 έως 1, ξεκινώντας από το 0 και τείνοντας προς το 1. [66]
Για να υπολογιστεί, οι τιμές της μεταβλητής ταξινομούνται σε αύξουσα σειρά και για κάθε σημείο εκτιμάται το ποσοστό των δειγμάτων που δεν την υπερβαίνουν. Με αυτόν τον τρόπο προκύπτει μια αυξανόμενη συνάρτηση, η οποία ξεκινά από 0 και τείνει προς 1, περιγράφοντας πώς κατανέμονται σωρευτικά οι πιθανότητες σε όλο το εύρος των τιμών. Η συνάρτηση που προκύπτει αντιπροσωπεύει την CDF.
Στο πλαίσιο του KS test, συγκρίνεται η CDF των πραγματικών δεδομένων με εκείνη των συνθετικών, και η μέγιστη διαφορά μεταξύ τους χρησιμοποιείται ως μέτρο απόκλισης

[65,66]. Με βάση αυτή τη μεθοδολογία είναι δυνατή η αξιολόγηση τόσο των ομοιοτήτων όσο και των διαφορών ανάμεσα στα παραγόμενα και τα πραγματικά δεδομένα.

Μαθηματικές υλοποιήσεις των ορών και τα σημεία υλοποίησης τους στον κώδικα.

Μαθηματική Υλοποίηση (Min-Max Scaler)

Για να είναι συγκρίσιμα τα πραγματικά και συνθετικά δεδομένα ανά χαρακτηριστικό, εφαρμόζεται κλιμάκωση Min–Max σε [0,1].
Για κάθε τιμή x ενός χαρακτηριστικού:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Σημείο υλοποίησης στον κώδικα:

```
scaler = MinMaxScaler()
realScaled = scaler.fit_transform(dfData[numericCols])
#ένας πίνακας NumPy με κανονικοποιημένες τιμές για κάθε feature και κάθε δείγμα
synthScaled = scaler.transform(syntheticData[numericCols])
```

Μαθηματική Υλοποίηση (Euclidean distance)

Για τον υπολογισμό της Ευκλείδειας απόστασης, αρχικά υπολογίστηκε η μέση τιμή κάθε χαρακτηριστικού τόσο για το πραγματικό όσο και για το συνθετικό σύνολο δεδομένων. Οι τιμές αυτές χρησιμοποιήθηκαν στη συνέχεια για τον υπολογισμό της απόστασης μεταξύ των δύο συνόλων.

Για τα πραγματικά δεδομένα:

$$r_i = \frac{\text{Άθροισμα όλων των τιμών του χαρακτηριστικού i στα πραγματικά δεδομένα}}{\text{Πλήθος δειγμάτων πραγματικών δεδομένων}}$$

Για τα συνθετικά δεδομένα:

$$s_i = \frac{\text{Άθροισμα όλων των τιμών του χαρακτηριστικού i στα συνθετικά δεδομένα}}{\text{Πλήθος δειγμάτων συνθετικών δεδομένων}}$$

Δεδομένου ότι η Ευκλείδεια απόσταση ορίζεται γενικά για πολυδιάστατα διανύσματα (δηλαδή για τον συνδυασμό πολλών χαρακτηριστικών), στην περίπτωση ενός μόνο χαρακτηριστικού η μετρική αυτή απλοποιείται και ισοδυναμεί με την απόλυτη διαφορά των μέσων τιμών. Συνεπώς, για κάθε χαρακτηριστικό η απόσταση υπολογίστηκε ως:

$$d_i = |r_i - s_i|$$

Απολυτή διαφορά ανά χαρακτηριστικό μεταξύ των μέσων τιμών των πραγματικών και των συνθετικών δεδομένων.

Σημείο υλοποίησης στον κώδικα:

```
meanReal = realScaled.mean(axis=0)
meanSynth = synthScaled.mean(axis=0)
absDiffPerFeature = np.abs(meanReal - meanSynth)
```

Συνολική (μέση) απόλυτη διαφορά των μέσων τιμών όλων των χαρακτηριστικών μεταξύ πραγματικών και συνθετικών δεδομένων.

Σημείο υλοποίησης στον κώδικα:

```
Absolute Mean Difference Overall": round(float(absDiffPerFeature.mean()), 6)
```

Ευκλείδεια απόσταση μεταξύ των διανυσμάτων των μέσων τιμών των πραγματικών και συνθετικών δεδομένων.

$$d_E = \sqrt{\sum_{i=1}^{n} (r_i - s_i)^2}$$

Σημείο υλοποίησης στον κώδικα:

```
euclidean = float(np.linalg.norm(meanReal - meanSynth, ord=2))# L2 norm (Ευκλείδεια)
```

Μαθηματική Υλοποίηση (RMSD)

Για τον υπολογισμό της μέσης τετραγωνικής απόκλισης (Mean Squared Deviation , MSD), αρχικά υπολογίστηκε η μέση τιμή κάθε χαρακτηριστικού, καθώς και η αντίστοιχη διακύμανση. Η διακύμανση ενός συνόλου m ισοπίθανων τιμών ορίζεται ως ο μέσος όρος των τετραγώνων των αποκλίσεων κάθε τιμής από τη μέση τιμή, όπως φαίνεται και στον παρακάτω τύπο:

$$\mathrm{Var_R}(i) = \frac{\sum_{j=1}^{m} \left(R'_{j,i} - r_i\right)^2}{m}$$

,όπου $R'_{j,i}$είναι η τιμή του χαρακτηριστικού i στο j-οστό δείγμα των πραγματικών δεδομένων, r_i η μέση τιμή του χαρακτηριστικού στα πραγματικά δεδομένα και m το πλήθος των δειγμάτων.

$$\mathrm{Var_S}(i) = \frac{\sum_{j=1}^{m} \left(S'_{j,i} - s_i\right)^2}{m}$$

Η διακύμανση ενός χαρακτηριστικού i στα πραγματικά δεδομένα μπορεί να εκφραστεί και πιο απλά ως:

$$\mathrm{Var_R(i)} = \frac{\text{Άθροισμα των τετράγωνων των διαφόρων των τιμών του χαρακτηριστικού i από τη μέση τιμή στα πραγματικά δεδομένα}}{\text{Πλήθος δειγμάτων πραγματικών δεδομένων}}$$

$$\mathrm{Var_s}(i) = \frac{\text{Άθροισμα των τετράγωνων των διαφόρων των τιμών του χαρακτηριστικού i από τη μέση τιμή στα συνθετικά δεδομένα}}{\text{Πλήθος δειγμάτων συνθετικών δεδομένων}}$$

Σημείο υλοποίησης στον κώδικα:

```
varReal = realScaled.var(axis=0)
varSynth = synthScaled.var(axis=0)
```

Από τον ορισμό, η μέση τετραγωνική διαφορά (MSE) μεταξύ δύο τυχαίων μεταβλητών X και Y είναι:

$$\mathrm{MSE} = E[(X - Y)^2]$$

,όπου E είναι η μαθηματική προσδοκία.

Στα θεωρητικά μαθηματικά όταν μιλάμε για expectation (προσδοκία), αναφερόμαστε το σταθμισμένο μέσο όλων των πιθανών τιμών μιας μεταβλητής.

Η ανάπτυξη τετράγωνου δίνεται ως:

$$(X - Y)^2 = X^2 - 2XY + Y^2$$

Επομένως :

$$\mathrm{E}[(\mathrm{X} - \mathrm{Y})^2] = \mathrm{E}[\mathrm{X}^2] - 2\mathrm{E}[\mathrm{XY}] + \mathrm{E}[\mathrm{Y}^2]$$

Ο ορισμός της διακύμανσης μίας μεταβλητής είναι:

$$\mathrm{Var}(\mathrm{X}) = \mathrm{E}[\mathrm{X}^2] - (\mathrm{E}[\mathrm{X}])^2$$

$$\Leftrightarrow$$

$$\mathrm{E}[\mathrm{X}^2] = \mathrm{Var}(\mathrm{X}) + (\mathrm{E}[\mathrm{X}])^2$$

$$\mathrm{Var}(\mathrm{Y}) = \mathrm{E}[\mathrm{Y}^2] - (\mathrm{E}[\mathrm{Y}])^2$$

$$\Leftrightarrow$$

$$E[Y^2] = \mathrm{Var}(Y) + (E[Y])^2$$

Αντικαθιστώντας στον MSE τύπο προκύπτει:

$$\mathrm{MSE} = [\mathrm{Var}(\mathrm{X}) + (\mathrm{E}[\mathrm{X}])^2] - 2\mathrm{E}[\mathrm{X}Y] + [\mathrm{Var}(\mathrm{Y}) + (\mathrm{E}[\mathrm{Y}])^2]$$

Υποθέτοντας ότι οι μεταβλητές X και Y είναι ανεξάρτητες ισχύει:

$$\mathrm{E}[\mathrm{XY}] = \mathrm{E}[\mathrm{X}] \cdot \mathrm{E}[\mathrm{Y}]$$

Επομένως ο παραπάνω τύπος γίνεται:

$$\mathrm{MSE} = \mathrm{Var}(\mathrm{X}) + (\mathrm{E}[\mathrm{X}])^2 - 2\mathrm{E}[\mathrm{X}]\mathrm{E}[\mathrm{Y}] + \mathrm{Var}(\mathrm{Y}) + (\mathrm{E}[\mathrm{Y}])^2$$

Από τον τύπο ανάπτυξης τετράγωνου:

$$(\mathrm{E}[\mathrm{X}])^2 - 2\mathrm{E}[\mathrm{X}]\mathrm{E}[\mathrm{Y}] + (\mathrm{E}[\mathrm{Y}])^2 = (\mathrm{E}[\mathrm{X}] - \mathrm{E}[\mathrm{Y}])^2$$

Πλέον ο τελικός τύπος είναι:

$$\mathrm{MSE} = \mathrm{Var}(\mathrm{X}) + \mathrm{Var}(\mathrm{Y}) + (\mathrm{E}[\mathrm{X}] - \mathrm{E}[\mathrm{Y}])^2$$

Σημείο υλοποίησης στον κώδικα:

```
meanSquaredDiff = varReal + varSynth + (meanReal - meanSynth) ** 2
rootMeanSquaredDiff = np.sqrt(meanSquaredDiff)
```

Για τον συνοπτικό δείκτη του RMSD υπολογίστηκε το μέσο όλων των RMSD τιμών.

Σημείο υλοποίησης στον κώδικα:

```
"RMSD Mean": round(float(rootMeanSquaredDiff.mean()), 6)
```

(Υποκεφάλαιο 4.2) Παράμετροι και αποτελέσματα CTGAN και TVAE

(Ενότητα 4.2.α Ανάλυση Παραμέτρων CTGAN και TVAE)

Ανάλυση Παραμέτρων CTGAN

Παρακάτω αναλύονται οι παράμετροι που χρησιμοποιήθηκαν στην υλοποίηση του κώδικα, για την παραγωγή συνθετικών δεδομένων με την χρήση CTGAN Synthesizer.[56]

Παράμετρος generator_dim

Η παράμετρος generator_dim καθορίζει το μέγεθος και το βάθος των κρυφών επιπέδων του Generator, δηλαδή του δικτύου που παράγει τα συνθετικά δεδομένα

Παράμετρος discriminator_dim

Η παράμετρος discriminator_dim καθορίζει τη δομή του Discriminator, δηλαδή του δικτύου που ξεχωρίζει τα πραγματικά από τα συνθετικά δείγματα. Για να υπάρχει ισορροπία ορίζεται συμμετρικά με του generator_dim. Έτσι ο Discriminator είναι εξίσου ισχυρός ώστε να αναγνωρίζει τα λάθη του Generator, αλλά όχι υπερβολικά ώστε να απορρίπτει κάθε παραγόμενο δείγμα και να υπάρξει αστάθεια.

Παράμετρος pac

Το pac ορίζει πόσα δείγματα ομαδοποιούνται και περνάνε ταυτόχρονα στον Discriminator. Με αυτόν τον τρόπο μειώνεται το φαινόμενο του mode collapse, δηλαδή η παραγωγή επαναλαμβανόμενων και μη ρεαλιστικών δειγμάτων.

Παράμετρος batch_size

Η παράμετρος batch_size καθορίζει πόσα δείγματα θα επεξεργάζεται το δίκτυο σε κάθε βήμα εκπαίδευσης. Το batch_size προσαρμόζεται πάντα ώστε να είναι πολλαπλάσιο του pac, καθώς έτσι απαιτείται από τον μηχανισμό του.

Ανάλυση Παραμέτρων TVAE

Παρακάτω αναλύονται οι παράμετροι που χρησιμοποιήθηκαν στην εργασία αυτή, για την παραγωγή συνθετικών δεδομένων με την χρήση TVAE Synthesizer.[57]

Παράμετρος embedding_dim

Η παράμετρος embedding_dim καθορίζει τη διάσταση του latent space, δηλαδή το μέγεθος του χώρου στον οποίο τα δεδομένα συμπιέζονται πριν αναπαραχθούν. Όσο μεγαλύτερος είναι αυτός ο χώρος-διάσταση, τόσο περισσότερη πληροφορία μπορεί να αποθηκεύσει το μοντέλο για τα δεδομένα.

Παράμετρος compress_dims

Η παράμετρος compress_dims ορίζει πόσα επίπεδα και πόσους νευρώνες θα έχει ο encoder, αφορά το τμήμα που συμπιέζει τα δεδομένα στον latent χώρο.

Παράμετρος decompress_dims

Η παράμετρος decompress_dims είναι το αντίστροφο του encoder, δηλαδή ορίζει τη δομή του decoder, του τμήματος που παίρνει τα δεδομένα από τον latent space και αναπαράγει τα συνθετικά. Ορίζεται συμμετρικά με τον compress_dims, ώστε να υπάρχει ισορροπία στη συμπίεση των δεδομένων (με τον encoder) και στην αποκωδικοποίησή τους (με τον decoder).

Παράμετρος batch_size

Η παράμετρος batch_size στον TVAE παρουσιάζει ίδιο ρόλο όπως και στο CTGAN, δηλαδή καθορίζει πόσα δείγματα θα χρησιμοποιηθούν σε κάθε βήμα εκπαίδευσης. Με μια βασική διαφορά ότι στα Variational Autoencoders (VAE) δεν εφαρμόζεται η παράμετρος pac, άρα το batch size δεν χρειάζεται να προσαρμόζεται.

Πίνακές με τιμές ανά παράμετρο

Οι παράμετροι προσαρμόζονται δυναμικά ανάλογα με το μέγεθος του dataset, ώστε να εξασφαλιστεί σταθερότητα στην εκπαίδευση, να αποφευχθεί η υπερπροσαρμογή και να βελτιωθεί η ικανότητα του μοντέλου να παράγει ρεαλιστικά και ποικίλα συνθετικά δεδομένα.

Πίνακας-CTGAN

Μέγεθος Δεδομένων (n)	pac	generator_dim	discriminator_dim	batch_size
n < 100	1	(32,)	(32,)	min(4, n)
100 ≤ n < 1,000	2	(128, 128)	(128, 128)	128
1,000 ≤ n < 10,000	4	(256, 256)	(256, 256)	256
10,000 ≤ n < 100,000	6	(256, 256, 128)	(256, 256, 128)	512
100,000 ≤ n < 1,000,000	8	(512, 512, 256)	(512, 512, 256)	1024
n ≥ 1,000,000	10	(1024, 1024, 512)	(1024, 1024, 512)	2048

Πίνακας-TVAE

Μέγεθος Δεδομένων (n)	embedding_dim	compress_dims	decompress_dims	batch_size
n < 100	16	(16,)	(16,)	min(4, n)
100 ≤ n < 1,000	32	(64,)	(64,)	128
1,000 ≤ n < 10,000	64	(128, 64)	(64, 128)	256

10,000 ≤ n < 100,000	128	(256, 128)	(128, 256)	512
100,000 ≤ n < 1,000,000	128	(512, 256, 128)	(128, 256, 512)	1024
n ≥ 1,000,000	128	(1024, 512, 256)	(256, 512, 1024)	2048

(Ενότητα 4.2.β Αποτελέσματα CTGAN και TVAE)

Η παραγωγή των συνθετικών δεδομένων πραγματοποιήθηκε σε τρία διαφορετικά σενάρια, χρησιμοποιώντας τα σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018. Τα αποτελέσματα των μετρικών της κάθε κατηγορίας παρουσιάζονται παρακάτω σε πίνακες μαζί με τον αριθμό των epochs που χρησιμοποιήθηκαν για την παραγωγή τους.

Σενάρια Παραγωγής δεδομένων με Synthesizers

I. **Χρήση αποκλειστικά με δεδομένα του CICIDS 2017.**
II. **Χρήση αποκλειστικά με δεδομένα του CSE-CICIDS 2018.**
III. **Συνδυασμός των δύο συνόλων δεδομένων CICIDS 2017 και CSE-CICIDS 2018, συγχωνεύοντας τα αρχεία που περιλαμβάνουν τις ίδιες ετικέτες κατηγορίας.**

Πίνακας CTGAN Σεναριο I

Κατηγορία	KS Score Avg	RMSD Mean	Absolute Mean Difference Overall	Euclidean	Duplicate Ratio	Internal Duplicate Ratio	Output Samples	Epochs
BENIGN	0.7488	0.0890	0.0116	0.2395	0.0	0.0	100000	100
Botnet Ares	0.8324	0.1330	0.0095	0.1478	0.0	0.0	100000	500
DDoS LOIC HTTP	0.8909	0.1044	0.0076	0.1296	0.0	0.0	100000	190
DoS GoldenEye	0.8710	0.2186	0.0167	0.2110	0.0	0.0	100000	500
DoS Hulk	0.8492	0.1176	0.0146	0.2216	0.0	0.0	100000	180

DoS Slowhttptest	0.8021	0.3193	0.0351	0.4377	0.0	0.0	100000	500
DoS Slowloris	0.8091	0.3504	0.0201	0.2646	0.0	0.0	100000	500
FTP-BruteForce-Patator	0.8882	0.1089	0.0241	0.3526	0.0	0.0	100000	500
Heartbleed	0.6849	0.2944	0.0592	1.2890	0.0	0.0	100000	500
Infiltration-Communication Victim Attacker	0.7568	0.2979	0.0361	0.7668	0.0	0.0	100000	500
Infiltration-NMAP Portscan	0.6544	0.1161	0.0226	0.3318	0.0	0.0	100000	210
Portscan	0.6654	0.0922	0.0264	0.4966	0.0	0.0	100000	190
SSH-BruteForce-Patator	0.8200	0.0854	0.0137	0.1857	0.0	0.0	100000	500
Web Attack-Brute Force	0.8493	0.1415	0.0226	0.8182	0.0	0.0	100000	500
Web Attack-SQL Injection	0.6818	0.3398	0.1451	2.7507	0.0	0.0	100000	500
Web Attack-XSS	0.7964	0.2160	0.0734	3.4743	0.0	0.0	100000	500

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του CTGAN βρίσκονται στον Πίνακα Δεδομένων 2017.

Πίνακας CTGAN Σεναριο II

Κατηγορία	KS Score Avg	RMSD Mean	Absolute Mean Difference Overall	Euclidean	Duplicate Ratio	Internal Duplicate Ratio	Output Samples	Epochs
BENIGN	0.8115	0.0766	0.0065	0.1451	0.0	0.0	100000	60
Botnet Ares	0.8329	0.0448	0.0062	0.1351	0.0	0.0	100000	190
DDoS HOIC	0.8960	0.0621	0.0113	0.2003	0.0	0.0	100000	60
DDoS LOIC HTTP	0.8843	0.0940	0.0102	0.1694	0.0	0.0	100000	200
DDoS LOIC UDP	0.9189	0.0404	0.0073	0.2675	0.0	0.0	100000	500
DoS GoldenEye	0.8317	0.1664	0.0154	0.2055	0.0	0.0	100000	400
DoS Hulk	0.8970	0.0741	0.0128	0.1848	0.0	0.0	100000	60
DoS Slowloris	0.7778	0.3543	0.0330	0.4360	0.0	0.0	100000	500
Infiltration-Communication Victim Attacker	0.7314	0.2441	0.0463	0.7030	0.0	0.0	100000	500
Infiltration-Dropbox Download	0.7705	0.2969	0.0198	0.4444	0.0	0.0	100000	600

Infiltration-NMAP Portscan	0.6613	0.1727	0.0374	0.4425	0.0	0.0	100000	300
SSH-BruteForce-Patator	0.9454	0.0770	0.0033	0.0596	0.0	0.0	100000	300
Web Attack-Brute Force	0.8058	0.1293	0.0290	0.8991	0.0	0.0	100000	300
Web Attack-SQL Injection	0.8261	0.2437	0.0243	0.7034	0.0	0.0	100000	600
Web Attack-XSS	0.8195	0.1836	0.0424	0.9352	0.0	0.0	100000	600

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του CTGAN βρίσκονται στον [Πίνακα Δεδομένων 2018](#).

Πίνακας CTGAN Σεναριο III

Κατηγορία	KS Score Avg	RMSD Mean	Absolute Mean Difference Overall	Euclidean	Duplicate Ratio	Internal Duplicate Ratio	Output Samples	Epochs
BENIGN	0.7806	0.0779	0.0068	0.1547	0.0	0.0	100000	60
Botnet Ares	0.8160	0.0511	0.0084	0.1685	0.0	0.0	100000	180
DDoS HOIC	0.8624	0.0627	0.0120	0.2048	0.0	0.0	100000	68
DDoS LOIC HTTP	0.8583	0.1665	0.0228	0.3210	0.0	0.0	100000	130

DDoS LOIC UDP	0.9303	0.0399	0.0045	0.1273	0.0	0.0	100000	300
DoS GoldenEye	0.7842	0.1872	0.0210	0.2669	0.0	0.0	100000	300
DoS Hulk	0.8051	0.0985	0.0258	0.4076	0.0	0.0	100000	65
DoS Slowhttptest	0.8105	0.3169	0.0297	0.3687	0.0	0.0	100000	300
DoS Slowloris	0.8029	0.3286	0.0274	0.3549	0.0	0.0	100000	300
FTP-BruteForce-Patator	0.8703	0.1154	0.0275	0.4125	0.0	0.0	100000	300
Heartbleed	0.7148	0.3180	0.0518	1.2026	0.0	0.0	100000	300
Infiltration-Communication Victim Attacker	0.7009	0.2228	0.0394	0.7102	0.0	0.0	100000	300
Infiltration-Dropbox Download	0.7512	0.2963	0.0289	0.5400	0.0	0.0	100000	300
Infiltration-NMAP Portscan	0.6577	0.1250	0.0227	0.2838	0.0	0.0	100000	400
Portscan	0.6960	0.0938	0.0272	0.5110	0.0	0.0	100000	170
SSH-BruteForce-Patator	0.8922	0.0549	0.0079	0.1161	0.0	0.0	100000	170

Web Attack-Brute Force	0.8023	0.2180	0.0295	0.7344	0.0	0.0	100000	300
Web Attack-SQL Injection	0.8191	0.2714	0.0327	0.7513	0.0	0.0	100000	300
Web Attack-XSS	0.8054	0.2021	0.0277	0.8237	0.0	0.0	100000	300

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του CTGAN βρίσκονται στους Πίνακες Δεδομένων 2017 και 2018.

Πίνακας TVAE Σεναριο I

Κατηγορία	KS Score Avg	RMSD Mean	Absolute Mean Difference Overall	Euclidean	Duplicate Ratio	Internal Duplicate Ratio	Output Samples	Epochs
Botnet Ares	0.8110	0.1721	0.0502	0.7026	0.0	0.0	100000	300
DoS GoldenEye	0.9169	0.1982	0.0105	0.1890	0.0	0.0	100000	300
DoS Slowhttptest	0.8628	0.3034	0.0166	0.2129	0.0	0.0	100000	300
DoS Slowloris	0.8509	0.3388	0.0197	0.2647	0.0	0.0	100000	300
FTP-BruteForce-Patator	0.9478	0.0650	0.0036	0.1665	0.0	0.0	100000	300
Heartbleed	0.6794	0.2632	0.0575	1.2371	0.0	0.0	100000	300

Infiltration - Communication Victim Attacker	0.7673	0.2524	0.0583	0.8885	0.0	0.0	100000	300
Infiltration - NMAP Portscan	0.7995	0.0647	0.0046	0.1208	0.0	0.0	100000	300
SSH-BruteForce-Patator	0.9165	0.0365	0.0015	0.0292	0.0	0.0	100000	300
Web Attack - Brute Force	0.8016	0.1825	0.0681	1.1586	0.0	0.0	100000	300
Web Attack - SQL Injection	0.7847	0.2922	0.0667	2.1773	0.0	0.0	100000	300
Web Attack - XSS	0.8130	0.2082	0.0740	3.4809	0.0	0.0	100000	300

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του TVAE βρίσκονται στον Πίνακα Δεδομένων 2017.

Πίνακας TVAE Σεναριο II

Κατηγορία	KS Score Avg	RMSD Mean	Absolute Mean Difference Overall	Euclidean	Duplicate Ratio	Internal Duplicate Ratio	Output Samples	Epochs
DDoS LOIC UDP	0.9756	0.0379	0.0058	0.2130	0.0000	0.0000	100000	360
DoS GoldenEye	0.9134	0.1348	0.0100	0.1320	0.0000	0.0000	80000	350
DoS Slowloris	0.8777	0.3277	0.0105	0.1266	0.0000	0.0000	100000	400

Infiltration - Communication Victim Attacker	0.7313	0.1953	0.0489	0.8332	0.0000	0.0000	100000	500
Infiltration - Dropbox Download	0.7444	0.3016	0.0500	0.7091	0.0000	0.0000	100000	500
Infiltration - NMAP Portscan	0.7750	0.1037	0.0068	0.1185	0.0000	0.0000	20000	300
Web Attack - Brute Force	0.8530	0.1250	0.0288	0.9134	0.0000	0.0000	100000	500
Web Attack - SQL Injection	0.7782	0.2401	0.0503	0.9507	0.0000	0.0000	100000	500
Web Attack - XSS	0.8682	0.1666	0.0359	0.9154	0.0000	0.0000	100000	500

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του TVAE βρίσκονται στον Πίνακα Δεδομένων 2018.

Πίνακας TVAE Σεναριο III

Κατηγορία	KS Score Avg	RMSD Mean	Absolute Mean Difference Overall	Euclidean	Duplicate Ratio	Internal Duplicate Ratio	Output Samples	Epochs
DDoS LOIC UDP	0.9785	0.0366	0.0039	0.2024	0.0000	0.0000	100000	300
DoS GoldenEye	0.9053	0.1571	0.0092	0.1268	0.0000	0.0000	100000	300
DoS Slowhttptest	0.8508	0.3059	0.0263	0.3111	0.0000	0.0000	100000	300

DoS Slowloris	0.8764	0.3093	0.0101	0.1315	0.0000	0.0000	100000	300
FTP-BruteForce-Patator	0.9459	0.0660	0.0040	0.1675	0.0000	0.0000	100000	300
Heartbleed	0.6945	0.2649	0.0614	1.2528	0.0000	0.0000	100000	300
Infiltration - Communication Victim Attacker	0.7187	0.2212	0.0645	0.9861	0.0000	0.0000	100000	500
Infiltration - Dropbox Download	0.7229	0.2950	0.0676	0.9343	0.0000	0.0000	100000	300
Web Attack - Brute Force	0.8501	0.2152	0.0341	0.7592	0.0000	0.0000	100000	300
Web Attack - SQL Injection	0.8102	0.2587	0.0597	0.9774	0.0000	0.0000	100000	300
Web Attack - XSS	0.8397	0.1785	0.0316	0.8261	0.0000	0.0000	100000	300

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του TVAE βρίσκονται στους Πίνακες Δεδομένων 2017 και 2018.

ΚΕΦΑΛΑΙΟ 5 Ανάπτυξη και Αξιολόγηση Μοντέλων Μηχανικής Μάθησης για Ανίχνευση Επιθέσεων

(Υποκεφάλαιο 5.1 Μοντέλα Μηχανικής Μάθησης για Ανίχνευση Επιθέσεων)

Σε αυτό το κεφάλαιο παρουσιάζονται τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν για την ανίχνευση επιθέσεων στα δεδομένα του CICIDS 2017 και CSE-CICIDS 2018.
Περιγράφεται η διαδικασία εκπαίδευσης, οι υπερπαράμετροι, καθώς και τα αποτελέσματα αξιολόγησης

(Ενότητα 5.1.α Επιλογή μοντέλου)

Για την ανίχνευση επιθέσεων εφαρμόστηκε ο **XGBoost** (Extreme Gradient Boosting), ένας από τους πιο ισχυρούς αλγόριθμους στην μηχανική μάθηση για ταξινόμηση και παλινδρόμηση [16,73]. Στην παρούσα εργασία το ζητούμενο ήταν η δημιουργία μοντέλων ικανών να αναγνωρίζουν αν μια διαδικτυακή ροή δεδομένων είναι κανονική ή κακόβουλη.
Στην περίπτωση κακόβουλης δραστηριότητας, το μοντέλο θα πρέπει επιπλέον να ταξινομεί την επίθεση σύμφωνα με το είδος της ώστε σε πραγματικά σενάρια κυβερνοασφάλειας να είναι δυνατή η άμεση λήψη στοχευμένων μέτρων αντιμετώπισης [76].
Ο XGBoost (Extreme Gradient Boosting) αποτελεί μία από τις πιο αποδοτικές και ευρέως χρησιμοποιούμενες υλοποιήσεις του gradient boosting για προβλήματα ταξινόμησης και παλινδρόμησης [16]. Ο αλγόριθμος βασίζεται στη λογική ότι πολλά απλά μοντέλα, τα οποία μεμονωμένα έχουν περιορισμένη ακρίβεια, μπορούν να συνδυαστούν ώστε να παραχθεί ένα ισχυρό και αξιόπιστο σύνολο [74].
Ο XGBoost βασίζεται σε θεμελιώδεις έννοιες της μηχανικής μάθησης, οι οποίες καθορίζουν τον τρόπο με τον οποίο κατασκευάζει και βελτιώνει τα μοντέλα του. Αρχικά, αξιοποιεί τους λεγόμενους weak learners, δηλαδή δέντρα απόφασης μικρού βάθους τα οποία μεμονωμένα έχουν περιορισμένη ακρίβεια [78]. Μέσω της διαδικασίας του boosting, τα αδύναμα αυτά μοντέλα συνδυάζονται προοδευτικά ώστε να παραχθεί ένας ισχυρός ταξινομητής με υψηλή ακρίβεια [78]. Στην πράξη, ο XGBoost χρησιμοποιεί μικρά δέντρα απόφασης που εκπαιδεύονται το ένα μετά το άλλο. Κάθε καινούριο δέντρο φτιάχνεται έτσι ώστε να διορθώνει τα λάθη που έκαναν τα προηγούμενα, ακολουθώντας τη λογική του boosting [74].
Το στοιχείο που διαφοροποιεί τον XGBoost είναι ότι εφαρμόζει την τεχνική του gradient boosting, κατά την οποία η εκπαίδευση κάθε νέου δέντρου καθοδηγείται από την κλίση της συνάρτησης κόστους, επιτρέποντας στο μοντέλο να ελαχιστοποιεί σταδιακά τα σφάλματα και να βελτιώνεται επαναληπτικά [16].
Στο πλαίσιο της παρούσας εργασίας αξιοποιήθηκαν συνδυαστικά οι τρεις κύριες παραλλαγές του gradient boosting που υποστηρίζει ο XGBoost, ώστε να επιτευχθεί μεγαλύτερη ακρίβεια και καλύτερη γενίκευση των αποτελεσμάτων. Αρχικά, εφαρμόστηκε η κλασική μορφή του Gradient Boosting Machine (GBM), όπου ο ρυθμός μάθησης (learning_rate) καθόριζε την επίδραση κάθε νέου δέντρου στο τελικό μοντέλο, εξασφαλίζοντας σταδιακή βελτίωση [16,74]. Η διαδικασία αυτή εμπλουτίστηκε με στοιχεία στοχαστικότητας μέσω του Stochastic Gradient Boosting. Για τον σκοπό αυτό χρησιμοποιήθηκε υποδειγματοληψία σε επίπεδο γραμμών (subsample) και σε επίπεδο χαρακτηριστικών (colsample_bytree), γεγονός που συνέβαλε στη μείωση του κινδύνου

υπερπροσαρμογής και ενίσχυσε την ικανότητα γενίκευσης [74]. Τέλος, αξιοποιήθηκε η Regularized Gradient Boosting, η οποία ενσωματώνει κανονικοποίηση τύπου L1 (reg_alpha) και L2 (reg_lambda), ώστε να ελέγχεται αποτελεσματικά η πολυπλοκότητα του μοντέλου και να αποτρέπεται η υπερβολική προσαρμογή στα δεδομένα εκπαίδευσης [78].
Η επιλογή του XGBoost δεν ήταν τυχαία, καθώς συνδυάζει υψηλή ακρίβεια με αποδοτικότητα και ευελιξία. Ιδιαίτερα σημαντικά για το συγκεκριμένο πρόβλημα ήταν η ανθεκτικότητά του στην αντιμετώπιση ανισόρροπων δεδομένων, χαρακτηριστικό που εμφανίζεται έντονα στα σύνολα δεδομένων της εργασίας αυτής, όπου οι επιθέσεις είναι πολύ λιγότερες από τις κανονικές ροές [76,77]. Ένα ακόμη σημαντικό πλεονέκτημα του XGBoost είναι η ενσωματωμένη δυνατότητά του για αξιοποίηση GPU, η οποία στην παρούσα εργασία χρησιμοποιήθηκε προκειμένου να μειωθεί δραστικά ο απαιτούμενος χρόνος τόσο για την εκπαίδευση όσο και για την πρόβλεψη [73]. Χάρη σε αυτά τα χαρακτηριστικά, ο XGBoost αποτελεί ιδανική επιλογή για την ανίχνευση και πολυκατηγορική ταξινόμηση κακόβουλων ροών δεδομένων.

(Ενότητα 5.1.β Σενάρια Υλοποίησης: πολυκατηγορικής ταξινόμησης και δυαδικής One vs All (OvA))

Στην παρούσα εργασία δοκιμάστηκαν δύο διαφορετικές προσεγγίσεις για την αναγνώριση κακόβουλων ροών δεδομένων: η πολυκατηγορική ταξινόμηση (Multiclass classification) και η προσέγγιση δυαδικής ταξινόμησης One vs All (OvA).
Η πρώτη προσέγγιση αφορά την εκπαίδευση ενός ενιαίου μοντέλου Multiclass XGBClassifier, το οποίο μπορεί να διακρίνει όλες τις διαθέσιμες κατηγορίες ταυτόχρονα. Το μοντέλο δέχεται ως είσοδο ένα σύνολο δεδομένων (CSV), όπου κάθε γραμμή αντιστοιχεί σε μία ροή δικτύου και κάθε στήλη σε ένα χαρακτηριστικό (feature). Ανάμεσα στα χαρακτηριστικά περιλαμβάνεται η στήλη Label, η οποία υποδεικνύει την κατηγορία της κάθε ροής (π.χ. Benign ή τύπος επίθεσης). Οι ετικέτες αυτές αξιοποιούνται κατά την εκπαίδευση, ώστε το μοντέλο να μάθει να αναγνωρίζει τα μοτίβα που αντιστοιχούν σε κάθε ετικ. Για τον λόγο αυτό, η προσέγγιση αυτή εντάσσεται στα μοντέλα εποπτευόμενης μάθησης (supervised learning).
Πριν από την εκπαίδευση, οι ετικέτες του πεδίου Label μετατράπηκαν σε ακέραιες τιμές μέσω Label Encoding, ώστε να είναι κατάλληλες για επεξεργασία από τον classifier. Κατά τη διαδικασία εκπαίδευσης, το μοντέλο έμαθε να συσχετίζει τα μοτίβα των χαρακτηριστικών (features) με την αντίστοιχη ετικέτα (Benign ή τύπο επίθεσης). Στη φάση πρόβλεψης, όταν δίνεται ένα νέο σύνολο δεδομένων χωρίς ετικέτες, το μοντέλο ταξινομεί κάθε ροή σε μία από τις γνωστές κατηγορίες ή την χαρακτηρίζει ως unknown όταν δεν υπάρχει επαρκής βεβαιότητα.
Η μέθοδος που επιλέχθηκε για την πρόβλεψη είναι η predict_proba(), η οποία επιστρέφει τις πιθανότητες κάθε ροής να ανήκει σε όλες τις διαθέσιμες κατηγορίες. Αν και η predict_proba() παρέχεται από το API της scikit-learn, ο XGBClassifier του XGBoost την υποστηρίζει πλήρως, καθώς είναι συμβατός με τη scikit-learn [75]. Αντίστοιχη μέθοδος είναι η predict(), η οποία εσωτερικά αξιοποιεί την predict_proba() και επιστρέφει κατευθείαν την κατηγορία με τη μεγαλύτερη πιθανότητα. Παρότι η predict() είναι πιο απλή στη χρήση, προτιμήθηκε η predict_proba(), ώστε να υπάρχει έλεγχος βεβαιότητας. Σε περιπτώσεις όπου η μέγιστη πιθανότητα δεν ξεπερνά ένα προκαθορισμένο όριο εμπιστοσύνης (threshold = 0.2), η ροή χαρακτηρίζεται ως unknown. Με αυτόν τον τρόπο αποφεύγονται λανθασμένες ταξινομήσεις σε περιπτώσεις χαμηλής βεβαιότητας ή όταν η ροή αντιστοιχεί σε άγνωστης μορφής επίθεση.

Στη δεύτερη προσέγγιση, αντί να εκπαιδευτεί ένα ενιαίο μοντέλο που προβλέπει όλες τις κατηγορίες ταυτόχρονα, εκπαιδεύτηκαν ξεχωριστά δυαδικά (binary) μοντέλα για κάθε κατηγορία (Benign ή τύπο επίθεσης). Κάθε μοντέλο εκπαιδεύεται ώστε να διακρίνει μία συγκεκριμένη κατηγορία έναντι όλων των υπολοίπων, οι οποίες συγκεντρώνονται σε μία κοινή κατηγορία με ετικέτα Other. Όλα τα μοντέλα, όπως και στην προηγούμενη προσέγγιση, δέχονται ως είσοδο σύνολα δεδομένων σε μορφή πινάκων, τα οποία περιλαμβάνουν στήλη με ετικέτες (Label) για κάθε κατηγορία. Εφόσον χρησιμοποιούν τις ετικέτες αυτές ως γνώση κατά την εκπαίδευση, χαρακτηρίζονται επίσης ως μοντέλα εποπτευόμενης μάθησης. Στην περίπτωση αυτή, καθώς εκπαιδεύονται δυαδικά μοντέλα, οι τιμές της στήλης Label μετατρέπονται σε δυαδικές: 1 για την κατηγορία που τίθεται ως στόχος και 0 για όλες τις υπόλοιπες. Συνεπώς, αν υπάρχουν N διαφορετικές κατηγορίες, απαιτείται η εκπαίδευση N ξεχωριστών μοντέλων.
Στη φάση της πρόβλεψης, κάθε δείγμα αξιολογείται από όλα τα εκπαιδευμένα δυαδικά μοντέλα, καθένα από τα οποία αποδίδει μια πιθανότητα το δείγμα να ανήκει στη δική του κατηγορία (στόχο). Η τελική πρόβλεψη προκύπτει από την κατηγορία με τη μεγαλύτερη πιθανότητα. Σε περιπτώσεις όπου η μέγιστη πιθανότητα είναι χαμηλότερη από ένα προκαθορισμένο όριο εμπιστοσύνης (threshold = 0.2), το δείγμα χαρακτηρίζεται ως Unknown, ώστε να αποφεύγονται λανθασμένες ταξινομήσεις. Η διαδικασία αυτή υλοποιήθηκε μέσω της μεθόδου predict_proba(), όπως και στο πολυκατηγορικό μοντέλο, επιτρέποντας την εφαρμογή μηχανισμού απόρριψης σε προβλέψεις χαμηλής βεβαιότητας.

(Ενότητα 5.1.γ Παραμέτροι μοντέλων)

Για την εκπαίδευση των μοντέλων, το XGBoost παρέχει ένα πλήθος παραμέτρων που επηρεάζουν άμεσα την απόδοσή τους, τόσο ως προς την ακρίβεια όσο και ως προς την ταχύτητα εκπαίδευσης. Οι παράμετροι που χρησιμοποιήθηκαν παρουσιάζονται στους παρακάτω πίνακες, μαζί με την τιμή που τους αποδόθηκε και περιγραφή του ρόλου τους. Οι επιλογές αυτές έγιναν μέσω πειραμάτων μέχρι να επιτευχθεί το επιθυμητό αποτέλεσμα γενίκευσης του μοντέλου και αποφυγή υπερεκπαίδευσης (overfitting).

Πίνακας Παραμέτρων Multiclass XGBClassifier

Παράμετρος	Τιμή	Περιγραφή
use_label_encoder	False	Απενεργοποιεί τον παλαιό ενσωματωμένο encoder του XGBoost, ώστε να χρησιμοποιείται ο LabelEncoder της scikit-learn που έχει δημιουργηθεί και αποθηκευτεί στο πλαίσιο της υλοποίησης.
device	"cuda"	Εκπαίδευση με χρήση GPU για ταχύτερη εκτέλεση.
tree_method	"hist"	Άλγόριθμος βασισμένος σε ιστογράμματα για την κατασκευή δέντρων (ιδανικός για μεγάλα δεδομένα όπως της παρούσης)

objective	"multi:softmax"	Για πολυκατηγορικό. Επιστρέφει κατευθείαν την κατηγορία με τη μεγαλύτερη πιθανότητα.[1]
eval_metric	"mlogloss"	Μετρική αξιολόγησης Multiclass.[2]
num_class	len(labelEncoder.classes_)	Ο αριθμός των διαφορετικών κατηγοριών στο dataset.
n_estimators	500	Αριθμός δέντρων (boosting rounds).[3]
max_depth	12	Μέγιστο βάθος κάθε δέντρου. Όσο μεγαλύτερο το βάθος τόσο πιο πολύπλοκα μοτίβα αλλά αυξημένος κίνδυνος overfitting.
learning_rate	0.03	Ρυθμός μάθησης. Μικρή τιμή υποδεικνύει σταθερότερη εκπαίδευση, αλλά χρειάζονται περισσότερα δέντρα.
subsample	0.8	Χρησιμοποιεί τυχαίο υποσύνολο των δειγμάτων (80%) σε κάθε γύρο εκπαίδευσης, δηλαδή κάθε φορά που δημιουργεί ένα νέο δέντρο.
colsample_bytree	0.8	Χρήση τυχαίου υποσυνόλου (80%) των χαρακτηριστικών κάθε φορά που κατασκευάζεται ένα δέντρο.
gamma	0.2	Το ελάχιστο κέρδος (μείωση στο loss) που πρέπει να επιτευχθεί για να γίνει ένα split σε έναν κόμβο του δέντρου.
reg_alpha	0.5	Προσθέτει ποινή 0.5 στα βάρη των χαρακτηριστικών, με αποτέλεσμα τα όχι τόσο σημαντικά χαρακτηριστικά να έχουν μικρότερο βάρος προς το 0 (θέλει να τα αποκλείσει).

[1] Κατά το στάδιο της εκπαίδευσης επιλέχθηκε η παράμετρος multi:softmax, καθώς επέτρεπε την άμεση επιστροφή της κατηγορίας με τη μεγαλύτερη πιθανότητα, γεγονός που διευκόλυνε τη γρήγορη αξιολόγηση της ορθότητας της ταξινόμησης. Αντίθετα, κατά το στάδιο της πρόβλεψης απαιτούνταν οι πιθανότητες συμμετοχής κάθε δείγματος σε όλες τις διαθέσιμες κατηγορίες, οπότε χρησιμοποιήθηκε η μέθοδος predict_proba(), η οποία επέτρεψε την εξαγωγή των πιθανών κατανομών και την εφαρμογή κατωφλίου εμπιστοσύνης (threshold). Με τον τρόπο αυτό συνδυάστηκαν τα πλεονεκτήματα της άμεσης ταξινόμησης στην εκπαίδευση με την πληρέστερη αξιολόγηση πιθανοτήτων στην πρόβλεψη.

[2] Η mlogloss (Multiclass logarithmic loss) αποτελεί μετρική που αξιολογεί το πόσο καλά οι πιθανότητες που προβλέπει το μοντέλο ευθυγραμμίζονται με τις πραγματικές κατηγορίες Όταν η σωστή κατηγορία αποδίδεται με υψηλή πιθανότητα, η συνεισφορά της στο loss είναι μικρή, όταν αποδίδεται με χαμηλή πιθανότητα, το loss αυξάνεται σημαντικά. Η μετρική αυτή τιμωρεί το μοντέλο όταν η πρόβλεψη είναι λανθασμένη και όταν η εκτίμησή του παρουσιάζει χαμηλή βεβαιότητα.

[3] Κάθε γύρος boosting προσθέτει ένα νέο δέντρο στο μοντέλο.

reg_lambda	1.0	Αντίστοιχά με το reg_alpha αλλά δεν τα μηδενίζει ποτέ. Απλά σταθεροποιεί το βάρος τους.

Πίνακας Παραμέτρων Binary XGBClassifier

Παράμετρος	Τιμή	Περιγραφή
n_estimators	300	Αριθμός δέντρων (boosting rounds).
tree_method	"hist"	Άλγόριθμος βασισμένος σε ιστογράμματα για την κατασκευή δέντρων.
device	"cuda"	Χρήση GPU.
objective	"binary:logistic"	Δυαδική ταξινόμηση. Επιστρέφει πιθανότητα σε κλειστό διάστημα 0–1.[4]
eval_metric	"auc"	για αξιολόγηση της ποιότητας του μοντέλου.
max_depth	6	Μέγιστο βάθος των δέντρων.
min_child_weight	1.0	Ελάχιστο βάρος που απαιτείται σε κάθε φύλλο για split. Στην περίπτωση αυτή το βάρος κάθε δείγματος είναι 1. Ορίζεται ότι ένα φύλλο μπορεί να χωριστεί ακόμα κι αν έχει μόνο 1 δείγμα.[5]
gamma	0.3	Ελάχιστη μείωση στο loss που απαιτείται για να γίνει split.
subsample	0.7	Κάθε δέντρο εκπαιδεύεται με τυχαίο υποσύνολο (70%) των δειγμάτων
colsample_bytree	0.7	Χρήση τυχαίου υποσυνόλου (70%) των χαρακτηριστικών σε κάθε δέντρο.
reg_alpha	2	Μηδενίζει ή μειώνει βάρη άχρηστων χαρακτηριστικών.
reg_lambda	3	Μικραίνει και σταθεροποιεί τα βάρη χωρίς να τα μηδενίζει.
scale_pos_weight	1.6	Δίνει μεγαλύτερο βάρος στη μειονοτική κατηγορία.
learning_rate	0.025	Ρυθμός μάθησης.
use_label_encoder	False	Απενεργοποιεί encoder του XGBoost.

(Υποκεφάλαιο 5.2 Πειράματα-Σενάρια και Αποτελέσματα

[4] Στη binary ταξινόμηση δεν έχει νόημα να χρησιμοποιηθεί μηχανισμός που επιστρέφει απευθείας την κατηγορία πρόβλεψης, όπως συμβαίνει με την παράμετρο multi:softmax στην πολυκατηγορική περίπτωση. Ο λόγος είναι ότι η αντίστοιχη επιλογή στο binary, η binary:logitraw, δεν εφαρμόζει καμία συνάρτηση πιθανοτήτων, αλλά επιστρέφει ακατέργαστες τιμές (raw scores) οι οποίες δεν έχουν ερμηνεία πιθανότητας. Αντίθετα, η multi:softmax βασίζεται πάντα σε πιθανότητες που κανονικοποιούνται μέσω της συνάρτησης softmax, εξασφαλίζοντας συνεπή ερμηνεία ως πιθανότητες.

[5] Η επιλογή αυτή έγινε επειδή υπήρχαν κατηγορίες με πολύ λίγα δείγματα, τα οποία διαφορετικά δεν θα λαμβάνονταν υπόψη κατά την εκπαίδευση.

(Ενότητα 5.2 α Πειράματα-Σενάρια μοντέλων)

Για την εκπαίδευση των πολυκατηγορικών και δυαδικών μοντέλων υλοποιήθηκαν έξι διαφορετικά πειραματικά σενάρια. Στόχος ήταν να εξεταστεί κατά πόσο τα εκπαιδευμένα μοντέλα μπορούν, στη φάση της πρόβλεψης, να πραγματοποιούν σωστή ταξινόμηση όταν έχουν εκπαιδευτεί με διαφορετικά σύνολα δεδομένων. Ο λόγος που εξετάστηκαν διαφορετικά σενάρια σχετίζεται με το γεγονός ότι, παρόλο που ορισμένες κατηγορίες επιθέσεων παρέμειναν κοινές στα έτη 2017 και 2018, παρατηρήθηκαν διαφοροποιήσεις στις τιμές των χαρακτηριστικών κάθε κατηγορίας. Το γεγονός αυτό είναι αναμενόμενο, καθώς τα δεδομένα συλλέχθηκαν σε διαφορετικές χρονικές περιόδους και τα πακέτα δικτύου δεν είναι ποτέ απολύτως ταυτόσημα.

Κατά τη διαδικασία εκπαίδευσης, μετά την ολοκλήρωση της μάθησης, πραγματοποιήθηκε μια αρχική πρόβλεψη με χρήση της μεθόδου predict() σε όλα τα διαθέσιμα δεδομένα, τόσο στα Multiclass όσο και στα binary μοντέλα. Στόχος ήταν να ελεγχθεί αν τα μοντέλα μπορούσαν να ταξινομήσουν σωστά τα δείγματα με τα οποία εκπαιδεύτηκαν. Αφού διαπιστώθηκε ότι εμφάνιζαν σχεδόν τέλεια ακρίβεια στα δεδομένα εκπαίδευσης, η διαδικασία προχώρησε στο στάδιο της τελικής πρόβλεψης.

Στη φάση της τελικής πρόβλεψης, η αξιολόγηση των μοντέλων πραγματοποιήθηκε αποκλειστικά σε συνθετικά δεδομένα που παρήχθησαν με τη χρήση GANs με στόχο να διαπιστωθεί ότι τα μοντέλα δεν υπέφεραν από φαινόμενα υπερεκπαίδευσης (overfitting).

Παρακάτω παρουσιάζονται τα σενάρια εκπαίδευσης και πρόβλεψης. Η ανάλυση χωρίστηκε σε επιμέρους στάδια, ώστε να είναι δυνατή η αντιστοίχιση με τα αποτελέσματα που θα παρουσιαστούν στη συνέχεια, ανάλογα με το σύνολο δεδομένων που χρησιμοποιήθηκε σε κάθε περίπτωση.

Σενάρια Εκπαίδευσής

Σενάριο	Εκπαίδευση	Πρόβλεψη
I	Δεδομένα έτους 2017	Συνθετικά δεδομένα GAN παραγόμενα από το dataset 2017
II	Δεδομένα έτους 2018	Συνθετικά δεδομένα GAN παραγόμενα από το dataset 2018
III	Συνδυασμός δεδομένων ετών 2017 και 2018	Συνθετικά δεδομένα GAN παραγόμενα από τον συνδυασμό σύνολα δεδομένων 2017 και 2018
IV	Δεδομένα έτους 2017 και συνθετικά TVAE παραγόμενα από το dataset 2017	Συνθετικά δεδομένα GAN παραγόμενα από το dataset 2017
V	Δεδομένα έτους 2018 και συνθετικά TVAE παραγόμενα από το dataset 2018	Συνθετικά δεδομένα GAN παραγόμενα από το dataset 2018
VI	Δεδομένα έτους 2017 και 2018 και συνθετικά TVAE	Συνθετικά δεδομένα GAN παραγόμενα από τον

	παραγόμενα από τον συνδυασμό των dataset 2017 και 2018	συνδυασμό σύνολα δεδομένων 2017 και 2018

Τα δεδομένα παρουσιάζονται στους Πίνακες Δεδομένων 2017 και 2018 και στην Ενότητα 4.2.β.

(Ενότητα 5.2 β Αξιολόγηση)

Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με τη χρήση του classification report, το οποίο υπολογίζει μετρικές απόδοσης όπως η ακρίβεια (Precision), η ανάκληση (Recall), το πλήθος των πραγματικών δειγμάτων (Support) και τον F1-score. Οι μετρικές αυτές προκύπτουν μέσω εσωτερικών υπολογισμών και συνοψίζονται σε αναφορά που αποτυπώνει την απόδοση του μοντέλου σε κάθε κατηγορία.
Παράλληλα, για μια πιο άμεση οπτική κατανόηση, δημιουργήθηκε ξεχωριστά ο πίνακας σύγχυσης (confusion matrix) σε μορφή heatmap, ο οποίος παρουσιάζει την κατανομή των σωστών και λανθασμένων ταξινομήσεων (TP, TN, FP, FN). Ο πίνακας σύγχυσης δεν αποτελεί πηγή των μετρικών, αλλά λειτουργεί ως συμπληρωματική απεικόνιση των αποτελεσμάτων.

Classification report

Για την υλοποίηση του report εφαρμόστηκε η συνάρτηση **classification_report()** της scikit-learn που υπολογίζει και επιστρέφει τις βασικές μετρικές απόδοσης ενός ταξινομητή για κάθε κατηγορία. Εσωτερικά παίρνει τις προβλέψεις και τις πραγματικές ετικέτες της κάθε κατηγορίας και κατασκευάζει μικρά confusion matrices. Στη συνέχεια βάσει αυτών υπολογίζει TP, FP, FN, TN και εφαρμόζει μαθηματικούς τύπους για Precision, Recall, Support και F1. Τέλος, υπολογίζει και μέσους όρους (accuracy ,macro, micro και weighted).

TP = True Positives (σωστές θετικές προβλέψεις)
FP = False Positives (λανθασμένες θετικές προβλέψεις)
FN = False Negatives (λανθασμένες αρνητικές προβλέψεις)
TN = True Negatives (Σωστές αρνητικές προβλέψεις)

Αυτοί οι ορισμοί ισχύουν αυστηρά μόνο στο binary, όπου γίνεται σαφής διάκριση μεταξύ θετικών και αρνητικών δειγμάτων (0 ή 1).

Στο Multiclass classification, τα TP, FP και FN μπορούν να οριστούν ξεχωριστά για κάθε κατηγορία (θετική = η υπό εξέταση κατηγορία, αρνητική = όλες οι υπόλοιπες), όμως το TN δεν έχει πρακτική εφαρμογή και για αυτό παραλείπεται.

Σημαντική Παρατήρηση
Η κατηγορία unknown δεν συμπεριλαμβάνεται στο classification report, καθώς δεν αποτελεί πραγματική κατηγορία του dataset. Είναι απλά μια ετικέτα που χρησιμοποιείται αποκλειστικά όταν η πιθανότητα πρόβλεψης ενός δείγματος δεν υπερβαίνει το προκαθορισμένο κατώφλι εμπιστοσύνης. Τα δείγματα αυτά δεν αξιολογούνται ως λανθασμένες προβλέψεις στο accuracy ούτε συμπεριλαμβάνονται στις μετρικές, αλλά παρουσιάζονται μόνο στο confusion matrix για λόγους πληρότητας.

Μαθηματική Υλοποίηση (Precision)

Η Precision (Ακρίβεια) ορίζει το ποσοστό των σωστών θετικών προβλέψεων ως προς όλες τις προβλέψεις μιας κατηγορίας.

Δίνεται από τον μαθηματικό τύπο:

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

Ο τύπος ισχύει και για binary και για Multiclass classification.

Μαθηματική Υλοποίηση (Recall)

Η Recall (Ανάκληση) ορίζει το ποσοστό των σωστών θετικών προβλέψεων ως προς όλα τα πραγματικά δείγματα μιας κατηγορίας.

Δίνεται από τον μαθηματικό τύπο:

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

Ο τύπος ισχύει και για binary και για Multiclass classification.

Μαθηματική Υλοποίηση (F1-score)

Το F1-score είναι ο αρμονικός μέσος του Precision και του Recall, που εξισορροπεί τα δύο μέτρα μιας κατηγορίας. Ο αρμονικός μέσος αναδεικνύει την πραγματική απόδοση του μοντέλου μόνο όταν και το Precision και το Recall είναι ταυτόχρονα υψηλές.

Δίνεται από τον μαθηματικό τύπο:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ο τύπος ισχύει και για binary και για Multiclass classification.

Μαθηματική Υλοποίηση (Accuracy)

Το Accuracy ορίζει το ποσοστό των σωστών προβλέψεων σε όλο το σύνολο.

Δίνεται από τον μαθηματικό τύπο:

Για Binary

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Για Multiclass

$$\text{Accuracy} = \frac{\text{Αριθμός σωστών προβλέψεων}}{\text{Συνολικός αριθμός δειγμάτων}}$$

Μαθηματική Υλοποίηση (Micro Average)

Το micro average υπολογίζεται μετρώντας τα True Positives (TP), False Positives (FP) και False Negatives (FN) για επιλεγμένες κατηγορίες.

Δίνεται από τον μαθηματικό τύπο:

$$\text{Micro-Avg} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

Τα TP,FP,FN είναι συνολικά μόνο για τις επιλεγμένες κατηγορίες.

Ο τύπος ισχύει και για binary και για Multiclass classification.

Παρατήρηση

Στο micro average, οι μετρικές υπολογίζονται ξεχωριστά για κάθε κατηγορία θεωρώντας την ως θετική, και στη συνέχεια αθροίζονται τα TP, FP και FN όλων των κατηγοριών.
Με αυτόν τον τρόπο, το micro average μετράει την απόδοση του μοντέλου συνολικά, αλλά σε επίπεδο επιλεγμένων κατηγοριών.
Στην παρούσα εργασία το micro average μπορεί να θεωρηθεί σαν accuracy χωρίς μετρήσεις για την ετικέτα unknown. Επομένως σε πολλές περιπτώσεις όπου έχει προβλεφθεί κάποιο δείγμα ως unknown στο classification report δεν εμφανίζει τιμή accuracy αλλά micro average.

Μαθηματική Υλοποίηση (Macro Average)

Ο απλός μέσος όρος των μετρικών όλων των κατηγοριών, χωρίς βάρη.

Δίνεται από τον μαθηματικό τύπο:

$$\text{Macro-Avg} = \frac{\text{Άθροισμα μετρικών όλων των κατηγοριών}}{\text{Αριθμός κατηγοριών}}$$

Μαθηματική Υλοποίηση (Weighted Average)

Ο σταθμισμένος μέσος όρος (weighted average) υπολογίζεται λαμβάνοντας υπόψη το πλήθος των δειγμάτων κάθε κατηγορίας, ώστε οι μεγαλύτερες κατηγορίες να έχουν μεγαλύτερη βαρύτητα στον τελικό δείκτη.

Δίνεται από τον μαθηματικό τύπο:

$$\text{Weighted-Avg} = \frac{\text{Άθροισμα (Μετρικών κατηγορίας} \times \text{Support κατηγορίας)}}{\text{Συνολικό πλήθος δειγμάτων}}$$

(Ενότητα 5.2 γ Αποτελέσματα)

Στην παρούσα ενότητα παρουσιάζονται, για κάθε σενάριο σύμφωνα με την παραπάνω αντιστοίχιση, τα αποτελέσματα της φάσης πρόβλεψης σε μορφή πινάκων και σε γραφικές απεικονίσεις.

ΑΠΟΤΕΛΕΣΜΑΤΑ MULTICLASS ΤΑΞΙΝΟΜΗΣΗΣ

MULTICLASS ΣΕΝΑΡΙΟ Ι

Στο σενάριο Ι το μοντέλο εκπαιδεύτηκε αποκλειστικά σε δεδομένα του έτους 2017 και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το ίδιο σύνολο.

Αποτελέσματα Classification Report

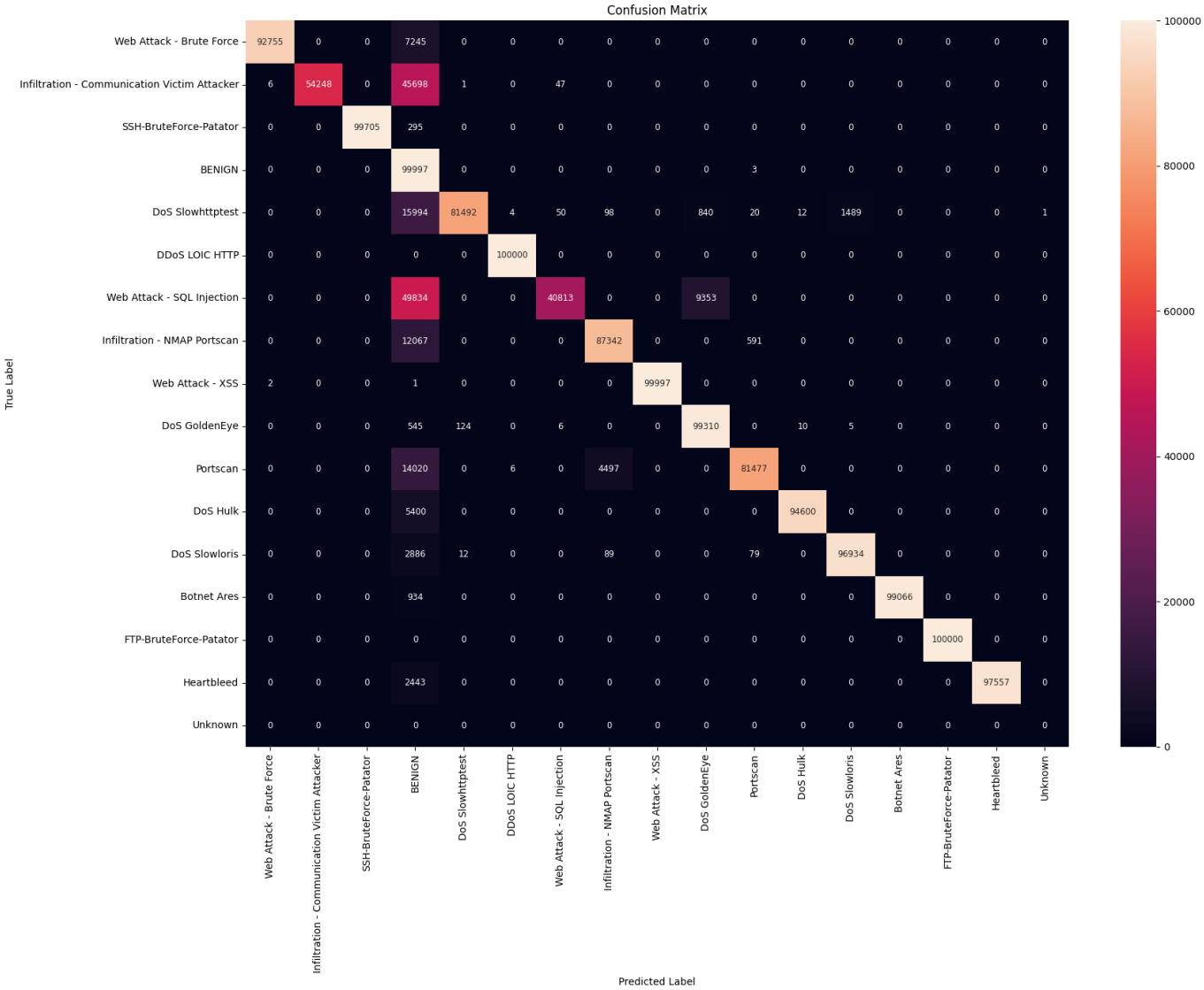
Κατηγορία	Precision	Recall	F1-score	Support
Web Attack - SQL Injection	1.00	0.41	0.58	100000
Web Attack - XSS	1.00	1.00	1.00	100000
FTP-BruteForce-Patator	1.00	1.00	1.00	100000
BENIGN	0.39	1.00	0.56	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DoS Hulk	1.00	0.95	0.97	100000
DoS Slowloris	0.98	0.97	0.98	100000
SSH-BruteForce-Patator	1.00	1.00	1.00	100000
DoS GoldenEye	0.91	0.99	0.95	100000
Web Attack - Brute Force	1.00	0.93	0.96	100000
Infiltration - Communication Victim Attacker	1.00	0.54	0.70	100000
DoS Slowhttptest	1.00	0.81	0.90	100000
Heartbleed	1.00	0.98	0.99	100000
Portscan	0.99	0.81	0.89	100000
Botnet Ares	1.00	0.99	1.00	100000
Infiltration - NMAP Portscan	0.95	0.87	0.91	100000
Micro avg	**0.89**	**0.89**	**0.89**	1600000
Macro avg	**0.95**	**0.89**	**0.90**	1600000
Weighted avg	**0.95**	**0.89**	**0.90**	1600000

Στο σενάριο αυτό, όπου η εκπαίδευση πραγματοποιήθηκε με τα δεδομένα του έτους 2017 χωρίς υπερδειγματοληψία, παρατηρείται σημαντική ανισορροπία μεταξύ των κλάσεων. Για παράδειγμα, οι κατηγορίες BENIGN (1.582.561 δείγματα), DoS Hulk (158.468 δείγματα) και Portscan (159.066 δείγματα) έχουν πολύ μεγάλο πλήθος δειγμάτων, σε αντίθεση με κατηγορίες όπως Web Attack - SQL Injection (13 δείγματα), Web Attack - XSS (18 δείγματα),

Infiltration - Communication Victim Attacker (32 δείγματα) και Heartbleed (11 δείγματα) που διαθέτουν ελάχιστα δείγματα.
Αυτό έχει ως συνέπεια το μοντέλο να αποδίδει πολύ καλά στις μεγάλες κλάσεις, ενώ οι μικρές κλάσεις παρουσιάζουν χαμηλά Recall και F1-score, αφού δεν υπάρχει αρκετή πληροφορία για να μάθει το μοντέλο τα μοτίβα τους. Χαρακτηριστικό παράδειγμα είναι οι κατηγορίες Web Attack - SQL Injection και Infiltration - Communication Victim Attacker, οι οποίες λόγω του εξαιρετικά μικρού αριθμού δειγμάτων δεν μπορούν να αναγνωριστούν με συνέπεια.
Παράλληλα, σε κάποιες κατηγορίες όπως η DoS GoldenEye (7.567 δείγματα) ή η SSH-BruteForce-Patator (2.961 δείγματα), η απόδοση παραμένει ικανοποιητική, καθώς το πλήθος των δειγμάτων, παρότι μικρότερο σε σχέση με τις μεγάλες κατηγορίες, είναι επαρκές για να αποτυπωθούν διακριτά μοτίβα. Εντυπωσιακό είναι το γεγονός ότι η κατηγορία Heartbleed, παρότι περιλάμβανε μόλις 11 δείγματα και τα συνθετικά του δεδομένα παρουσίασαν μέτρια ποιότητα, ταξινομήθηκε με σχεδόν τέλεια ακρίβεια, γεγονός που δείχνει ότι τα χαρακτηριστικά της ήταν ιδιαίτερα διακριτά και εύκολα αναγνωρίσιμα από το μοντέλο.
Σημαντικό να σημειωθεί ότι η ποιότητα των συνθετικών δεδομένων [*] που χρησιμοποιούνται για την αξιολόγηση ή την πρόβλεψη παίζει κρίσιμο ρόλο. Όταν τα συνθετικά δείγματα δεν αποτυπώνουν με ακρίβεια την πραγματική κατανομή μιας κατηγορίας, αυξάνεται η πιθανότητα να συγχέονται με άλλες κλάσεις.
Συνολικά, στο Σενάριο Ι η έλλειψη εμπλουτισμού δεδομένων ορισμένων μειονοτήτων οδήγησε σε μοντέλο με υψηλή απόδοση για τις κυρίαρχες κλάσεις, αλλά σημαντικές απώλειες στην ακρίβεια για υποεκπροσωπούμενες κατηγορίες.

[*] Πίνακες αποτελεσμάτων CTGAN

Confusion Matrix



Οι μεγαλύτερες συγχύσεις εμφανίζονται με την κατηγορία Benign, καθώς ένα σημαντικό ποσοστό κακόβουλων ροών ταξινομήθηκε εσφαλμένα ως κανονική κίνηση. Το φαινόμενο αυτό είναι ιδιαίτερα έντονο σε κατηγορίες με μικρό αριθμό αλλά και σε ορισμένες επιθέσεις. Η σύγχυση μεταξύ διαφορετικών επιθέσεων ήταν περιορισμένη και όχι ιδιαίτερα έντονη. Αυτό δείχνει ότι το μοντέλο καταφέρνει να διακρίνει ικανοποιητικά τις διάφορες κατηγορίες επιθέσεων, με το κύριο πρόβλημα να εντοπίζεται στην ομοιότητα ορισμένων κακόβουλων ροών με την κανονική κίνηση. Σημαντικό να σημειωθεί ότι σχεδόν κανένα δείγμα δεν ταξινομήθηκε ως άγνωστο, γεγονός που δείχνει ότι το μοντέλο παρήγαγε πάντα κάποια πρόβλεψη, ακόμη και σε περιπτώσεις χαμηλής βεβαιότητας.

MULTICLASS ΣΕΝΑΡΙΟ ΙΙ

Στο σενάριο ΙΙ το μοντέλο εκπαιδεύτηκε αποκλειστικά σε δεδομένα του έτους 2018 και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGANs από το ίδιο σύνολο.

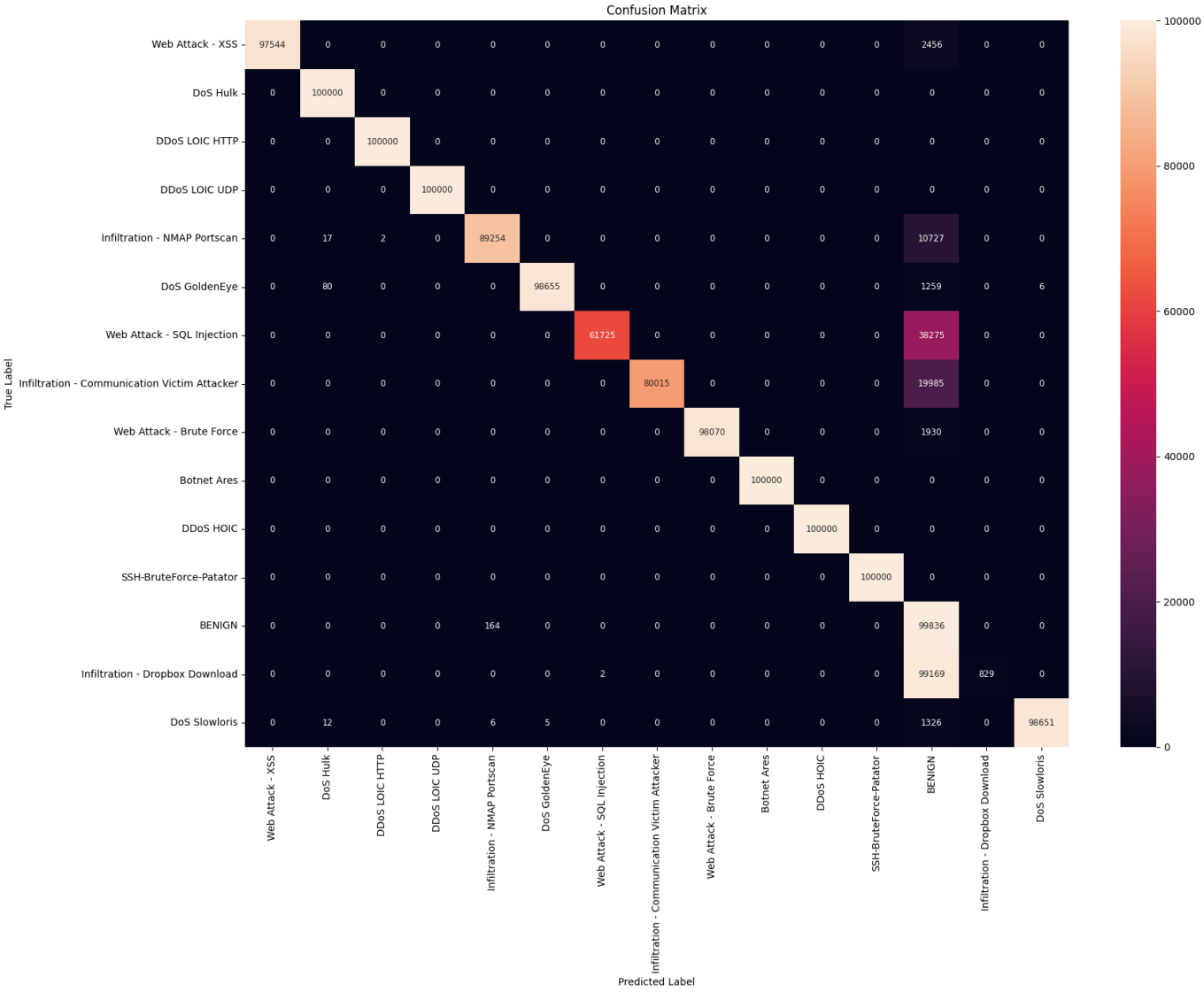
Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
Web Attack - XSS	1.00	0.98	0.99	100000
Web Attack - Brute Force	1.00	0.98	0.99	100000
DoS Slowloris	1.00	0.99	0.99	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
DoS Hulk	1.00	1.00	1.00	100000
SSH-BruteForce-Patator	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
BENIGN	0.36	1.00	0.53	100000
Infiltration - NMAP Portscan	1.00	0.89	0.94	100000
DoS GoldenEye	1.00	0.99	0.99	100000
Infiltration - Dropbox Download	1.00	0.01	0.02	100000
Infiltration - Communication Victim Attacker	1.00	0.80	0.89	100000
Web Attack - SQL Injection	1.00	0.62	0.76	100000
Accuracy			**0.88**	1500000
Macro avg	**0.96**	**0.88**	**0.87**	1500000
Weighted avg	**0.96**	**0.88**	**0.87**	1500000

Σε αυτό το σενάριο, όπου η εκπαίδευση πραγματοποιήθηκε με τα δεδομένα του έτους 2018 χωρίς εφαρμογή εμπλουτισμού δεδομένων (oversampling), παρατηρείται επίσης έντονη ανισορροπία μεταξύ των κλάσεων. Για παράδειγμα, οι κατηγορίες BENIGN (5.000.000 δείγματα), DoS Hulk (1.803.160 δείγματα) και DDoS HOIC (1.082.293 δείγματα) υπερισχύουν συντριπτικά, ενώ άλλες κατηγορίες διαθέτουν ελάχιστα δείγματα, όπως η Web Attack - SQL Injection (39 δείγματα), η Web Attack - XSS (113 δείγματα), η Web Attack - Brute Force (131 δείγματα) και η Infiltration - Dropbox Download (85 δείγματα). Η ανισορροπία αυτή έχει άμεσο αντίκτυπο στην απόδοση του μοντέλου. Οι κατηγορίες με μεγάλο πλήθος δειγμάτων αναγνωρίζονται με υψηλή ακρίβεια, καθώς το μοντέλο μαθαίνει αποτελεσματικά τα μοτίβα τους. Αντίθετα, κατηγορίες με εξαιρετικά μικρό αριθμό δειγμάτων δεν διαθέτουν επαρκή πληροφορία ώστε να γενικευτούν σωστά, με αποτέλεσμα πολύ χαμηλές τιμές Recall, όπως στην περίπτωση του Infiltration - Dropbox Download (Recall = 0.01) και της Web Attack - SQL Injection (Recall = 0.62). Παρόλα αυτά, υπάρχουν μικρές κατηγορίες που εμφάνισαν ικανοποιητική απόδοση, όπως η Web Attack - XSS (Recall = 0.98) και η Web Attack - Brute Force (Recall = 0.98), γεγονός που υποδεικνύει ότι τα χαρακτηριστικά τους είναι αρκετά διακριτά ώστε να αναγνωριστούν ακόμα και με περιορισμένα δείγματα. Συνεπώς, η απουσία υπερδειγματοληψίας οδηγεί σε άνιση συμπεριφορά του μοντέλου. Οι μεγάλες κλάσεις

κυριαρχούν και ταξινομούνται με ακρίβεια, ενώ οι μικρότερες, ειδικά εκείνες με ελάχιστα δείγματα, υποεκπροσωπούνται και παρουσιάζουν χαμηλές επιδόσεις.

Confusion Matrix



Οι μεγάλες κατηγορίες, όπως οι DoS Hulk, DDoS LOIC HTTP/UDP, Botnet Ares, DDoS HOIC, SSH-BruteForce-Patator παρουσιάζουν σχεδόν τέλεια απόδοση, με υψηλή ακρίβεια και ελάχιστες λανθασμένες ταξινομήσεις. Ωστόσο, στις μικρότερες κατηγορίες εμφανίζονται σημαντικές συγχύσεις. Η πιο χαρακτηριστική είναι στην κατηγορία Web Attack - SQL Injection, όπου μόλις 61.725 δείγματα ταξινομήθηκαν σωστά, ενώ 38.275 δείγματα μπερδεύτηκαν με την κατηγορία Benign. Παρόμοια, η κατηγορία Infiltration - Communication Victim Attacker εμφάνισε 80.015 σωστές ταξινομήσεις, αλλά 19.945 δείγματα ταξινομήθηκαν εσφαλμένα ως Benign. Επιπλέον, στην κατηγορία Infiltration - NMAP Portscan 10.727 δείγματα μπερδεύτηκαν με Benign, ενώ στην κατηγορία Infiltration - Dropbox Download 823 δείγματα ταξινομήθηκαν επίσης λανθασμένα ως Benign. Γενικά η κατηγορία Benign αναγνωρίστηκε τέλεια (υψηλό Precision), ωστόσο επειδή αρκετές επιθέσεις μπερδεύτηκαν με αυτήν, το Recall της έπεσε σημαντικά.

Συνολικά, το μοντέλο στο Σενάριο II εμφανίζει εξαιρετική απόδοση για τις κυρίαρχες επιθέσεις κίνηση, ωστόσο δυσκολεύεται στη διάκριση των πιο σπάνιων κατηγοριών από την

κατηγορία Benign, με αποτέλεσμα να καταγράφονται χαμηλότερα Recall και F1-scores σε αυτές τις περιπτώσεις.

MULTICLASS ΣΕΝΑΡΙΟ III

Στο σενάριο III το μοντέλο εκπαιδεύτηκε αποκλειστικά σε δεδομένα του έτους 2017 και 2018 και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το ίδιο σύνολο.

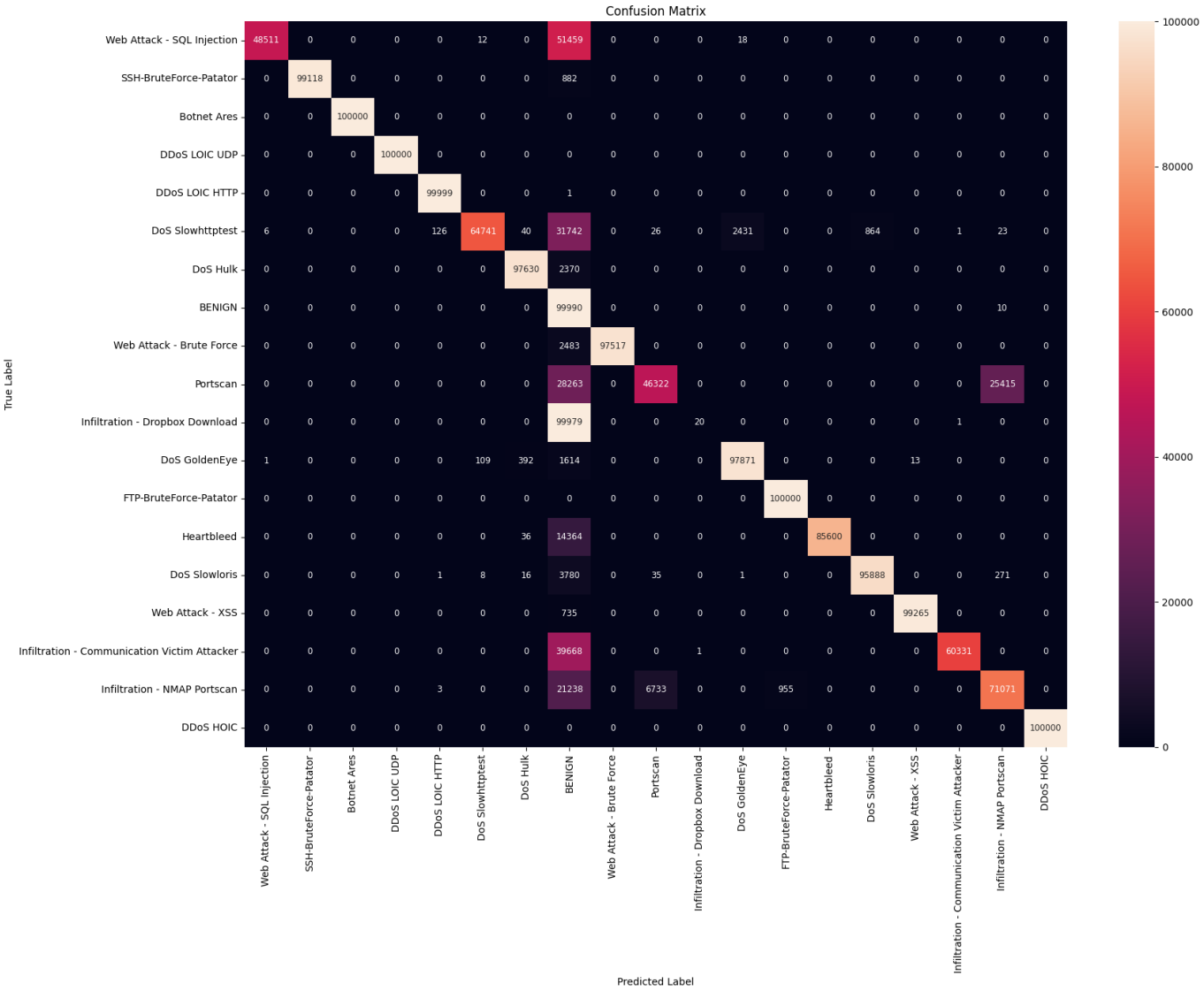
Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
Web Attack - SQL Injection	1.00	0.49	0.65	100000
SSH-BruteForce-Patator	1.00	0.99	1.00	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DoS Slowhttptest	1.00	0.65	0.79	100000
DoS Hulk	1.00	0.98	0.99	100000
BENIGN	0.25	1.00	0.40	100000
Web Attack - Brute Force	1.00	0.98	0.99	100000
Portscan	0.87	0.46	0.61	100000
Infiltration - Dropbox Download	0.95	0.00	0.00	100000
DoS GoldenEye	0.98	0.98	0.98	100000
FTP-BruteForce-Patator	0.99	1.00	1.00	100000
Heartbleed	1.00	0.86	0.92	100000
DoS Slowloris	0.99	0.96	0.97	100000
Web Attack - XSS	1.00	0.99	1.00	100000
Infiltration - Communication Victim Attacker	1.00	0.60	0.75	100000
Infiltration - NMAP Portscan	0.73	0.71	0.72	100000
DDoS HOIC	1.00	1.00	1.00	100000
Accuracy			**0.82**	1900000
Macro avg	**0.93**	**0.82**	**0.83**	1900000
Weighted avg	**0.93**	**0.82**	**0.83**	1900000

Στο Σενάριο III, η εκπαίδευση χωρίς εμπλουτισμό δεδομένων οδήγησε σε εμφανείς ανισορροπίες στην απόδοση του μοντέλου. Οι μεγάλες κατηγορίες όπως DoS Hulk (Recall 0.98, F1 0.99), DDoS LOIC HTTP/UDP (Recall 1.00, F1 1.00), Botnet Ares (F1 1.00) και FTP-BruteForce-Patator (F1 1.00) παρουσιάζουν σχεδόν τέλεια αποτελέσματα, καθώς διαθέτουν επαρκές πλήθος δειγμάτων για να μάθει το μοντέλο τα χαρακτηριστικά τους.
Αντίθετα, στις μικρότερες κατηγορίες παρατηρείται σημαντική πτώση. Η Web Attack – SQL Injection εμφανίζει χαμηλό Recall (0.49) και F1-score (0.65), με μεγάλο ποσοστό των δειγμάτων της να συγχέεται με την κλάση Benign. Αντίστοιχα, η Infiltration – Dropbox Download ουσιαστικά δεν αναγνωρίστηκε (Recall 0.00, F1 0.00), γεγονός που δείχνει ότι το μοντέλο δεν κατάφερε να μάθει καθόλου τα μοτίβα της λόγω του εξαιρετικά μικρού αριθμού δειγμάτων στην εκπαίδευση.

Η κλάση Benign, παρότι είχε Recall 1.00, εμφανίζει πολύ χαμηλή Precision (0.25) και F1-score (0.40), γεγονός που δείχνει ότι πολλές άλλες επιθέσεις μπερδεύτηκαν μαζί της. Το ίδιο συμβαίνει και με την Infiltration – Communication Victim Attacker (Recall 0.60, F1 0.75) και την Infiltration – NMAP Portscan (Recall 0.71, F1 0.72), όπου το περιορισμένο πλήθος δειγμάτων οδήγησε σε μερική σύγχυση με την κλάση Benign. Ειδικά στο Portscan, το μπέρδεμα με το NMAP Portscan δεν αποτελεί έντονο πρόβλημα, καθώς πρόκειται για επιθέση ίδιου τύπου όπως έχει προαναφερθεί.
Γενικά, οι συγχύσεις μεταξύ διαφορετικών επιθέσεων ήταν περιορισμένες, με το κύριο πρόβλημα να εντοπίζεται στη διάκριση των σπάνιων κατηγοριών από το Benign. Αυτό εξηγεί και την πτώση των συνολικών μετρικών (Accuracy 0.82, Macro avg Recall 0.82, F1-score 0.83). Συνεπώς, η ανισορροπία στα δεδομένα εκπαίδευσης, σε συνδυασμό με την περιορισμένη ποιότητα ορισμένων συνθετικών δειγμάτων, οδήγησε σε χαμηλή γενίκευση για τις υποεκπροσωπούμενες κατηγορίες.

Confusion Matrix



Οι περισσότερες λανθασμένες ταξινομήσεις αφορούν τη σύγχυση μικρών κατηγοριών με την κλάση Benign. Ένα χαρακτηριστικό παράδειγμα είναι το Portscan, το οποίο μπερδεύεται συχνά με το Infiltration - NMAP Portscan, το φαινόμενο όμως είναι αναμενόμενο, αφού πρόκειται για παρόμοιο τύπο επίθεσης. Γενικά, οι συγχύσεις μεταξύ διαφορετικών επιθέσεων παραμένουν περιορισμένες, με το κύριο πρόβλημα να εντοπίζεται στη διάκριση των σπάνιων κλάσεων από το Benign. Συνεπώς, η ανισορροπία των δεδομένων ορισμένων κατηγοριών εξηγεί τα χαμηλότερα Recall και F1-score στις υποεκπροσωπούμενες επιθέσεις.

MULTICLASS ΣΕΝΑΡΙΟ IV

Στο Σενάριο IV, το μοντέλο εκπαιδεύτηκε σε δεδομένα του έτους 2017 με εμπλουτισμό δεδομένων με τη μέθοδο TVAE και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το σύνολο 2017.

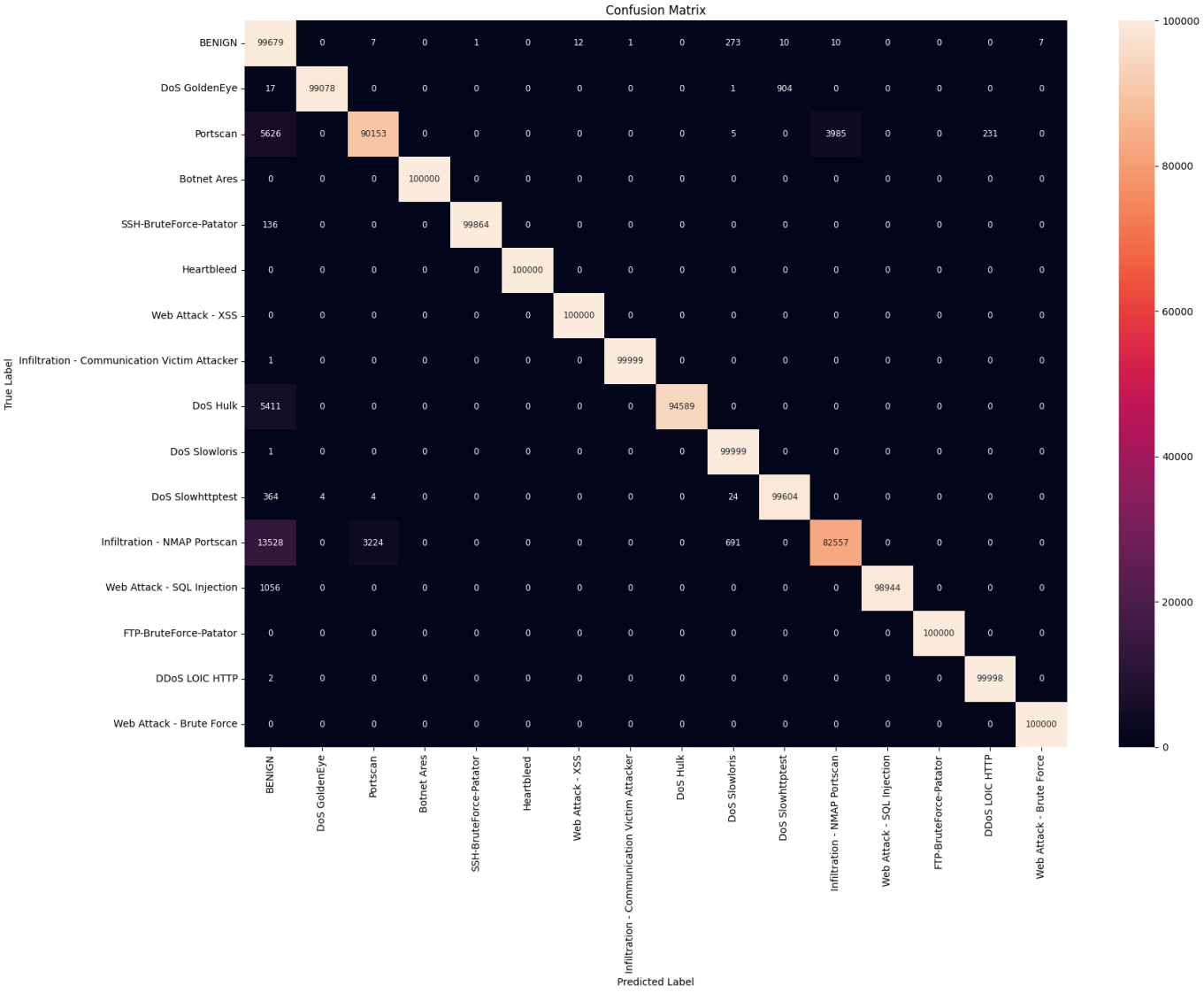
Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.79	1.00	0.88	100000
DoS GoldenEye	1.00	0.99	1.00	100000
Infiltration - Communication Victim Attacker	1.00	1.00	1.00	100000
DoS Hulk	1.00	0.95	0.97	100000
DoS Slowloris	0.99	1.00	1.00	100000
Portscan	0.97	0.90	0.93	100000
DoS Slowhttptest	0.99	1.00	0.99	100000
Infiltration - NMAP Portscan	0.95	0.83	0.89	100000
Web Attack - SQL Injection	1.00	0.99	0.99	100000
Botnet Ares	1.00	1.00	1.00	100000
FTP-BruteForce-Patator	1.00	1.00	1.00	100000
SSH-BruteForce-Patator	1.00	1.00	1.00	100000
Heartbleed	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
Web Attack - Brute Force	1.00	1.00	1.00	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Accuracy			**0.98**	1600000
Macro avg	**0.98**	**0.98**	**0.98**	1600000
Weighted avg	**0.98**	**0.98**	**0.98**	1600000

Στο Σενάριο IV, όπου εφαρμόστηκε εμπλουτισμός δεδομένων με TVAE, παρατηρείται σαφής βελτίωση στην απόδοση του μοντέλου σε σχέση με τα προηγούμενα σενάρια χωρίς εμπλουτισμό. Οι περισσότερες κατηγορίες επιθέσεων ταξινομούνται σχεδόν άψογα, με precision, recall και F1-score κοντά στο 1.00. Ωστόσο, εξακολουθούν να καταγράφονται μερικές αστοχίες. Η κατηγορία Benign αν και αναγνωρίζεται με recall 1.00, εμφανίζει χαμηλότερο precision (0.79), πράγμα που σημαίνει ότι εξακολουθεί να δέχεται λανθασμένες ταξινομήσεις από άλλες κατηγορίες. Οι κατηγορίες DoS Hulk, Portscan και Infiltration - NMAP Portscan παρουσιάζουν ελαφρώς χαμηλότερα recall (0.95, 0.90 και 0.83 αντίστοιχα),

δείχνοντας ότι μέρος των δειγμάτων τους συγχέεται κυρίως με την κατηγορία Benign ή με συγγενείς επιθέσεις (Portscan, Infiltration - NMAP Portscan).
Συνολικά, η χρήση TVAE συνέβαλε στον εξισορροπισμό των κατηγοριών και στην καλύτερη εκμάθηση των σπανιότερων επιθέσεων, μειώνοντας τα προβλήματα που παρατηρήθηκαν στα προηγούμενα σενάρια. Το μοντέλο φτάνει σε πολύ υψηλή συνολική ακρίβεια (Accuracy = 0.98, Macro/Weighted Avg F1 = 0.98), γεγονός που επιβεβαιώνει ότι η ενίσχυση με συνθετικά δεδομένα βελτιώνει την ικανότητα γενίκευσης και τη διάκριση επιθέσεων από Benign κίνηση.

Confusion Matrix



MULTICLASS ΣΕΝΑΡΙΟ V

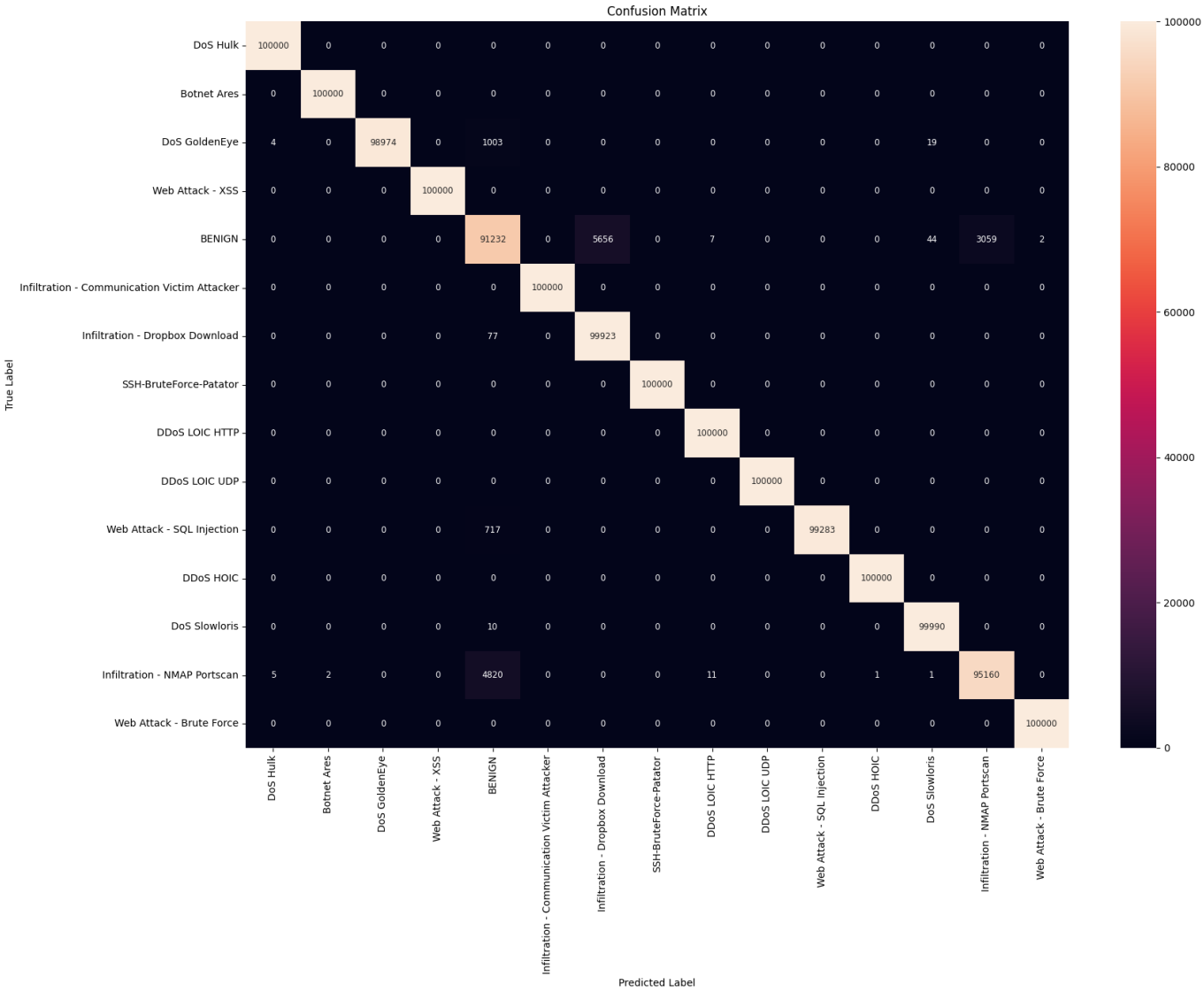
Στο Σενάριο V, το μοντέλο εκπαιδεύτηκε σε δεδομένα του έτους 2018 με εμπλουτισμό δεδομένων με τη μέθοδο TVAE και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το σύνολο 2018.

Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
DoS Hulk	1.00	1.00	1.00	100000
Botnet Ares	1.00	1.00	1.00	100000
BENIGN	0.93	0.91	0.92	100000
Infiltration - Communication Victim Attacker	1.00	1.00	1.00	100000
Infiltration - Dropbox Download	0.95	1.00	0.97	100000
SSH-BruteForce-Patator	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
Web Attack - SQL Injection	1.00	0.99	1.00	100000
Web Attack - Brute Force	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
DoS Slowloris	1.00	1.00	1.00	100000
Infiltration - NMAP Portscan	0.97	0.95	0.96	100000
DoS GoldenEye	1.00	0.99	0.99	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Accuracy			**0.99**	1500000
Macro avg	**0.99**	**0.99**	**0.99**	1500000
Weighted avg	**0.99**	**0.99**	**0.99**	1500000

Τα αποτελέσματα δείχνουν εντυπωσιακή συνολική απόδοση, με Accuracy 0.99 και Macro/Weighted F1-score επίσης 0.99. Οι περισσότερες κατηγορίες επιθέσεων, καταγράφουν σχεδόν τέλεια ταξινόμηση (Precision/Recall/F1 = 1.00 ή πολύ κοντά στο 1.00). Παράλληλα, καταγράφονται κάποιες περιορισμένες συγχύσεις.
Η κατηγορία Benign διατηρεί υψηλή απόδοση (Precision = 0.93, Recall = 0.91, F1 = 0.92), αλλά ένα ποσοστό των δειγμάτων της μπερδεύεται με άλλες κατηγορίες, γεγονός που μειώνει το precision.
Η κατηγορία Infiltration - NMAP Portscan εμφανίζει Recall = 0.95, με μικρή διαρροή προς την κατηγορία Benign, αλλά παραμένει σε πολύ υψηλά επίπεδα.
Στην DoS GoldenEye καταγράφεται επίσης μικρή απώλεια (Recall = 0.99), αλλά χωρίς ουσιαστικό αντίκτυπο στη συνολική ακρίβεια.

Confusion Matrix



Η confusion matrix επιβεβαιώνει ότι τα περισσότερα λάθη σχετίζονται με συγχύσεις μεταξύ Benign και επιθέσεων, όπως φάνηκε και σε προηγούμενα σενάρια, αλλά σε μικρότερο βαθμό. Οι συγχύσεις μεταξύ διαφορετικών τύπων επιθέσεων είναι σχεδόν ανύπαρκτες, κάτι που δείχνει ότι το μοντέλο εκπαιδεύτηκε αποτελεσματικά και ότι ο συνδυασμός TVAE (για εμπλουτισμό) και CTGAN (για αξιολόγηση) βοήθησε να αποτυπωθούν με μεγαλύτερη ακρίβεια οι κατανομές ακόμα και των πιο σπάνιων επιθέσεων.

Συνολικά, το Σενάριο V είναι το πιο ισχυρό μέχρι στιγμής, δείχνοντας ότι η χρήση TVAE για εκπαίδευση και CTGAN για παραγωγή δεδομένων αξιολόγησης οδηγεί σε πολύ σταθερό και γενικεύσιμο μοντέλο.

MULTICLASS ΣΕΝΑΡΙΟ VI

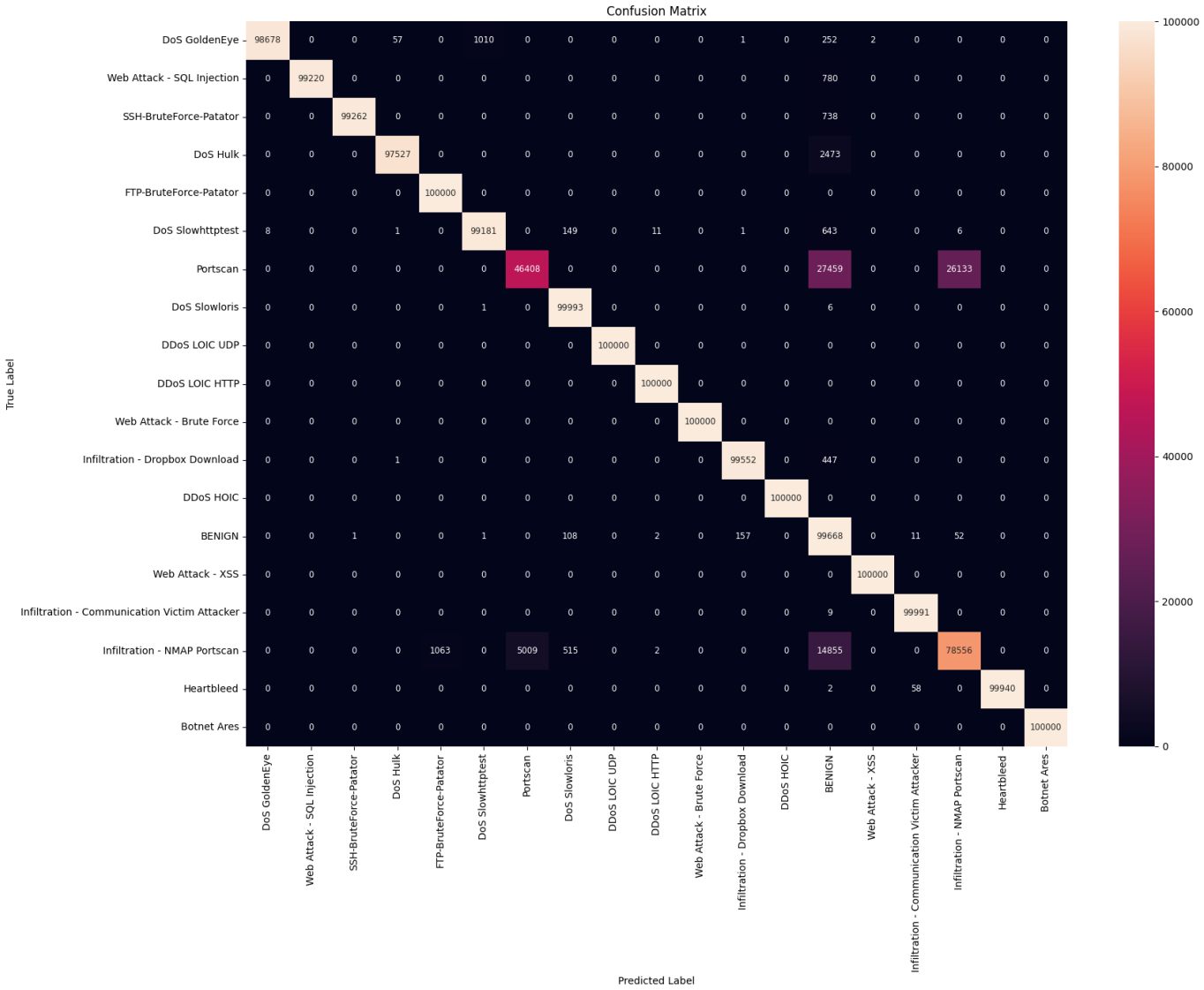
Στο Σενάριο VI, το μοντέλο εκπαιδεύτηκε σε δεδομένα του έτους 2017 και 2018 με εμπλουτισμό δεδομένων με τη μέθοδο TVAE και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το σύνολο 2017 και 2018.

Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
DoS GoldenEye	1.00	0.99	0.99	100000
Web Attack - SQL Injection	1.00	0.99	1.00	100000
SSH-BruteForce-Patator	1.00	0.99	1.00	100000
DoS Hulk	1.00	0.98	0.99	100000
FTP-BruteForce-Patator	0.99	1.00	0.99	100000
DoS Slowhttptest	0.99	0.99	0.99	100000
Portscan	0.90	0.46	0.61	100000
DoS Slowloris	0.99	1.00	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
Web Attack - Brute Force	1.00	1.00	1.00	100000
Infiltration - Dropbox Download	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
BENIGN	0.68	1.00	0.81	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Infiltration - Communication Victim Attacker	1.00	1.00	1.00	100000
Infiltration - NMAP Portscan	0.75	0.79	0.77	100000
Heartbleed	1.00	1.00	1.00	100000
Botnet Ares	1.00	1.00	1.00	100000
Accuracy			**0.96**	1900000
Macro avg	**0.96**	**0.96**	**0.95**	1900000
Weighted avg	**0.96**	**0.96**	**0.95**	1900000

Τα συνολικά αποτελέσματα είναι πολύ ισχυρά (Accuracy = 0.96, Macro F1 = 0.95), ωστόσο προκύπτουν ορισμένες ενδιαφέρουσες παρατηρήσεις. Οι περισσότερες επιθέσεις, καταγράφουν σχεδόν τέλεια Precision και Recall. Αυτό δείχνει ότι η γενίκευση του μοντέλου για τις πιο καθαρές και μαζικές επιθέσεις είναι εξαιρετική. Αντίθετα, εμφανίζονται δύο κύριες αδυναμίες. Η κατηγορία Benign ενώ έχει Recall 1.00, το Precision πέφτει στο 0.68. Αυτό σημαίνει ότι πολλές επιθέσεις με περιορισμένα ή λιγότερο πιστά συνθετικά δείγματα μπερδεύτηκαν ως Benign, κάτι που παρατηρείται συστηματικά σε όλα τα σενάρια χωρίς ισχυρή ισορροπία δεδομένων. Η κατηγορία Portscan παρουσιάζει σημαντικό πρόβλημα, με Recall 0.46 και F1 = 0.61.

Confusion Matrix



Η confusion matrix δείχνει ότι μεγάλο ποσοστό δειγμάτων Portscan ταξινομήθηκαν ως NMAP Portscan, κάτι που θεωρείται αναμενόμενο καθώς πρόκειται για ιδίους τύπους επιθέσεων με όμοια χαρακτηριστικά. Η Infiltration - NMAP Portscan, εμφανίζει επίσης συγχύσεις με Portscan και σε μικρότερο βαθμό με Benign.

Σε σχέση με τα προηγούμενα σενάρια, η απόδοση είναι σαφώς βελτιωμένη χάρη στον εμπλουτισμό με TVAE, όμως οι κατηγορίες με χαμηλότερη ποιότητα συνθετικών δεδομένων (χαμηλά KS scores, <0.7) συνεχίζουν να είναι πιο ευαίσθητες σε σύγχυση, κυρίως με Benign. Το Σενάριο VI δείχνει ότι η συνδυαστική χρήση δεδομένων 2017 και 2018 με TVAE οδηγεί σε ένα πολύ σταθερό μοντέλο με ακρίβεια 96%, αλλά τα προβλήματα παραμένουν σε δύο βασικά σημεία με τα Benign να εμπλέκονται με σπάνιες ή λιγότερο καλά συνθετικές επιθέσεις και τα Portscan να συγχέονται με Infiltration - NMAP Portscan.

ΑΠΟΤΕΛΕΣΜΑΤΑ BINARY ΤΑΞΙΝΟΜΗΣΗΣ

ONE VS ALL ΣΕΝΑΡΙΟ Ι

Στο σενάριο Ι το μοντέλο εκπαιδεύτηκε αποκλειστικά σε δεδομένα του έτους 2017 χωρίς εμπλουτισμό (υπερδειγματοληψία /TVAE) και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το ίδιο σύνολο.

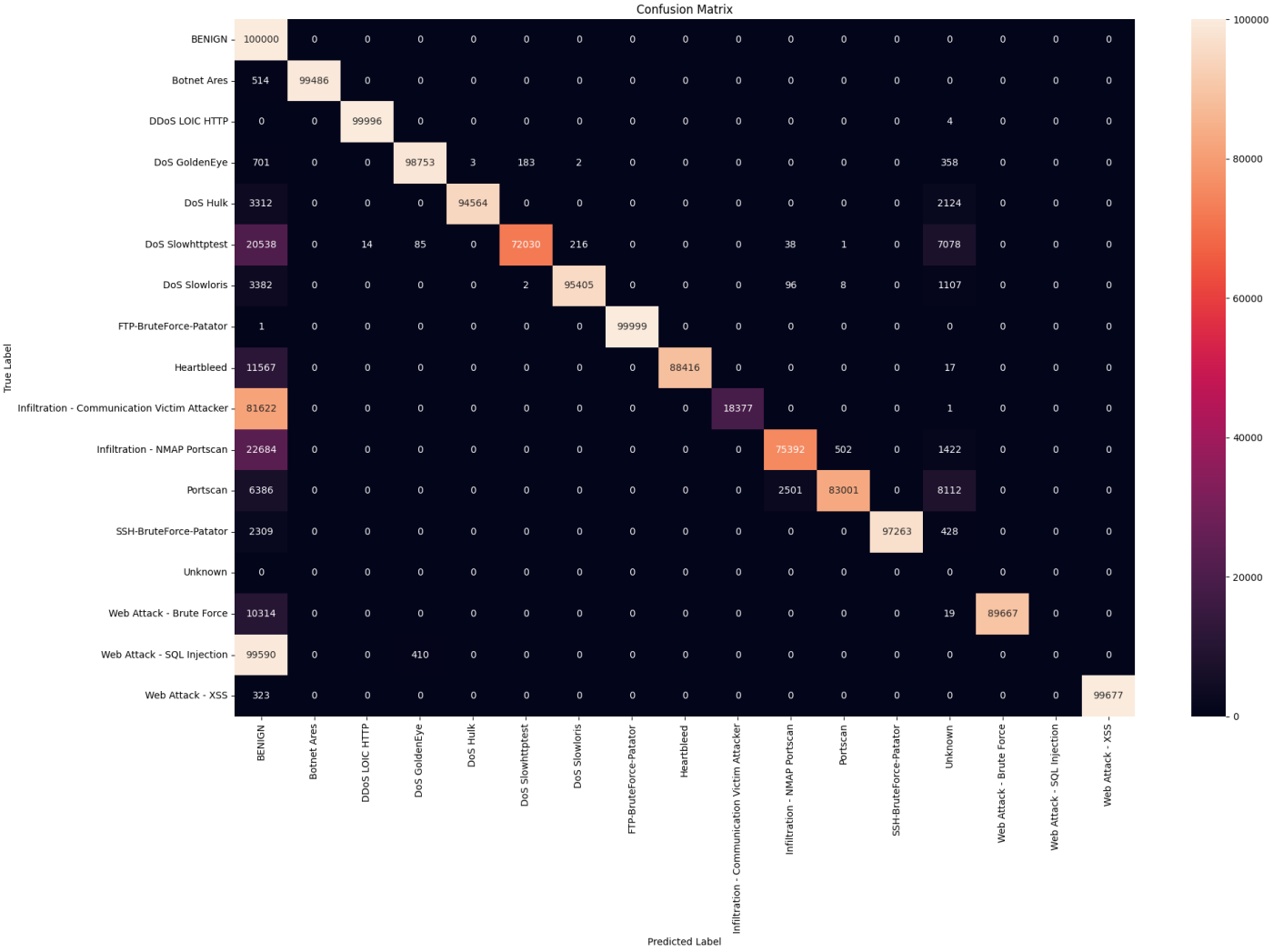
Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.28	1.00	0.43	100000
Botnet Ares	1.00	0.99	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DoS GoldenEye	1.00	0.99	0.99	100000
DoS Hulk	1.00	0.95	0.97	100000
DoS Slowhttptest	1.00	0.72	0.84	100000
DoS Slowloris	1.00	0.95	0.98	100000
FTP-BruteForce-Patator	1.00	1.00	1.00	100000
Heartbleed	1.00	0.88	0.94	100000
Infiltration - Communication Victim Attacker	1.00	0.18	0.31	100000
Infiltration - NMAP Portscan	0.97	0.75	0.85	100000
Portscan	0.99	0.83	0.90	100000
SSH-BruteForce-Patator	1.00	0.97	0.99	100000
Web Attack - Brute Force	1.00	0.90	0.95	100000
Web Attack - SQL Injection	0.00	0.00	0.00	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Micro avg	**0.83**	**0.82**	**0.83**	1600000
Macro avg	**0.89**	**0.82**	**0.82**	1600000
Weighted avg	**0.89**	**0.82**	**0.82**	1600000

Η συνολική επίδοση είναι μέτρια (Accuracy = 0.82, Macro F1 = 0.82), με πολύ καλή αναγνώριση για αρκετές επιθέσεις, αλλά και σημαντικές αδυναμίες σε άλλες. Η κατηγορία Benign παρουσιάζει Recall = 1.00 αλλά Precision μόλις 0.28. Αυτό σημαίνει ότι το μοντέλο αναγνωρίζει όλα τα κανονικά δείγματα, αλλά ταξινομεί μεγάλο αριθμό επιθέσεων ως Benign, οδηγώντας σε υψηλή σύγχυση. Η Web Attack - SQL Injection απέτυχε πλήρως (Precision/Recall = 0), κάτι που εξηγείται από το ελάχιστο πλήθος αρχικών δειγμάτων (μόλις 13 στο 2017), αλλά και από τη χαμηλή ποιότητα των συνθετικών δειγμάτων που παρήχθησαν (χαμηλά KS scores).Η Infiltration - Communication Victim Attacker επίσης είχε πολύ χαμηλή επίδοση (Recall = 0.18, F1 = 0.31) λόγω σπάνιων και δύσκολων μοτίβων. Κατηγορίες με μέτρια απόδοση είχαν οι DoS Slowhttptest (Recall = 0.72) και DoS Slowloris (Recall = 0.95) καθώς και η Infiltration - NMAP Portscan και Portscan που εμφάνισαν αλληλεπικάλυψη (F1 = 0.85 και 0.90 αντίστοιχα), κάτι αναμενόμενο λόγω της φύσης των επιθέσεων (παρόμοια χαρακτηριστικά σάρωσης). Ισχυρές κατηγορίες είναι οι Botnet Ares, DDoS LOIC HTTP/UDP,

SSH/FTP-BruteForce-Patator, Heartbleed, Web Attack - XSS που διατηρούν σχεδόν τέλεια αποτελέσματα (F1 > 0.94) όπως και στη πολυκατηγορική (Multiclass) προσέγγιση.
Το σενάριο αυτό ανέδειξε την έντονη ανισορροπία του dataset του 2017. Οι μεγάλες κατηγορίες (DoS Hulk, Portscan, Benign) αποδίδουν καλά, ενώ οι μικρές και πιο σπάνιες (Web Attack - SQL Injection, Infiltration - Communication Victim Attacker) καταρρέουν, κυρίως λόγω του μικρού πλήθους πραγματικών δειγμάτων και της περιορισμένης ποιότητας των συνθετικών δεδομένων. Το αποτέλεσμα είναι ένα μοντέλο που μαθαίνει καλά τις συχνές επιθέσεις, αλλά αποτυγχάνει στις σπάνιες, με αποτέλεσμα την πτώση του συνολικού F1-score στο 0.82.
Συγκριτικά το πολυκατηγορικό (Multiclass) μοντέλο στο αντίστοιχο σενάριο υπερέχει, γιατί έστω και με λίγα δείγματα καταφέρνει να αναγνωρίσει κάποιες σπάνιες επιθέσεις με μερική επιτυχία, ενώ το One vs All δυαδικό (binary) μοντέλο παρουσιάζει πλήρεις αποτυχίες (Web Attack - SQL Injection με Recall=0). Και στα δύο σενάρια, οι μεγαλύτερες συγχύσεις έγιναν με Benign με το Multiclass μοντέλο να πετυχαίνει υψηλότερα συνολικά σκορ (Micro avg 0.89)

Confusion Matrix



Οι περισσότερες συγχύσεις παρατηρούνται με την κατηγορία Benign, καθώς αρκετές επιθέσεις ταξινομήθηκαν εσφαλμένα ως κανονική κίνηση. Οι εμπλοκές μεταξύ διαφορετικών επιθέσεων είναι περιορισμένες και όχι τόσο έντονες, γεγονός που δείχνει ότι το μοντέλο έχει μάθει να ξεχωρίζει τα μοτίβα τους με σχετική επιτυχία.

Ωστόσο, καταγράφεται και ένα υψηλό ποσοστό αγνώστων ταξινομήσεων, κάτι που φανερώνει ότι σε περιπτώσεις χαμηλής βεβαιότητας το μοντέλο αδυνατεί να αποδώσει σε συγκεκριμένη κατηγορία και παράγει ασαφείς προβλέψεις ,κάτι που δεν παρατηρήθηκε καθόλου στα μοντέλα της Multiclass προσέγγισης .
Αυτό αποτελεί ένδειξη ότι για ορισμένες κατηγορίες τα χαρακτηριστικά δεν αναπαρίστανται επαρκώς στα δεδομένα εκπαίδευσης ή και στα συνθετικά δείγματα, με αποτέλεσμα το μοντέλο να διστάζει να αποδώσει μια πρόβλεψη.

ONE VS ALL ΣΕΝΑΡΙΟ ΙΙ

Στο σενάριο ΙΙ το μοντέλο εκπαιδεύτηκε αποκλειστικά σε δεδομένα του έτους 2018 χωρίς εμπλουτισμό (oversampling/TVAE) και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το ίδιο σύνολο.

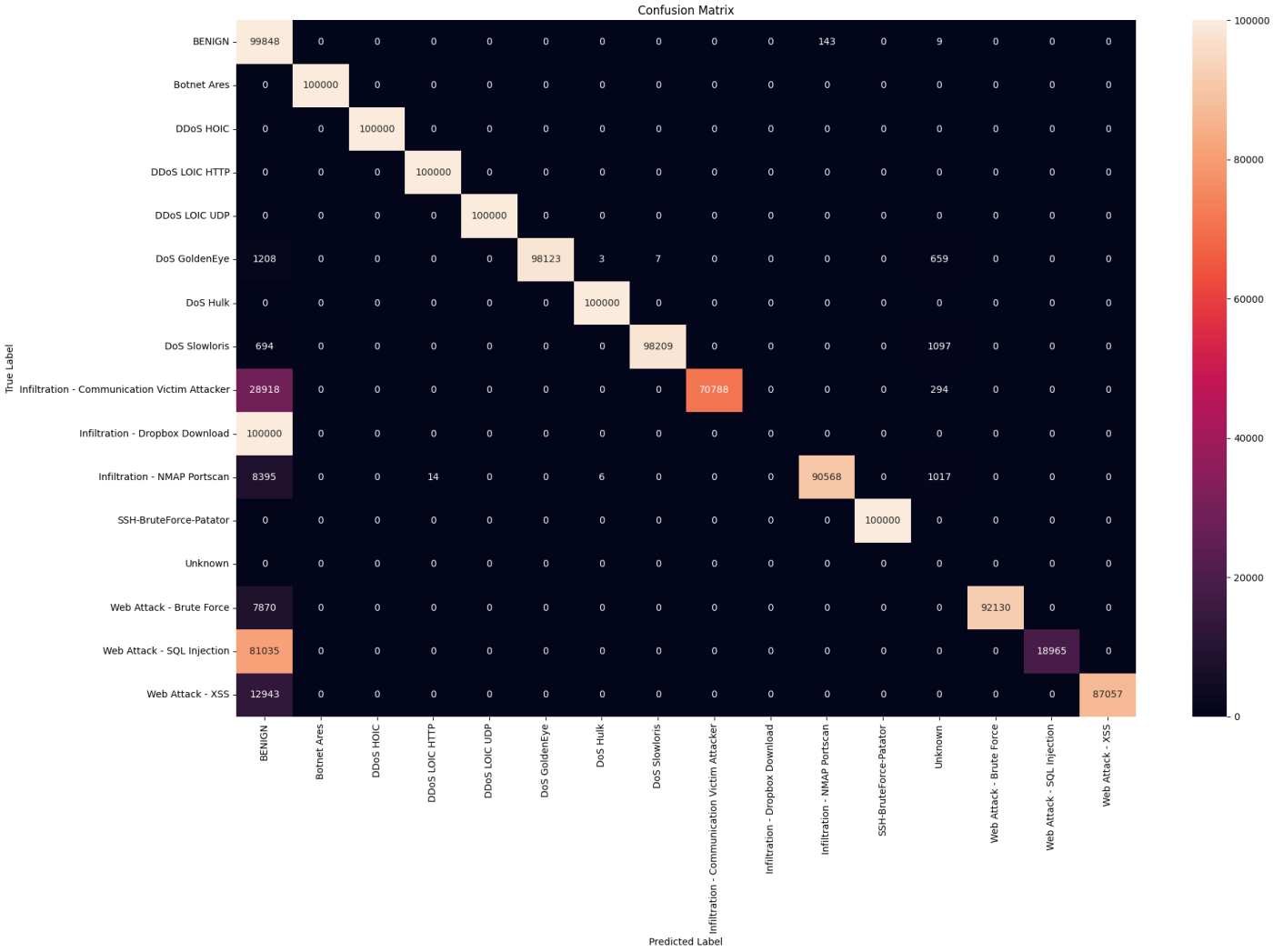
Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.29	1.00	0.45	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
DoS GoldenEye	1.00	0.98	0.99	100000
DoS Hulk	1.00	1.00	1.00	100000
DoS Slowloris	1.00	0.98	0.99	100000
Infiltration - Communication Victim Attacker	1.00	0.71	0.83	100000
Infiltration - Dropbox Download	0.00	0.00	0.00	100000
Infiltration - NMAP Portscan	1.00	0.91	0.95	100000
SSH-BruteForce-Patator	1.00	1.00	1.00	100000
Web Attack - Brute Force	1.00	0.92	0.96	100000
Web Attack - SQL Injection	1.00	0.19	0.32	100000
Web Attack - XSS	1.00	0.87	0.93	100000
Micro avg	**0.84**	**0.84**	**0.84**	1500000
Macro avg	**0.89**	**0.84**	**0.83**	1500000
Weighted avg	**0.89**	**0.84**	**0.83**	1500000

Οι μεγάλες κατηγορίες, όπως οι DoS Hulk, DDoS LOIC HTTP/UDP, HOIC, Botnet Ares και SSH-BruteForce-Patator, αποδίδουν τέλεια ή σχεδόν τέλεια με Precision/Recall κοντά στο 1.00. Αντίθετα, η κατηγορία Benign εμφανίζει Recall ίσο με 1.00, αλλά πολύ χαμηλό Precision (0.29), γεγονός που σημαίνει ότι μεγάλο ποσοστό επιθέσεων ταξινομήθηκαν λανθασμένα ως κανονική κίνηση.
Οι πιο σπάνιες κατηγορίες παρουσίασαν σημαντικές δυσκολίες. Συγκεκριμένα, η Infiltration - Communication Victim Attacker εμφάνισε καλύτερη επίδοση σε σχέση με το αντίστοιχο σενάριο του 2017 (Recall=0.71, F1=0.83), ενώ η SQL Injection εξίσου κατέρρευσε πλήρως (Recall=0.19, F1=0.32) και η Dropbox Download δεν αναγνωρίστηκε καθόλου (Recall=0). Το φαινόμενο αυτό αποδίδεται στο εξαιρετικά μικρό πλήθος δειγμάτων και στην αδυναμία των

συνθετικών δεδομένων να αναπαραστήσουν σωστά την πραγματική κατανομή. Στις κατηγορίες Web Attacks (Brute Force, XSS) παρατηρήθηκαν μέτριες απώλειες σε Recall (0.92 και 0.87 αντίστοιχα), διατηρώντας ωστόσο υψηλή ακρίβεια.

Confusion Matrix



Οι περισσότερες συγχύσεις εντοπίζονται με την κατηγορία Benign, ενώ οι συγχύσεις μεταξύ διαφορετικών επιθέσεων παραμένουν περιορισμένες. Συνολικά, το μοντέλο πέτυχε Micro Accuracy 0.84 και Macro F1 =0.83, χαμηλότερα σε σχέση με το αντίστοιχο Multiclass σενάριο, γεγονός που επιβεβαιώνει ότι η ανισορροπία των δεδομένων και η ποιότητα των συνθετικών δειγμάτων επηρεάζουν έντονα τις υποεκπροσωπούμενες κατηγορίες. Επιπλέον, και σε αυτό το σενάριο της συγκεκριμένης προσέγγισης παρατηρείται ένας μεγάλος αριθμός προβλέψεων στην κατηγορία unknown (άγνωστες ροές).

ONE VS ALL ΣΕΝΑΡΙΟ III

Στο σενάριο III το μοντέλο εκπαιδεύτηκε αποκλειστικά σε δεδομένα του έτους 2017 και 2018 χωρίς εμπλουτισμό (oversampling/TVAE) και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το ίδιο σύνολο.

Αποτελέσματα Classification Report

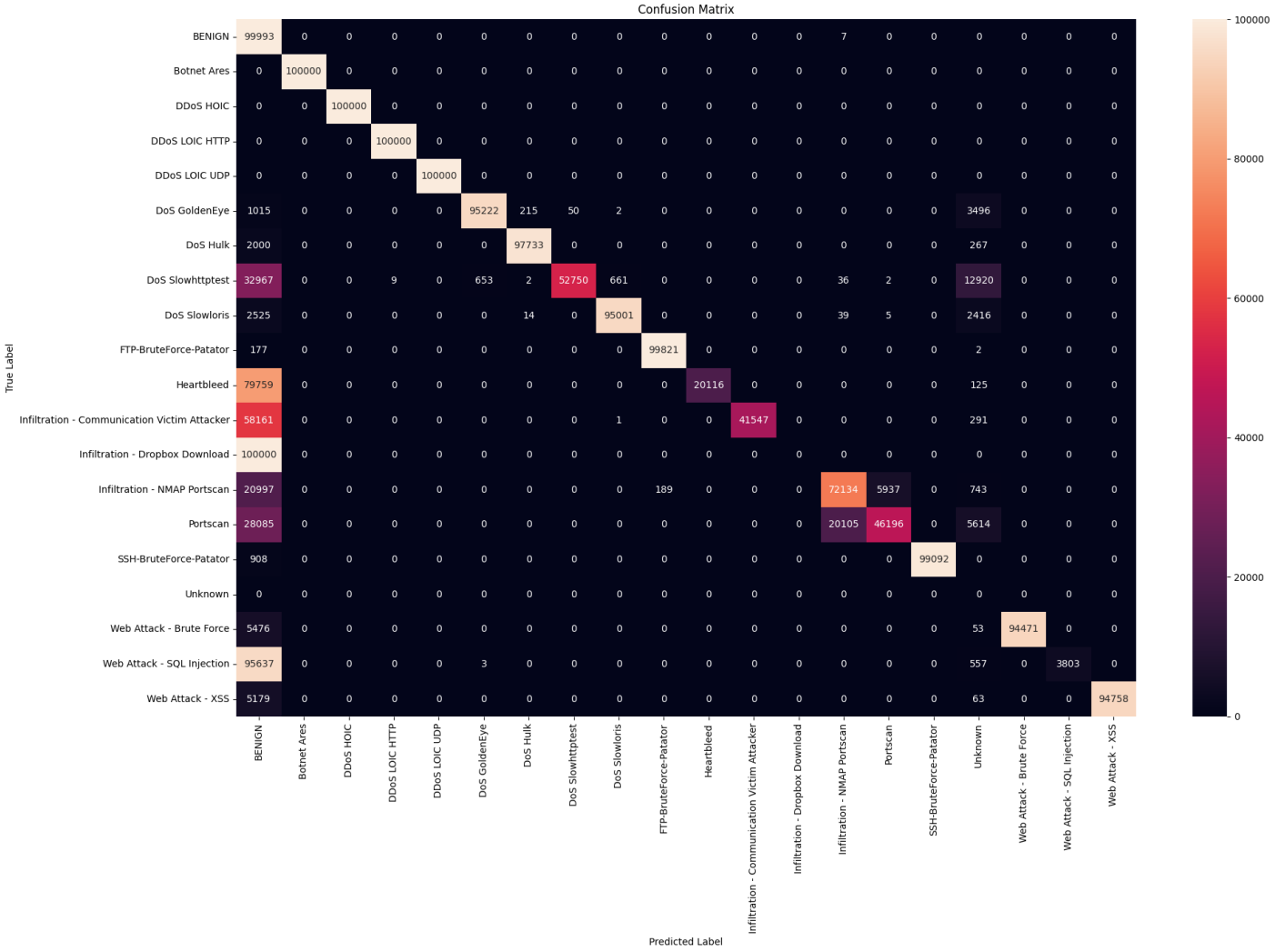
Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.19	1.00	0.32	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
DoS GoldenEye	0.99	0.95	0.97	100000
DoS Hulk	1.00	0.98	0.99	100000
DoS Slowhttptest	1.00	0.53	0.69	100000
DoS Slowloris	0.99	0.95	0.97	100000
FTP-BruteForce-Patator	1.00	1.00	1.00	100000
Heartbleed	1.00	0.20	0.33	100000
Infiltration - Communication Victim Attacker	1.00	0.42	0.59	100000
Infiltration - Dropbox Download	0.00	0.00	0.00	100000
Infiltration - NMAP Portscan	0.78	0.72	0.75	100000
Portscan	0.89	0.46	0.61	100000
SSH-BruteForce-Patator	1.00	0.99	1.00	100000
Web Attack - Brute Force	1.00	0.94	0.97	100000
Web Attack - SQL Injection	1.00	0.04	0.07	100000
Web Attack - XSS	1.00	0.95	0.97	100000
Micro avg	**0.75**	**0.74**	**0.75**	1900000
Macro avg	**0.89**	**0.74**	**0.75**	1900000
Weighted avg	**0.89**	**0.74**	**0.75**	1900000

Οι μεγάλες κατηγορίες όπως DoS Hulk, DDoS HOIC/LOIC (HTTP, UDP), Botnet Ares και SSH-BruteForce-Patator αποδίδουν εξαιρετικά (Precision/Recall ≈ 1.00), όπως και στα προηγούμενα σενάρια, γεγονός που δείχνει ότι το μοντέλο μπορεί να μάθει αποτελεσματικά τις πιο συχνές επιθέσεις. Η κατηγορία Benign παρουσιάζει Recall=1.00 αλλά εξαιρετικά χαμηλό Precision (0.19), που σημαίνει ότι ένα πολύ μεγάλο ποσοστό επιθέσεων ταξινομήθηκε εσφαλμένα ως κανονική κίνηση. Αυτό την καθιστά την κυρίαρχη πηγή σύγχυσης στο σενάριο. Όσον αφορά τις σπάνιες κατηγορίες, η επίδοση είναι απογοητευτική: το Heartbleed κατέρρευσε σχεδόν πλήρως (Recall=0.20, F1=0.33) παρότι διαθέτει διακριτά χαρακτηριστικά, το Dropbox Download δεν αναγνωρίστηκε καθόλου (Recall=0), ενώ το SQL Injection εμφάνισε οριακή αναγνώριση (Recall=0.04, F1=0.07). Καλύτερη αλλά μέτρια ήταν η απόδοση για το Infiltration - Communication Victim Attacker (Recall=0.42, F1=0.59).
Στις ενδιάμεσες κατηγορίες, το DoS Slowhttptest επηρεάστηκε σημαντικά (Recall=0.53), το Portscan (Recall=0.46, F1=0.61) ξανά εμφάνισε αρκετή σύγχυση με Infiltration - NMAP Portscan και Benign, ενώ το NMAP Portscan παρέμεινε σε μέτριο επίπεδο (F1=0.75).

Οι συγχύσεις εμφανίζονται κυρίως με την κατηγορία Benign, ενώ οι συγχύσεις μεταξύ διαφορετικών επιθέσεων είναι πιο περιορισμένες. Παράλληλα, σε σύγκριση με τα Σενάρια I και II παρατηρούνται περισσότερα unknown, ένδειξη ότι το μοντέλο συχνά διστάζει να αποδώσει μια κατηγορία, λόγω της πολυπλοκότητας που εισάγει ο συνδυασμός των datasets.

Το Σενάριο III εμφανίσε τη χαμηλότερη συνολική επίδοση (Micro F1=0.75) από τα τρία σενάρια. Αυτό οφείλεται κυρίως στην έντονη σύγχυση με την κατηγορία Benign, στην αποτυχία αναγνώρισης των πολύ σπάνιων κλάσεων (SQL Injection, Dropbox, Heartbleed) και στην αυξημένη πολυπλοκότητα που δημιουργείται από τον συνδυασμό των datasets χωρίς εμπλουτισμό. Παρότι το μοντέλο συνεχίζει να διατηρεί υψηλή απόδοση στις πιο συχνές επιθέσεις, χάνει τις υποεκπροσωπούμενες κατηγορίες, περιορίζοντας έτσι την ισορροπία και τη γενίκευσή του.

Confusion Matrix



ONE VS ALL ΣΕΝΑΡΙΟ IV

Στο Σενάριο IV, το μοντέλο εκπαιδεύτηκε σε δεδομένα του έτους 2017 με εμπλουτισμό δεδομένων με τη μέθοδο TVAE και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το σύνολο 2017.

Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.73	0.99	0.84	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	0.99	1.00	100000
DoS GoldenEye	1.00	0.99	1.00	100000
DoS Hulk	1.00	0.94	0.97	100000
DoS Slowhttptest	0.99	0.99	0.99	100000
DoS Slowloris	1.00	1.00	1.00	100000
FTP-BruteForce-Patator	1.00	1.00	1.00	100000
Heartbleed	1.00	1.00	1.00	100000
Infiltration - Communication Victim Attacker	1.00	1.00	1.00	100000
Infiltration - NMAP Portscan	0.60	0.82	0.69	100000
Portscan	1.00	0.16	0.28	100000
SSH-BruteForce-Patator	1.00	0.97	0.99	100000
Web Attack - Brute Force	1.00	1.00	1.00	100000
Web Attack - SQL Injection	1.00	0.98	0.99	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Micro avg	**0.94**	**0.93**	**0.93**	1600000
Macro avg	**0.96**	**0.93**	**0.92**	1600000
Weighted avg	**0.96**	**0.93**	**0.92**	1600000

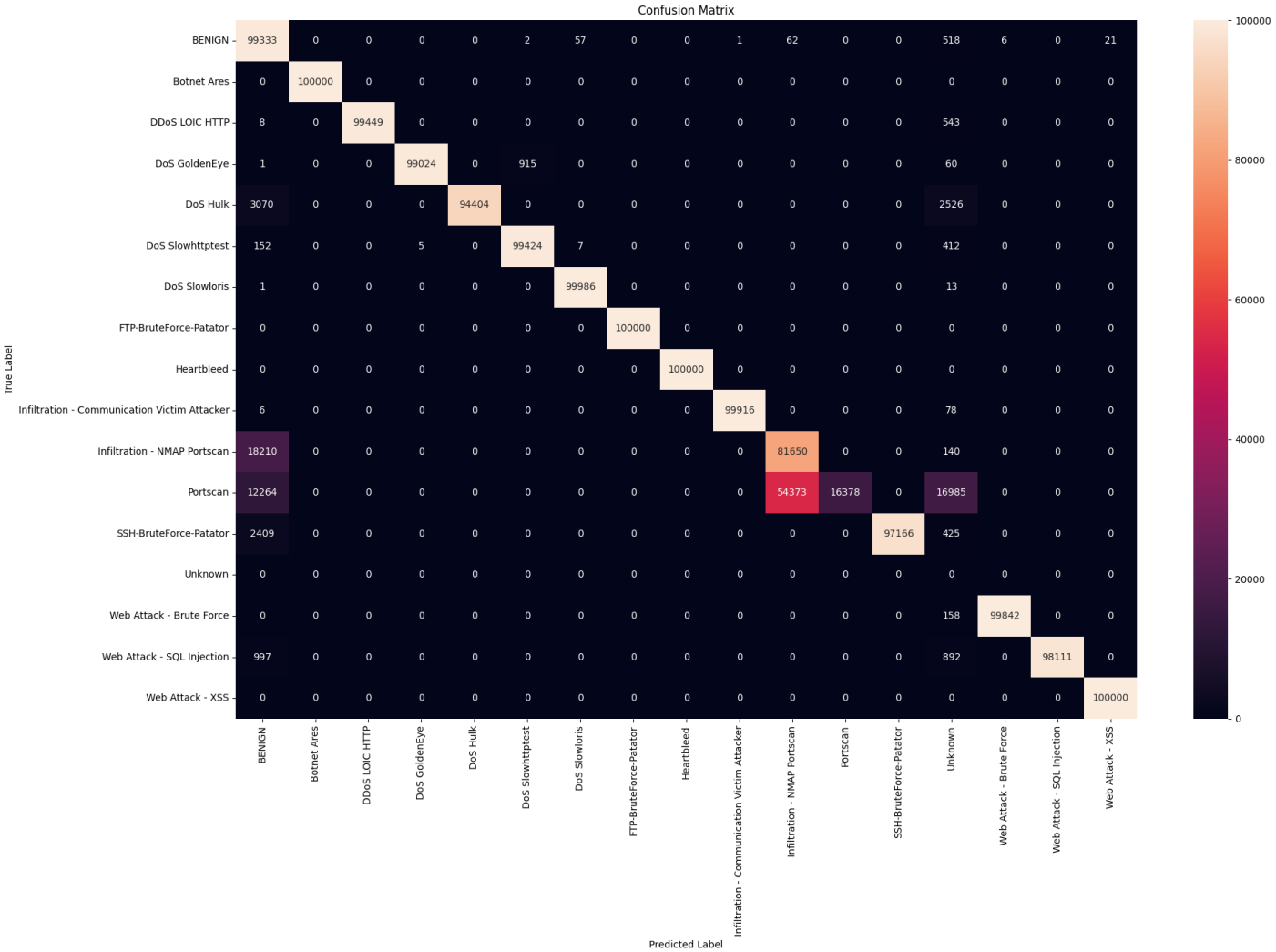
Οι μεγάλες και μεσαίες κατηγορίες όπως DoS GoldenEye, DoS Hulk, DoS Slowloris, DoS Slowhttptest, Botnet Ares, SSH/FTP-BruteForce-Patator, Heartbleed, Web Attack - Brute Force/SQL Injection/XSS, DDoS LOIC HTTP παρουσιάζουν πολύ υψηλή απόδοση, με Precision/Recall/F1 κοντά στο 1.00. Ο εμπλουτισμός με TVAE βελτίωσε αισθητά την ικανότητα του μοντέλου να γενικεύει ακόμα και σε πιο δύσκολες κατηγορίες, μειώνοντας τα προβλήματα που παρατηρήθηκαν σε προηγούμενα σενάρια χωρίς εμπλουτισμό. Η κατηγορία Benign εμφανίζει Recall=0.99 αλλά Precision=0.73, γεγονός που δείχνει ότι εξακολουθούν να ταξινομούνται λανθασμένα επιθέσεις ως κανονική κίνηση. Παρότι βελτιωμένο σε σχέση με τα προηγούμενα σενάρια.

Στις πιο δύσκολες κατηγορίες, το Portscan παρουσίασε πτώση (Recall=0.16, F1=0.28), γεγονός που δείχνει αδυναμία αναγνώρισης παρά τον εμπλουτισμό, πιθανώς λόγω ελλιπούς διαφοροποίησης στα συνθετικά δεδομένα. Αντίθετα, το NMAP Portscan εμφάνισε μέτρια απόδοση (F1=0.69).

Οι συγχύσεις εμφανίζονται κυρίως μεταξύ των κατηγοριών Portscan, Infiltration - NMAP Portscan και Benign, ενώ οι συγχύσεις μεταξύ διαφορετικών επιθέσεων παραμένουν περιορισμένες. Η εισαγωγή του εμπλουτισμού με TVAE φαίνεται να έχει βοηθήσει ώστε οι περισσότερες επιθέσεις να αναγνωρίζονται με πολύ υψηλή συνέπεια.

Το Σενάριο IV δείχνει ξεκάθαρη βελτίωση σε σχέση με τα Σενάρια I–III, με μικρότερη απώλεια απόδοσης σε σπάνιες κατηγορίες και συνολική Micro F1 = 0.93.

Confusion Matrix



ONE VS ALL ΣΕΝΑΡΙΟ V

Στο Σενάριο V, το μοντέλο εκπαιδεύτηκε σε δεδομένα του έτους 2018 με εμπλουτισμό δεδομένων με τη μέθοδο TVAE και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το σύνολο 2018.

Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.94	0.96	0.95	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	1.00	1.00	100000

DDoS LOIC UDP	1.00	1.00	1.00	100000
DoS GoldenEye	1.00	0.98	0.99	100000
DoS Hulk	1.00	1.00	1.00	100000
DoS Slowloris	1.00	1.00	1.00	100000
Infiltration - Communication Victim Attacker	1.00	1.00	1.00	100000
Infiltration - Dropbox Download	0.99	1.00	0.99	100000
Infiltration - NMAP Portscan	0.99	0.95	0.97	100000
SSH-BruteForce-Patator	1.00	1.00	1.00	100000
Web Attack - Brute Force	1.00	1.00	1.00	100000
Web Attack - SQL Injection	1.00	0.98	0.99	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Micro avg	**0.99**	**0.99**	**0.99**	1500000
Macro avg	**0.99**	**0.99**	**0.99**	1500000
Weighted avg	**0.99**	**0.99**	**0.99**	1500000

Οι μεγάλες κατηγορίες ταξινομούνται με άριστη ακρίβεια (Precision/Recall σχεδόν 1.00), αποδεικνύοντας ότι το μοντέλο είναι ιδιαίτερα αποτελεσματικό όταν υπάρχει επαρκής όγκος δεδομένων.

Η κατηγορία Benign εμφανίζει πολύ βελτιωμένη συμπεριφορά συγκριτικά με τα Σενάρια I–IV. Με Precision=0.94 και Recall=0.96, το μοντέλο αποφεύγει σε μεγάλο βαθμό τις υπερβολικές συγχύσεις, αν και εξακολουθούν να υπάρχουν μεμονωμένες λανθασμένες ταξινομήσεις επιθέσεων ως κανονική κίνηση.

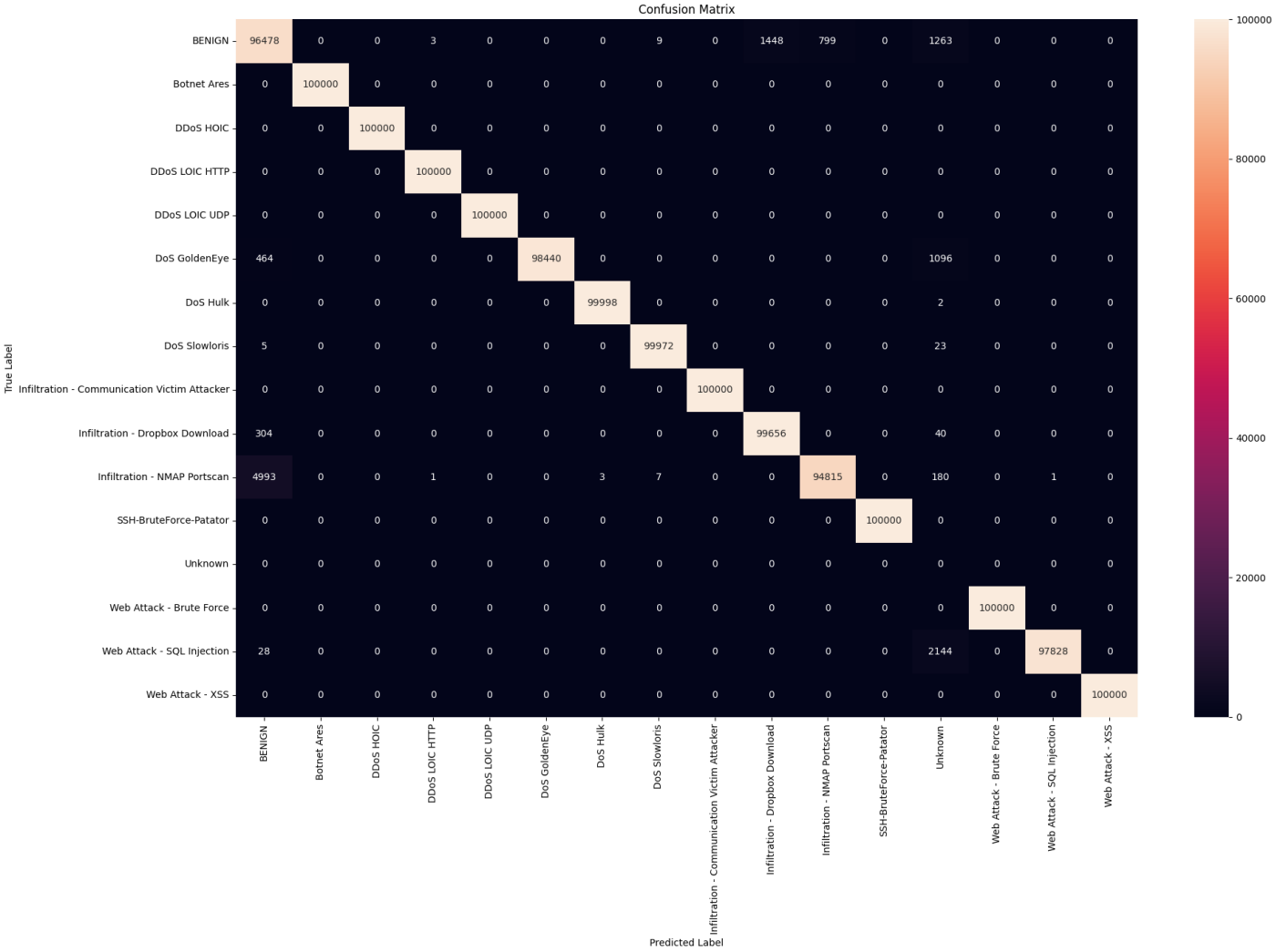
Οι σπάνιες κατηγορίες επωφελούνται ιδιαίτερα από τον εμπλουτισμό με TVAE. Συγκεκριμένα, η Web Attack - SQL Injection (Recall=0.98, F1=0.99) και η Dropbox Download (Recall=1.00, F1=0.99) αποδίδουν πλέον εξαιρετικά, δείχνοντας ότι η ενίσχυση των δεδομένων αντιμετώπισε αποτελεσματικά το πρόβλημα της υποεκπροσώπησης.

Οι ενδιάμεσες κατηγορίες, όπως το DoS GoldenEye και το Infiltration – NMAP Portscan, διατηρούν υψηλά ποσοστά ακρίβειας, με μικρές μόνο απώλειες στο Recall (0.95–0.98), χωρίς ωστόσο να επηρεάζεται σημαντικά η συνολική τους επίδοση.

Ορισμένες συγχίσεις εντοπίζονται κυρίως στην κατηγορία Benign, αλλά σε σαφώς χαμηλότερα ποσοστά, ενώ δεν παρατηρούνται συστηματικές συγχύσεις μεταξύ διαφορετικών τύπων επιθέσεων.

Το Σενάριο V αποτελεί το πιο επιτυχημένο σενάριο στα δυαδικά μοντέλα, με συνολικό F1-score = 0.99. Ο εμπλουτισμός με TVAE εξάλειψε σε μεγάλο βαθμό τις αδυναμίες των προηγούμενων σεναρίων, ενώ η απουσία της κατηγόριας Portscan παίζει μεγάλο ρόλο στα αποτελέσματα.

Confusion Matrix



ONE VS ALL ΣΕΝΑΡΙΟ VI

Στο Σενάριο VI, το μοντέλο εκπαιδεύτηκε σε δεδομένα του έτους 2017 και 2018 με εμπλουτισμό δεδομένων με τη μέθοδο TVAE και αξιολογήθηκε σε συνθετικά δεδομένα που παρήχθησαν με CTGAN από το σύνολο 2017 και 2018.

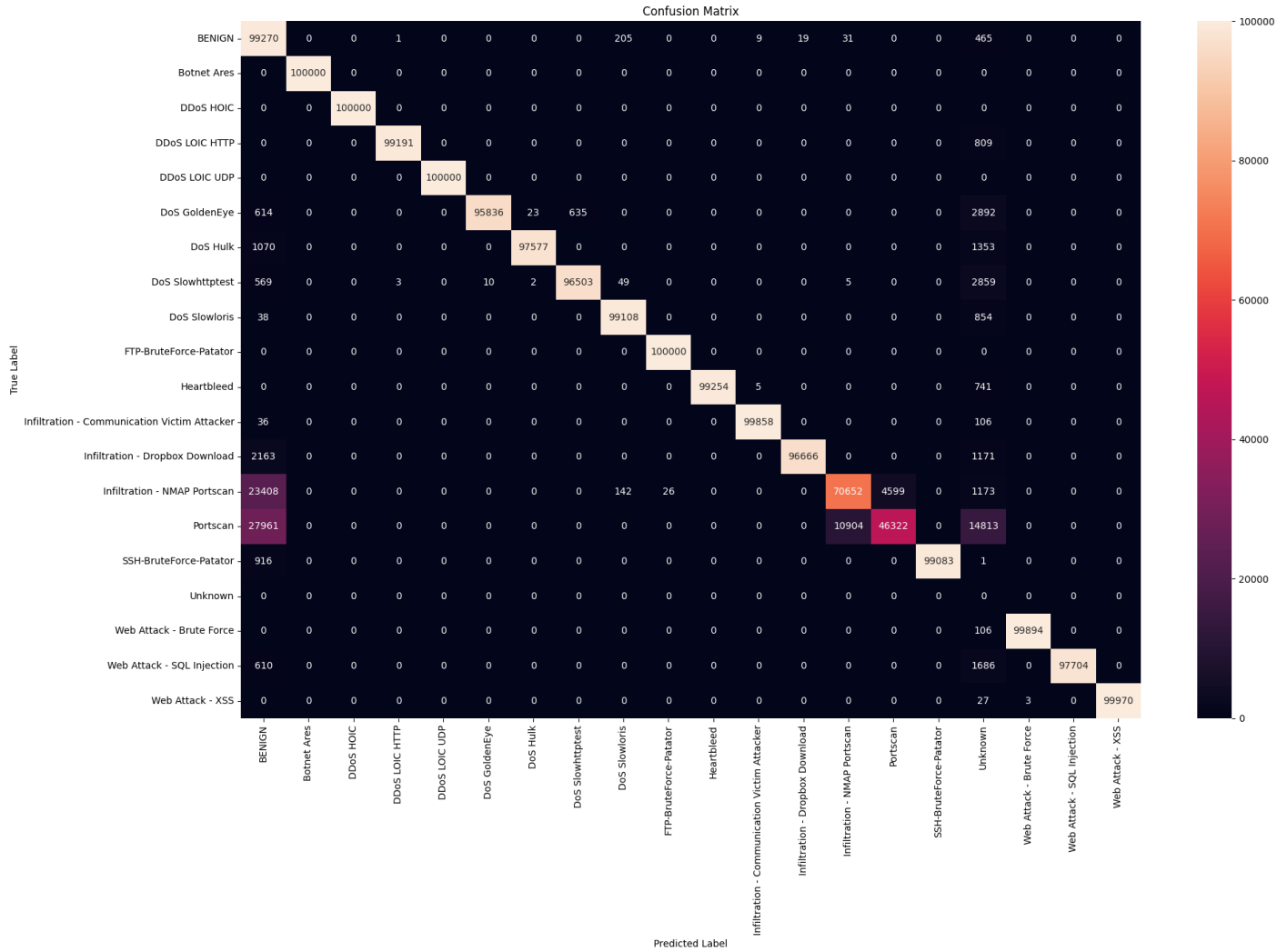
Αποτελέσματα Classification Report

Κατηγορία	Precision	Recall	F1-score	Support
BENIGN	0.63	0.99	0.77	100000
Botnet Ares	1.00	1.00	1.00	100000
DDoS HOIC	1.00	1.00	1.00	100000
DDoS LOIC HTTP	1.00	0.99	1.00	100000
DDoS LOIC UDP	1.00	1.00	1.00	100000
DoS GoldenEye	1.00	0.96	0.98	100000
DoS Hulk	1.00	0.98	0.99	100000
DoS Slowhttptest	0.99	0.97	0.98	100000

DoS Slowloris	1.00	0.99	0.99	100000
FTP-BruteForce-Patator	1.00	1.00	1.00	100000
Heartbleed	1.00	0.99	1.00	100000
Infiltration - Communication Victim Attacker	1.00	1.00	1.00	100000
Infiltration - Dropbox Download	1.00	0.97	0.98	100000
Infiltration - NMAP Portscan	0.87	0.71	0.78	100000
Portscan	0.91	0.46	0.61	100000
SSH-BruteForce-Patator	1.00	0.99	1.00	100000
Web Attack - Brute Force	1.00	1.00	1.00	100000
Web Attack - SQL Injection	1.00	0.98	0.99	100000
Web Attack - XSS	1.00	1.00	1.00	100000
Micro avg	**0.96**	**0.95**	**0.95**	1900000
Macro avg	**0.97**	**0.95**	**0.95**	1900000
Weighted avg	**0.97**	**0.95**	**0.95**	1900000

Η συνολική επίδοση είναι ιδιαίτερα υψηλή (Micro/Weighted F1 = 0.95), με σημαντικές βελτιώσεις σε πολλές κατηγορίες συγκριτικά με τα προηγούμενα σενάρια.
Οι μεγάλες κατηγορίες, διατηρούν εξαιρετικά αποτελέσματα με Precision και Recall ≈ 1.00, επιβεβαιώνοντας ότι η επάρκεια δεδομένων και ο εμπλουτισμός επιτρέπουν στο μοντέλο να μάθει με μεγάλη ακρίβεια τα αντίστοιχα μοτίβα.
Η κατηγορία Benign παρουσιάζει σημαντική βελτίωση σε σχέση με τα Σενάρια I–III, ωστόσο εξακολουθεί να αποτελεί πηγή σύγχυσης. Με Precision=0.63 και Recall=0.99, το Recall παραμένει πολύ υψηλό, αλλά το χαμηλό Precision υποδηλώνει ότι αρκετές επιθέσεις ταξινομήθηκαν εσφαλμένα ως κανονική κίνηση. Παρότι το πρόβλημα δεν έχει εξαλειφθεί, η κατάσταση είναι αισθητά βελτιωμένη σε σχέση με προηγούμενες δοκιμές χωρίς εμπλουτισμό.
Στις σπάνιες κατηγορίες, όπως Web Attack - SQL Injection (Recall=0.98, F1=0.99), το Infiltration - Dropbox Download (Recall=0.97, F1=0.98) και το Heartbleed (Recall=0.99, F1=1.00), παρατηρείται πλέον εξαιρετική απόδοση. Αυτό δείχνει ότι η χρήση TVAE κατά την εκπαίδευση κατάφερε να ενισχύσει επαρκώς τα υποεκπροσωπούμενα δεδομένα, επιτρέποντας στο μοντέλο να διακρίνει τα μοτίβα τους με συνέπεια.
Οι ενδιάμεσες κατηγορίες παρουσιάζουν μεικτή εικόνα: το DoS Slowhttptest (Recall=0.97, F1=0.98) και το DoS Slowloris (Recall=0.99, F1=0.99) αποδίδουν πολύ καλά, ενώ το NMAP Portscan (Recall=0.71, F1=0.78) και το Portscan (Recall=0.46, F1=0.61) εξακολουθούν να παρουσιάζουν αδυναμίες, με συχνές συγχύσεις προς την κατηγορία Benign και άλλες ροές.
Οι συγχύσεις παραμένουν συγκεντρωμένες κυρίως στην κατηγορία Benign, αν και σε μικρότερο βαθμό από τα πρώτα σενάρια. Οι μεταξύ διαφορετικών επιθέσεων συγχύσεις είναι περιορισμένες και μη συστηματικές, γεγονός που δείχνει ότι το μοντέλο διατηρεί καλή ικανότητα διάκρισης ανάμεσα στις κακόβουλες ροές.
Το Σενάριο VI παρουσιάζει συνολικά πολύ υψηλή επίδοση (F1 = 0.95), με το μοντέλο να διακρίνει με συνέπεια τις περισσότερες επιθέσεις. Οι μεγάλες κατηγορίες αποδίδουν τέλεια, οι σπάνιες ενισχύονται σημαντικά χάρη στον εμπλουτισμό με TVAE, και οι συγχύσεις με το Benign μειώνονται, αν και παραμένουν το κύριο αδύναμο σημείο. Η απόδοση στις κατηγορίες NMAP Portscan και Portscan συνεχίζει να υστερεί, γεγονός που υποδηλώνει ότι είτε ο διαχωρισμός τους ως ξεχωριστές κλάσεις δεν προσφέρει ουσιαστικό όφελος, είτε ότι τα συνθετικά τους δεδομένα δεν αποτυπώνουν με επάρκεια την πραγματική κατανομή των ροών, οδηγώντας σε περιορισμένη δυνατότητα διάκρισης από το μοντέλο.

Confusion Matrix



Συμπεράσματα

Η παρούσα εργασία είχε ως στόχο την ανάπτυξη και αξιολόγηση ενός Συστήματος Ανίχνευσης Εισβολών (IDS) βασισμένου σε μεθόδους Μηχανικής Μάθησης, με χρήση των σύνολα δεδομένων CICIDS 2017 και CSE-CICIDS 2018. Πραγματοποιήθηκαν πολλαπλά πειραματικά σενάρια (Multiclass και One vs All binary), με και χωρίς εμπλουτισμό δεδομένων (oversampling), ώστε να εξεταστεί η συμπεριφορά των μοντέλων υπό διαφορετικές συνθήκες δεδομένων και ανισορροπίας. Η αξιολόγηση υλοποιήθηκε με χρήση τυπικών μετρικών (Precision, Recall, F1-score) και ανάλυσης confusion matrix, παρέχοντας ποσοτική και οπτική εικόνα της απόδοσης.

Αποτελέσματα Multiclass Ταξινόμησης

Από τη σύγκριση των Multiclass σεναρίων προέκυψε, ότι τα μοντέλα χωρίς εμπλουτισμό δεδομένων (Σενάρια I–III) εμφάνισαν μέτρια συνολική απόδοση, με Micro avg/Accuracy F1 στο εύρος 0.82–0.89**.** Οι κατηγορίες με επαρκή αριθμό δειγμάτων ταξινομήθηκαν με σχετικά καλή ακρίβεια, ενώ οι μειονοτικές κατηγορίες εμφάνισαν χαμηλότερο Recall, συχνά συγχέοντας τις επιθέσεις με κανονικές ροές (Benign). Αυτό δείχνει ότι χωρίς πρόσθετη ενίσχυση, οι ταξινομητές μαθαίνουν καλύτερα τα ισχυρά μοτίβα (π.χ. DoS/DDoS) αλλά δυσκολεύονται με τις πιο λεπτές ή σπάνιες επιθέσεις.

Στο Σενάριο I (εκπαίδευση και GAN από 2017), οι κατηγορίες Web Attack - SQL Injection και Infiltration - Communication Victim Attacker ταξινομήθηκαν σε μεγάλο βαθμό ως Benign, παρουσιάζοντας χαμηλό Recall. Το γεγονός αυτό συνδέεται με τη χαμηλή αντιπροσωπευτικότητα των δειγμάτων τους αλλά και με τη μορφολογική ομοιότητα ορισμένων χαρακτηριστικών με κανονική κίνηση. Αντίθετα, άλλες κατηγορίες με περιορισμένα δείγματα αλλά πιο ξεκάθαρο προφίλ ροής απέδωσαν πολύ υψηλά αποτελέσματα.

Στο Σενάριο II (εκπαίδευση και GAN από 2018), οι ίδιες κατηγορίες (SQL Injection, Infiltration - Communication Victim Attacker) παρουσίασαν ξανά χαμηλές επιδόσεις. Επιπλέον, η κατηγορία Infiltration-Dropbox Download ταξινομήθηκε σχεδόν αποκλειστικά ως Benign (Recall 0.01). Αυτό εξηγείται από τη φύση της επίθεσης, που ισοδυναμεί με απλό κατέβασμα μολυσμένου αρχείου και συνεπώς είναι δύσκολο να ξεχωρίσει από φυσιολογική δραστηριότητα.

Στο Σενάριο III (εκπαίδευση και GAN από 2017 και 2018), χαμηλές επιδόσεις καταγράφηκαν στην κατηγορία DoS Slowhttptest (Recall 0.65), η οποία σε αρκετές περιπτώσεις αναγνωρίστηκε λανθασμένα ως Benign. Η κατηγορία PortScan επίσης ταξινομήθηκε ως Benign αλλά και ως Infiltration - NMAP Portscan.

Ωστόσο, η εμπλοκή αυτών των δύο κατηγοριών δεν θεωρείται ουσιαστικό σφάλμα, αφού όπως έχει ήδη αναφερθεί, η εμπλοκή των δύο αυτών κατηγοριών μεταξύ τους δεν θεωρείται ουσιαστικό σφάλμα, αφού στην πράξη σχετίζονται στενά και ο διαχωρισμός τους έγινε αποκλειστικά για λόγους πειραματικού ελέγχου *.

Στα Σενάρια IV-VI όπου εφαρμόστηκε εμπλουτισμός με δεδομένα TVAE, τα αποτελέσματα ήταν εξαιρετικά, με Accuracy F1 0.96–0.99.

Συγκεκριμένα στο Σενάριο IV (εκπαίδευση και GAN από 2017) οι κατηγορίες PortScan και Infiltration - NMAP Portscan διαχωρίστηκαν με επιτυχία, ωστόσο παρατηρήθηκε σύγχυση της NMAP με Benign.

Στο Σενάριο V (εκπαίδευση και GAN από 2018) τα αποτελέσματα ήταν σχεδόν άριστα, με συνολικό F1 0.99. Η κατηγορία Infiltration - NMAP Portscan δεν παρουσίασε ουσιαστικά σφάλματα, ούτε σύγχυση με Benign. Το γεγονός αυτό πιθανόν να οφείλεται στην απουσία της απλής κατηγορίας PortScan στα δεδομένα 2018 καθώς και στη σαφέστερη ροή των δεδομένων της κατηγορίας. Το CSE-CICIDS 2018 θεωρείται βελτιωμένη έκδοση του CICIDS 2017, με αποτέλεσμα οι ροές των επιθέσεων να είναι πιο καθαρές και χαρακτηριστικές.
Τελικό σενάριο της πολυκατηγορικής προσέγγισης (Multiclass), είναι το Σενάριο VI (εκπαίδευση και GAN από 2017 και 2018). Στο συνδυασμένο σενάριο τα αποτελέσματα ήταν ελαφρώς υποδεέστερα σε σχέση με των IV και V. Συγκεκριμένα, η κατηγορία Infiltration-NMAP Portscan και PortScan παρουσίασαν σημαντική σύγχυση με Benign, ενώ η PortScan ταξινομήθηκε και ως Infiltration - NMAP Portscan.

MULTICLASS

ΣΕΝΑΡΙΟ I

	Precision	Recall	F1-score
Micro avg	0.89	0.89	0.89
Macro avg	0.95	0.89	0.90
Weighted avg	0.95	0.89	0.90

ΣΕΝΑΡΙΟ II

	Precision	Recall	F1-score
Accuracy			0.88
Macro avg	0.96	0.88	0.87
Weighted avg	0.96	0.88	0.87

ΣΕΝΑΡΙΟ III

	Precision	Recall	F1-score
Accuracy			0.82
Macro avg	0.93	0.82	0.83
Weighted avg	0.93	0.82	0.83

ΣΕΝΑΡΙΟ IV

	Precision	Recall	F1-score
Accuracy			0.98
Macro avg	0.98	0.98	0.98
Weighted avg	0.98	0.98	0.98

ΣΕΝΑΡΙΟ V

	Precision	Recall	F1-score
Accuracy			0.99
Macro avg	0.99	0.99	0.99
Weighted avg	0.99	0.99	0.99

ΣΕΝΑΡΙΟ VI

	Precision	Recall	F1-score
Accuracy			0.96
Macro avg	0.96	0.96	0.95
Weighted avg	0.96	0.96	0.95

Αποτελέσματα Binary Ταξινόμησης

Από τη σύγκριση των σεναρίων της binary ταξινόμησης (One vs All) προέκυψε ότι τα μοντέλα χωρίς εμπλουτισμό δεδομένων (Σενάρια I–III) είχαν μέτρια απόδοση με Macro/Weighted F1 0.75–0.84. Σε γενικές γραμμές, οι κατηγορίες με μεγάλο όγκο δεδομένων ταξινομήθηκαν με υψηλή ακρίβεια, ενώ οι μειονοτικές κατηγορίες εμφάνισαν χαμηλότερα Recall με κάποιες εξαιρέσεις.

Στο Σενάριο I (εκπαίδευση και GAN από 2017), η κατηγορία Web Attack - SQL Injection δεν ανιχνεύθηκε καθόλου (Recall 0.00), ενώ η Infiltration - Communication Victim Attacker είχε χαμηλό Recall (0.18), γεγονός που δείχνει αδυναμία αναγνώρισης ροών με περιορισμένα

δεδομένα εκπαίδευσης. Στο ίδιο σενάριο, η DoS Slowhttptest εμφάνισε Recall 0.72, με αρκετά δείγματα να ταξινομούνται ως Benign.
Στο Σενάριο II (εκπαίδευση και GAN από 2018), παρατηρήθηκε βελτίωση σε ορισμένες κατηγορίες, ωστόσο η Infiltration - Dropbox Download δεν αναγνωρίστηκε (Recall 0.00), ενώ το Web Attack - SQL Injection διατήρησε χαμηλό Recall (0.19). Η κατηγορία Infiltration - Communication Victim Attacker εμφάνισε καλύτερη επίδοση (Recall 0.71), σε σχέση με το Σενάριο I.
Στο Σενάριο III (εκπαίδευση και GAN από 2017 και 2018), τα αποτελέσματα παρέμειναν μέτρια. Οι κατηγορίες Infiltration - Dropbox Download και Web Attack - SQL Injection παρουσίασαν και πάλι εξαιρετικά χαμηλά Recall (0.00 και 0.04 αντίστοιχα). Επιπλέον, η κατηγόρια Heartbleed είχε Recall 0.20, ενώ η Infiltration - Communication Victim Attacker σημείωσε Recall 0.42. Αντίστοιχα με το πολυκατηγορικό μοντέλο στο αντίστοιχο σενάριο, η PortScan και η Infiltration - NMAP Portscan εμφάνισαν σημαντική σύγχυση μεταξύ τους αλλά και με Benign (Recall 0.46 και 0.72 αντίστοιχα).
Αντίθετα, στα Σενάρια IV–VI, όπου εφαρμόστηκε εμπλουτισμός με δεδομένα TVAE, παρατηρείται σαφής βελτίωση με Macro/Weighted F1 0.92–0.99.
Στο Σενάριο IV (εκπαίδευση και GAN από 2017), η πλειονότητα των κατηγοριών ταξινομήθηκε σχεδόν άριστα, με εξαίρεση την PortScan (Recall 0.16), η οποία μπερδεύτηκε κυρίως με την Infiltration - NMAP Portscan και Benign.
Στο Σενάριο V (εκπαίδευση και GAN από 2018) τα αποτελέσματα ήταν σχεδόν τέλεια (F1 ≈ 0.99). Ακόμα και κατηγορίες που προηγουμένως εμφάνιζαν χαμηλά Recall (Web Attack - SQL Injection, Infiltration - Dropbox Download) παρουσίασαν υψηλή απόδοση. Η Infiltration - NMAP Portscan είχε F1 0.97 χωρίς ουσιαστική σύγχυση με Benign.
Τέλος, στο Σενάριο VI (εκπαίδευση και GAN από 2017 και 2018), τα αποτελέσματα ήταν ελαφρώς υποδεέστερα σε σχέση με το Σενάριο V, με Macro/Weighted F1 = 0.95. Σημαντική σύγχυση παρατηρήθηκε και πάλι μεταξύ PortScan και Infiltration - NMAP Portscan (Recall 0.46 και 0.71 αντίστοιχα), καθώς και με Benign. Ωστόσο, όπως και στα Multiclass σενάρια, η σύγχυση των δύο αυτών κατηγοριών δεν θεωρείται ουσιαστικό σφάλμα, αφού στην πράξη όπως προαναφέρθηκε πρόκειται για το ίδιο είδος επίθεσης

BINARY ONE VS ALL

ΣΕΝΑΡΙΟ I

	Precision	Recall	F1-score
Micro avg	0.83	0.82	0.83
Macro avg	0.89	0.82	0.82
Weighted avg	0.89	0.82	0.82

ΣΕΝΑΡΙΟ II

	Precision	Recall	F1-score
Micro avg	0.84	0.84	0.84
Macro avg	0.89	0.84	0.83
Weighted avg	0.89	0.84	0.83

ΣΕΝΑΡΙΟ III

	Precision	Recall	F1-score
Micro avg	0.75	0.74	0.75
Macro avg	0.89	0.74	0.75
Weighted avg	0.89	0.74	0.75

ΣΕΝΑΡΙΟ IV

	Precision	Recall	F1-score
Micro avg	0.94	0.93	0.93
Macro avg	0.96	0.93	0.92
Weighted avg	0.96	0.93	0.92

ΣΕΝΑΡΙΟ V				ΣΕΝΑΡΙΟ VI			
	Precision	Recall	F1-score		Precision	Recall	F1-score
Micro avg	0.99	0.99	0.99	Micro avg	0.96	0.95	0.95
Macro avg	0.99	0.99	0.99	Macro avg	0.97	0.95	0.95
Weighted avg	0.99	0.99	0.99	Weighted avg	0.97	0.95	0.95

ΣΥΖΗΤΗΣΗ

Συνολικά, σημαντική επίδραση στα αποτελέσματα των μοντέλων είχαν τα συνθετικά δεδομένα που παρήχθησαν μέσω CTGANs και TVAE. Στην περίπτωση των CTGANs, ο στόχος ήταν η δημιουργία συνθετικών δειγμάτων πάνω στα οποία έγινε πρόβλεψη, με σκοπό να αξιολογηθεί ο βαθμός στον οποίο το εκπαιδευμένο μοντέλο μπορούσε να γενικεύσει σε δεδομένα που δεν προέρχονταν απευθείας από το αρχικό σύνολο. Η ενσωμάτωση συνθετικών δειγμάτων βοήθησε στη βελτίωση της αναλογίας μεταξύ των κατηγοριών, ιδίως στις περιπτώσεις επιθέσεων με περιορισμένη παρουσία πραγματικών δεδομένων, προσφέροντας έτσι καλύτερη εκπαίδευση και γενίκευση. Ωστόσο, τα αποτελέσματα έδειξαν ότι η απόδοση των μοντέλων εξαρτήθηκε άμεσα από την ποιότητα και τη ρεαλιστικότητα των παραγόμενων δεδομένων: όταν τα δείγματα ήταν πιο αντιπροσωπευτικά, η γενίκευση βελτιωνόταν, ενώ σε περιπτώσεις μεγαλύτερης απόκλισης παρατηρήθηκαν μειώσεις στην ακρίβεια. Συνεπώς, τα GANs και τα TVAE λειτούργησαν ως κρίσιμοι μηχανισμοί που επηρέασαν τόσο θετικά όσο και αρνητικά την τελική απόδοση των μοντέλων.

Τα αποτελέσματα δείχνουν ότι τα προτεινόμενα συστήματα είναι ικανά να διαχωρίζουν αποτελεσματικά τις περισσότερες κατηγορίες επιθέσεων όταν υπάρχει επαρκής αντιπροσώπευση δειγμάτων. Σε σενάρια όπου τα δεδομένα ήταν πλούσια, το μοντέλο παρουσίασε πολύ υψηλά scores για πλήθος επιθέσεων. Όταν όμως ορισμένες κατηγόριες ήταν υποεκπροσωπούμενες, η απόδοση χωρίς υπερδειγματοληψία ήταν σημαντικά φτωχότερη. Το αποτέλεσμα ήταν αναμενόμενο, καθώς το μοντέλο δεν είχε αρκετά παραδείγματα για να μάθει τα συγκεκριμένα μοτίβα. Σε αυτή την περίπτωση, ακόμα και η ανίχνευση λίγων κακόβουλων ροών έχει πρακτική αξία. Το να εντοπίσει ένα IDS έστω και μερικές ροές από μια επίθεση μπορεί να αρκεί για να ενεργοποιήσει περαιτέρω μέτρα αντίδρασης και ανάλυσης. Για αυτόν τον λόγο η λειτουργική προτεραιότητα του συστήματος ήταν η ικανότητα του να αποκρίνεται σωστά σε πραγματικές ροές και να μην ενεργοποιεί αδικαιολόγητους συναγερμούς σε κανονική κίνηση. Επομένως, τα χαμηλά Precision της κατηγορίας Benign δεν επηρεάζουν ουσιαστικά τη λειτουργικότητα του συστήματος, εφόσον το Recall παραμένει υψηλό και κανονικές ροές δεν χαρακτηρίζονται επιθέσεις.

Συγκριτικά, η πολυκατηγορική (Multiclass) προσέγγιση παρουσίασε συνολικά καλύτερα αποτελέσματα, όπως φαίνεται και από τα υψηλότερα F1-scores. Ακόμη και κατηγορίες με περιορισμένο αριθμό δειγμάτων εντοπίστηκαν, έστω και μερικώς, γεγονός που είναι σημαντικό για τη λειτουργικότητα ενός IDS. Παρόλα αυτά, το βασικό μειονέκτημα της πολυκατηγορικής προσέγγισης ήταν ότι αντί να κατατάσσει δείγματα στην κατηγόρια unknown, παρουσίαζε μεγαλύτερη σύγχυση μεταξύ διαφορετικών κατηγοριών επιθέσεων (αν και όχι σε ακραίο βαθμό). Αντίθετα, τα δυαδικά μοντέλα One vs All είχαν την τάση να αποδίδουν περισσότερα δείγματα στην κατηγορία unknown και να μην εμπλέκουν επιθέσεις

μεταξύ τους. Το γεγονός αυτό πιθανόν να οφείλεται στο ότι χρησιμοποιήθηκε κοινό κατώφλι εμπιστοσύνης, ώστε να είναι εφικτή η βέλτιστη και δίκαιη σύγκριση μεταξύ των σεναρίων.
Είναι προφανές λοιπόν ότι και οι δύο προσεγγίσεις αποδίδουν εξίσου ικανοποιητικά αποτελέσματα. Η κάθε προσέγγιση έχει διαφορετικά πλεονεκτήματα και μειονεκτήματα, ανάλογα με τον στόχο και τη λειτουργία του εκάστοτε IDS. Σε πιο ρεαλιστικά σενάρια, όπου τα δεδομένα θα προέρχονται από stress testing και θα περιλαμβάνουν μεγαλύτερο όγκο καθώς και μεγαλύτερη ποικιλία ροών ανά κατηγορία, μπορούν να αναμένονται ακόμη πιο αντιπροσωπευτικά αποτελέσματα. Η επιλογή των κατάλληλων παραμέτρων, όπως το κατώφλι εμπιστοσύνης, μπορεί να βελτιστοποιηθεί μέσα από δοκιμές, ενώ στην παρούσα εργασία η ανάλυση περιορίστηκε στα διαθέσιμα δεδομένα και σε συγκεκριμένες ροές με περιορισμένες παραλλαγές ανά κατηγορία.
Αξίζει να σημειωθεί ότι πραγματοποιήθηκε και ανάλυση ερμηνευσιμότητας με SHAP, η οποία ανέδειξε τα χαρακτηριστικά που συμβάλλουν περισσότερο στη διάκριση μεταξύ των κατηγοριών. Επειδή η πλήρης ανάλυση αποδείχθηκε ιδιαίτερα εκτεταμένη και απαιτούσε ξεχωριστή μελέτη, δεν συμπεριλήφθηκε στην παρούσα εργασία. Ωστόσο, τα αποτελέσματα αυτής της προσέγγισης μπορούν να αποτελέσουν τη βάση για μελλοντική ερευνητική προσπάθεια που θα εστιάζει στη κατανόηση των αποφάσεων του IDS και στην συσχέτιση τους με τα μοτίβα την κάθε ροής.
Τέλος, αναγνωρίζονται οι περιορισμοί και δυνατότητες επέκτασης. Εάν υπήρχε γνώση και υποστήριξη για stress testing, θα μπορούσε να εμπλουτιστεί το πειραματικό μέρος με ρεαλιστικά σενάρια, παραλλαγές επιθέσεων και περισσότερά δεδομένα επιθέσεων. Η μη ένταξη τέτοιων δοκιμών οφείλεται στην ανάγκη για καθοδήγηση και την τεχνική πολυπλοκότητα και κινδυνότητα που συνεπάγεται το stress testing χωρίς κατάλληλο περιβάλλον και εποπτεία.
Θεωρώ ότι μια μελλοντική, στοχευμένη εργασία που θα επικεντρωθεί στην προσομοίωση πραγματικών συνθηκών (stress and load testing) και στην παρουσίαση SHAP αναλύσεων θα εμπλουτίσει σημαντικά τα αποτελέσματα και θα ενισχύσει την αξιοπιστία της πρότασης.
Συνοψίζοντας, η εργασία απέδειξε τη λειτουργική ικανότητα των Μοντέλων Μηχανικής Μάθησης σε ποικίλα σενάρια, ανέδειξε τους περιορισμούς στις υποεκπροσωπημένες κατηγορίες και πρότεινε ρεαλιστικές κατευθύνσεις για μελλοντική έρευνα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Alasmary, W., & Alharthi, R. (2024). From Creeper to Ransomware: The Evolution of Malware. Review paper.. https://doi.org/10.13140/RG.2.2.17919.83369

2. Symantec. (2021). Internet Security Threat Report. Symantec. https://docs.broadcom.com/doc/istr-03-jan-en

3. Li, Y., Guo, L., Guo, Y., Chen, C., & Qin, Z. (2021). A comprehensive review study of cyber-attacks and defenses. Sustainable Cities and Society, 72, 103005. https://doi.org/10.1016/j.scs.2021.103005

4. Oz, H., Aris, A., Levi, A., & Uluagac, A. S. (2021). A Survey on Ransomware: Evolution, Taxonomy, and Defense Solutions. arXiv preprint arXiv:2102.06249. https://arxiv.org/pdf/2102.06249

5. Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. Computers & Security, 38, 97–102. https://doi.org/10.1016/j.cose.2013.04.004

6. Gyamfi, E., & Jurcut, A. (2022). Intrusion Detection in Internet of Things Systems: A Review on Design Approaches Leveraging Multi-Access Edge Computing, Machine Learning, and Σύνολα δεδομένων. Sensors, 22(10), 3744. https://doi.org/10.3390/s22103744

7. Ahmed, U., Nazir, M., Sarwar, A., Ali, T., Aggoune, E. H. M., Shahzad, T., & Khan, M. A. (2025). Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering. *Scientific Reports*, 15, 85866. https://doi.org/10.1038/s41598-025-85866-7

8. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. ICISSP 2018 – 4th International Conference on Information Systems Security and Privacy, 108-116. https://www.scitepress.org/papers/2018/66398/66398.pdf

9. Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://ieeexplore.ieee.org/document/7307098

10. UNB. IDS-2017. https://www.unb.ca/cic/σύνολα δεδομένων/ids-2017.html

11. UNB. IDS-2018. https://www.unb.ca/cic/σύνολα δεδομένων/ids-2018.html

12. Liu, Lisa; Engelen, Gints; Lynar, Timothy; Essam, Daryl; Joosen, Wouter. (2022). Error Prevalence in NIDS: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018. IEEE Conference on Communications and Network Security (CNS). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9947235

    - CIC-IDS 2017 dataset. (2022). Improved CIC-IDS 2017. https://intrusion-detection.distrinet-research.be/CNS2022/CICIDS2017.html

- CSE-CIC-IDS 2018 (Improved). (2022). Improved CSE-CIC-IDS 2018. https://intrusion-detection.distrinet-research.be/CNS2022/CSECICIDS2018.html

13. Gu, Y., & McCallum, A. (2010). Network anomaly detection: Flow-based or packet-based approach arXiv. https://arxiv.org/abs/1007.1266

14. Umer, M. F. (2017). Flow-based intrusion detection: Techniques and challenges. Computers & Security. https://doi.org/10.1016/j.cose.2017.11.012

15. Yang, B., Arshad, M. H., & Zhao, Q. (2022). Packet-Level and Flow-Level Network Intrusion Detection Based on Reinforcement Learning and Adversarial Training. Algorithms, 15(12), 453. https://doi.org/10.3390/a15120453

16. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 785–794. https://doi.org/10.1145/2939672.2939785

17. Ring, M., Wunderlich, S., Grüdl, D., Landes, D., & Hotho, A. (2019). Flow-based network traffic generation using Generative Adversarial Networks. Computers & Security, 82, 156–172. https://doi.org/10.1016/j.cose.2018.12.012

18. CISA. (2013). Low Orbit Ion Cannon (LOIC) denial of service attack tool. Cybersecurity and Infrastructure Security Agency. https://www.cisa.gov/news-events/analysis-reports/ar13-164a

19. Imperva. (n.d.). Low Orbit Ion Cannon (LOIC). Imperva Cybersecurity Learning Center. https://www.imperva.com/learn/application-security/loic-low-orbit-ion-cannon/

20. CISA. (2013). High Orbit Ion Cannon (HOIC) denial of service tool. Cybersecurity & Infrastructure Security Agency. https://www.cisa.gov

21. Imperva. (n.d.). High Orbit Ion Cannon (HOIC). Imperva Cybersecurity Knowledge Base. https://www.imperva.com

22. Mirkovic, J., & Reiher, P. (2004). A taxonomy of DDoS attack and DDoS defense mechanisms. ACM SIGCOMM Computer Communication Review, 34(2), 39–53. https://www.princeton.edu/~rblee/ELE572Papers/Fall04Readings/DDoSmirkovic.pdf

23. Cloudflare Learning. (n.d.). Slowloris – what it is and how it works. https://www.cloudflare.com/learning/ddos/ddos-attack-tools/slowloris/

24. Wallarm. (2024). What is HULK – HTTP Unbearable Load King https://www.wallarm.com/what/what-is-hulk-http-unbearable-load-king

25. Singh, S. K., et al. (2024). Brute-Force SSH Attacks In The Wild And How To Stop Them. In Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation. https://www.usenix.org/system/files/nsdi24-singh-sachin.pdf

26. Najafabadi, M. M., Khoshgoftaar, T. M., Calvert, C., & Kemp, C. (2015). Detection of SSH brute force attacks using aggregated Netflow data. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 283–288. https://doi.org/10.1109/ICMLA.2015.20

27. Hamza, A. A. (2024). Detecting Brute Force Attacks Using Machine Learning. BIO Web Conf.Volume 97, 2024 Fifth International Scientific Conference of Alkafeel University (ISCKU 2024). https://doi.org/10.1051/bioconf/20249700045

28. National Cyber Security Centre. (n.d.). Denial of service (DoS) – guidance collection. GOV.UK. https://www.ncsc.gov.uk/collection/denial-service-dos-guidance-collection

29. Kali Linux (2024). slowhttptest. https://www.kali.org/tools/slowhttptest/

30. Kali Linux (2024). GoldenEye (Kali Tools). https://www.kali.org/tools/goldeneye/

31. Kali Linux. Heartleech. https://www.kali.org/tools/heartleech/

32. Kali Linux. (n.d.). Patator Kali Tools documentation. https://www.kali.org/tools/patator/

33. National Institute of Standards and Technology. (n.d.). Brute Force Password Attack (CSRC glossary). https://csrc.nist.gov/glossary/term/brute_force_password_attack

34. OWASP Foundation. (n.d.). Brute-force attack. OWASP. https://owasp.org/www-community/attacks/Brute_force_attack

35. George, R. (2025). Detecting Brute-Force Attacks on SSH and FTP Protocol Using Machine Learning –thesis/report. DiVA portal (pdf). https://www.diva-portal.org/smash/get/diva2%3A1981478/FULLTEXT01.pdf

36. Kali Linux. (2025, June 16). dvwa. Kali Linux tools. https://www.kali.org/tools/dvwa/

37. Halfond, W. G. J., Viegas, J., & Orso, A. (2006). A classification of SQL injection attacks and countermeasures (conference paper). https://viterbiweb.usc.edu/~halfond/papers/halfond06issse.pdf

38. Melicher, W., et al. (2018). Toward Detecting and Preventing DOM Cross-Site Scripting (NDSS 2018). https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_07A-4_Melicher_paper.pdf

39. National Institute of Standards and Technology. (n.d.). Backdoor. In Computer Security Resource Center Glossary. https://csrc.nist.gov/glossary/term/backdoor

40. Rapid7. (n.d.). Metasploit Overview. https://docs.rapid7.com/metasploit/msf-overview/

41. Command & Control: Understanding, Denying and Detecting J. Gardiner et al. (2014). https://arxiv.org/pdf/1408.1136

42. Abuse of Cloud-Based and Public Legitimate Services as Command-and-Control (C&C) Infrastructure: A Systematic Literature Review – Turki Al lelah, George Theodorakopoulos et al. (2023). https://doi.org/10.3390/jcp3030027

43. Nmap. (n.d.). Nmap Reference Guide. https://nmap.org/book/man.html

44. M. I. Kareem, M. J. K. Abood, and K. Ibrahim, "Machine learning-based PortScan attacks detection using OneR classifier," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 6, pp. 3690–3696, Dec. 2023. https://doi.org/10.11591/eei.v12i6.4142

45. Lee, C. B., Roedel, C., & Silenok, E. (2003). Detection and characterization of port scan attacks. University of California, San Diego. https://cseweb.ucsd.edu/~clbailey/PortScans.pdf

46. Codenomicon. (2014). The Heartbleed Bug. https://heartbleed.com/

47. Durumeric, Z., Kasten, J., Adrian, D., Halderman, J. A., Bailey, M., Li, F., Weaver, N., Amann, J., Beekman, J., Payer, M., Paxson, V., & Sommer, R. (2014). The matter of Heartbleed. Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14), 475–488. https://doi.org/10.1145/2663716.2663755

48. Silva, S. S., Silva, R. M., Pinto, R. C., & Salles, R. M. (2013). Botnets: A survey. Computer Networks, 57(2), 378–403. https://doi.org/10.1016/j.comnet.2012.07.021

49. Bergmann, D., & Stryker, C. (2024, June 12). What is a variational autoencoder? IBM Think. https://www.ibm.com/think/topics/variational-autoencoder

50. Varughese, J. (2025, May 06). What are generative adversarial networks (GANs)? IBM Think. https://www.ibm.com/think/topics/generative-adversarial-networks

51. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. arXiv. https://doi.org/10.48550/arXiv.1907.00503

52. Kapoor, N. (2024, September 3). How to create synthetic data with CTGAN. Sustainability Methods. https://sustainabilitymethods.org/index.php/How_To_Create_Synthetic_Data_with_CTGAN

53. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 27, 2672–2680. http://dx.doi.org/10.1145/3422622

54. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1312.6114

55. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

56. SDV Developers. (n.d.). CTGANSynthesizer. Synthetic Data Vault (SDV). https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer

57. SDV Developers. (n.d.). TVAESynthesizer. Synthetic Data Vault (SDV). https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/tvaesynthesizer

58. Fonseca, J. R., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. Journal of Big Data, 10, Article number:115. https://doi.org/10.1186/s40537-023-00792-7

59. Hansen, C., Seedat, N., van der Schaar, M., & Petrovic, J. (2023). Reimagining synthetic tabular data generation. In Advances in Neural Information Processing Systems,36 https://doi.org/10.48550/arXiv.2310.16981

60. Khan, A., Malik, S., & Almotiri, J. (2024). Rigorous experimental analysis of tabular data generated using generative models. International Journal of Advanced Computer Science and Applications (IJACSA), 15 (4), 1250–1262. https://thesai.org/Downloads/Volume15No4/Paper_125-Rigorous_Experimental_Analysis_of_Tabular_Data_Generated.pdf

61. Zhu, J., Zhao, X., Birke, R., & Chen, L. Y. (2022). Permutation-invariant tabular data synthesis. arXiv preprint. https://arxiv.org/abs/2211.09286

62. Abdelrahman, Y., & Kaya, M. (2025). A comparative study on synthetic tabular data generation. arXiv preprint. https://arxiv.org/abs/2502.14523

63. Murthy, S., & Schmidt, F. (2023). MargCTGAN: A “marginally” better CTGAN for the low sample regime. In Proceedings of the German Conference on Pattern Recognition (GCPR 2023) . https://www.dagm-gcpr.de/fileadmin/dagm-gcpr/pictures/2023_Heidelberg/Paper_MainTrack/065.pdf

64. Ahmad, W., Khan, I., & Rehman, A. U. (2024). Challenges of using synthetic data generation methods for tabular data. Applied Sciences, 14(14), 5975. https://doi.org/10.3390/app14145975

65. Arize AI. (2023, August 21). Kolmogorov–Smirnov test for AI. https://arize.com/blog-course/kolmogorov-smirnov-test/

66. U.S. Department of Commerce, National Institute of Standards and Technology. (n.d.). Kolmogorov-Smirnov Goodness-of-Fit Test. In NIST/SEMATECH e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

67. Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21–33. https://www.researchgate.net/publication/267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests

68. Singh, A., Yadav, A., & Rana, A. (2013). K-means with three different distance metrics. International Journal of Computer Applications, 67(10), 13–17. https://doi.org/10.5120/11430-6785

69. Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences, 1(4), 300–307. http://www.fisica.edu.uy/~cris/teaching/Cha_pdf_distances_2007.pdf

70. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

71. Wikipedia. (n.d.). Mean squared error. https://en.wikipedia.org/wiki/Mean_squared_error

72. Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. International Advanced Research Journal in Science, Engineering and Technology, 2(3), 20–22. https://doi.org/10.17148/IARJSET.2015.2305

73. Mitchell, R., Adinets, A., Rao, T., & Frank, E. (2018). XGBoost: Scalable GPU accelerated learning. arXiv preprint, arXiv:1806.11248. https://arxiv.org/abs/1806.11248

74. Wiens, M. (2025). A tutorial and use case example of the eXtreme Gradient Boosting (XGBoost) algorithm for classification and regression. Clinical and Translational Science. https://doi.org/10.1111/cts.70172

75. XGBoost Developers. XGBoost documentation. https://xgboost.readthedocs.io

76. Gouveia, A., & Correia, M. (2020). Network intrusion detection with XGBoost. In Recent advances in security, privacy, and trust for Internet of Things (IoT) and cyber-physical systems (CPS) (pp. 137–166). Chapman & Hall/CRC. https://doi.org/10.1201/9780429270567-6

77. Hosain, Y., & Çakmak, M. (2025). XAI-XGBoost: An innovative explainable intrusion detection approach for securing Internet of Medical Things systems. Scientific Reports, 15, Article 22278. https://doi.org/10.1038/s41598-025-07790-0

78. Ali, Z. A., Abduljabbar, Z. H., Tahir, H. A., Sallow, A. B., & Almufti, S. M. (2023). Exploring the power of eXtreme Gradient Boosting algorithm in machine learning: A review. Academic Journal of Nawroz University (AJNU), 12(2). https://doi.org/10.25007/ajnu.v12n2a1612