Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Διατμηματικο Προγραμμα Μεταπτυχιακων Σπουδων
Επιστημη Δεδομενων και Μηχανικη Μαθηση

Χρήση Federated Learning για τη Συνεργατική Ανίχνευση Δικτυακών Απειλών με Υβριδικό Μοντέλο Βαθιάς Μάθησης

Μεταπτυχιακη Διπλωματικη Εργασια

του

ΜΠΙΜΠΑ Η. ΗΛΙΑ

Επιβλέπων: Βασίλειος Μάγκλαρης
Ομότιμος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Διατμηματικο Προγραμμα Μεταπτυχιακων Σπουδων
Επιστημη Δεδομενων και Μηχανικη Μαθηση

Χρήση Federated Learning για τη Συνεργατική Ανίχνευση Δικτυακών Απειλών με Υβριδικό Μοντέλο Βαθιάς Μάθησης

Μεταπτυχιακη Διπλωματικη Εργασια

του

ΜΠΙΜΠΑ Η. ΗΛΙΑ

Επιβλέπων: Βασίλειος Μάγκλαρης
Ομότιμος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Ιουλίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Βασίλειος Μάγκλαρης	Ευστάθιος Συκάς	Γεώργιος Στάμου
Ομότιμος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Διατμηματικο Προγραμμα Μεταπτυχιακων Σπουδων
Επιστημη Δεδομενων και Μηχανικη Μαθηση

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

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

(Υπογραφή)

.....
Ηλίας Μπίμπας

3 Ιουλίου 2022

Περίληψη

Η συνεχής ανάπτυξη των δικτύων υπολογιστικών συστημάτων συνοδεύεται από την ανάδειξη νέων απειλών για την ασφαλή λειτουργία τους. Σε αυτό το πλαίσιο, τα συστήματα ανίχνευσης εισβολής έχουν σημαντικό ρόλο στην καταπολέμηση των απειλών και την προστασία των συστημάτων που διασυνδέονται. Η πρόοδος στον τομέα της βαθιάς μάθησης οδήγησε το ερευνητικό ενδιαφέρον στην μελέτη χρήσης της για την δημιουργία αξιόπιστων συστημάτων ανίχνευσης εισβολής. Αυτά τα συστήματα ασφαλείας χρειάζονται συχνά σημαντικό όγκο δεδομένων για να εκπαιδευτούν, όμως ο διαμοιρασμός δεδομένων και η αξιοποίηση τους ενέχει προκλήσεις για την διασφάλιση της ιδιωτικότητας. Αυτό επιχειρεί να αντιμετωπίσει η εφαρμογή federated learning στα πλαίσια ανάπτυξης συστημάτων ανίχνευσης εισβολής. Η συγκεκριμένη τεχνική μάθησης προσφέρει ένα περιβάλλον εκπαίδευσης μοντέλων βαθιάς μάθησης, στο οποίο οι συμμετέχοντες συνεργάζονται για την εκπαίδευση ενός κοινού μοντέλου, χωρίς ωστόσο να μοιράζονται τα πιθανώς ευαίσθητα δεδομένα εκπαίδευσης που διαθέτουν.

Στην παρούσα διπλωματική εργασία, σχεδιάζεται αρχικά ένα υβριδικό μοντέλο βαθιάς μάθησης για την χρήση του ως σύστημα ανίχνευσης εισβολής, που αξιοποιεί τόσο επισημασμένα όσο και μη επισημασμένα δεδομένα, τα οποία απαντώνται πιο συχνά στην πράξη. Στην συνέχεια, υλοποιείται αρχιτεκτονική federated learning για την συνεργατική εκπαίδευσή του. Στόχος είναι η παρουσίαση των πλεονεκτημάτων της χρήσης federated learning, στα πλαίσια εκπαίδευσης σύνθετων δικτύων για συστήματα ανίχνευσης εισβολής, και της ανάδειξης του γεγονότος ότι μπορεί να προστατευτεί η ιδιωτικότητα των συνεργαζόμενων υποσυστημάτων χωρίς σημαντική μείωση στην επίδοση του μοντέλου που προκύπτει. Η αξιολόγηση του συνεργατικού μοντέλου πραγματοποιείται ως προς την ανίχνευση εισβολής αλλά και ως προς την ταξινόμηση των απειλών που εντοπίζονται. Ακόμα, εξετάζεται η συμπεριφορά αυτού του μοντέλου κατά την εκπαίδευση, καθώς και η επίδοσή του όταν τα δεδομένα που χρησιμοποιούνται δεν ικανοποιούν την i.i.d. υπόθεση. Τα αποτελέσματα των πειραμάτων δείχνουν πως το συνεργατικό μοντέλο παραμένει αξιόπιστο όταν εκπαιδεύεται στα πλαίσια federated learning, ενώ η επίδοση του είναι ανώτερη από αυτήν όμοιων μοντέλων που εκπαιδεύονται ατομικά, όταν δεν ικανοποιείται η i.i.d. υπόθεση για τα δεδομένα.

Λέξεις Κλειδιά

Σύστημα ανίχνευσης εισβολής, βαθιά μάθηση, υβριδικό μοντέλο, ομοσπονδιακή μάθηση, μη ανεξάρτητα και ταυτόσημα κατανεμημένα δεδομένα

Abstract

The continuous development of computer networks is followed by the emergence of new threats to their secure operation. In this context, intrusion detection systems have a vital role in the mitigation of such threats and the protection of the connected systems. The advances in deep learning led the scientific community to study its use for reliable intrusion detection systems. These security systems usually rely on a significant amount of data for their training. However, data handling and sharing pose a significant risk to privacy. Federated learning attempts to face this challenge in the context of intrusion detection systems. This learning technique provides an environment for training deep learning models, without the need for the cooperating participants to share their sensitive training data.

In this diploma thesis, a baseline hybrid model is initially designed for its use in intrusion detection, that utilizes both labeled and unlabeled data, that are usually more available. Afterwards, a federated learning architecture is implemented for the cooperative training of the hybrid model. The goal is to demonstrate the advantages of using federated learning for training deep networks to be used as intrusion detection systems, and to show that the privacy of the participants can be protected with minimal impact on the model's performance. The evaluation of the cooperative model targets both intrusion detection and threat classification. The behavior of the model is examined, as well as its performance when the data used for training do not satisfy the i.i.d. assumption. From the experiments, it is demonstrated that the federated learning model remains reliable, while outperforming the same standalone models when they are trained with non-i.i.d. data whose distribution varies.

Keywords

Intrusion detection system, deep learning, hybrid model, federated learning, non-i.i.d. data

στον πατέρα μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον ομότιμο καθηγητή κ. Μάγκλαρη για την επίβλεψη της παρούσης εργασίας, καθώς και τον υποψήφιο διδάκτορα κ. Κωστόπουλο για τις χρήσιμες συμβουλές του.

Ευχαριστώ θερμά την οικογένειά μου, για την στήριξη που μου παρείχε κατά την εκπόνηση αυτής της εργασίας, αλλά και για την βοήθεια που μου προσέφερε καθ' όλη την διάρκεια των σπουδών μου.

Αθήνα, Ιούλιος 2022

Ηλίας Μπίμπας

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η συνεχής ανάπτυξη των δικτύων και των δυνατοτήτων τους, καθώς και των υπολογιστικών συστημάτων που τα χρησιμοποιούν, οδηγούν ολοένα και περισσότερους τομείς της ανθρώπινης καθημερινότητας στο να βασίζονται στην ομαλή και ασφαλή λειτουργία τους. Μερικά μόνο από τα καθιερωμένα συστήματα που στηρίζονται στην ορθή παροχή υπηρεσιών τους είναι ακαδημαϊκά ιδρύματα, data centers, cloud based service providers, κρατικές δομές, υπηρεσίες πληροφοριών, επιχειρήσεις ανάπτυξης λογισμικού, ιδρύματα παροχής υπηρεσιών υγείας καθώς και ετερογενή δίκτυα του Internet of Things. Η αξιοποίηση όμως τέτοιων συστημάτων συνοδεύεται και από σημαντικές προκλήσεις. Για παράδειγμα, είναι πλέον αναγκαία η διαχείριση σημαντικού όγκου δεδομένων, είτε αυτός προέρχεται από edge devices, είτε από χρήστες. Η προστασία αυτών των δεδομένων, καθώς και η απρόσκοπτη λειτουργία των υπολογιστικών συστημάτων εξαρτάται από την ελαχιστοποίηση απειλών που προσβλέπουν στην παραβίαση των βασικών αρχών ασφαλούς λειτουργίας τους: εμπιστευτικότητα, ακεραιότητα και διαθεσιμότητα (Confidentiality, Integrity, and Availability, CIA) [1]. Η αύξηση των δεδομένων καθώς και το γεγονός ότι οι επιθέσεις σε αυτά τα δίκτυα γίνονται συνεχώς πιο εκλεπτυσμένες, έχει οδηγήσει την ακαδημαϊκή κοινότητα και την βιομηχανία σε αναζήτηση καινοτόμων τεχνικών ασφαλείας για να τις αντιμετωπίσουν [2].

Σε αυτήν την προσπάθεια, εξέχοντα ρόλο έχουν τα συστήματα ανίχνευσης εισβολής (intrusion detection systems, IDS). Με αφετηρία πρωτοπόρες προσπάθειες (όπως στο [3]), έχουν γίνει σημαντικά βήματα για την μοντελοποίηση και τον ορισμό των απειλών ασφαλείας, της ανίχνευσης και της αντιμετώπισής τους, όπως για παράδειγμα η δημοσίευση [4] από το National Institute of Standards and Technology (NIST). Ανάλογα με τις σχεδιαστικές επιλογές και τον τρόπο λειτουργίας τους, τα συστήματα ανίχνευσης εισβολής μπορούν να κατηγοριοποιηθούν με διαφορετικά κριτήρια. Ένας διαχωρισμός που συναντάται συχνά στην πράξη τα διαχωρίζει βάσει των πόρων που προστατεύουν και για τον οποίο αντλούν πληροφορίες. Τα *host-based* IDS (HIDS) στοχεύουν κυρίως στην προστασία συγκεκριμένου μηχανήματος που κρίνεται ευάλωτο σε επιθέσεις. Προς αυτόν το σκοπό, αξιοποιούν τοπικές πληροφορίες σχετικά με την λειτουργία του host (π.χ. αρχεία καταγραφής συστήματος και παρακολούθηση πόρων) και των χρηστών που αλληλεπιδρούν με αυτό. Η αύξηση όμως της διασύνδεσης συστημάτων, που ακολούθησε την ανάπτυξη των δικτύων, οδήγησε στην ανάγκη σχεδιασμού συστημάτων ανίχνευσης που θα παρείχαν προστασία στα επιμέρους διασυνδεδεμένα υποσυστήματα και την μεταξύ τους επικοινωνία από νέες απειλές που προέκυπταν. Αυτόν τον ρόλο έχουν τα *network-based* IDS (NIDS) τα οποία δεν περιορίζονται στην προστασία μόνο συ-

γκεκριμένου host, αλλά αξιοποιούν πληροφορίες από την δικτυακή κίνηση, για παράδειγμα από τα πακέτα που ανταλλάσσονται μεταξύ υποσυστημάτων. Ασφαλώς, έχουν υλοποιηθεί και μεικτού τύπου IDS που συνδυάζουν τεχνικές και από τις δύο βασικές κατηγορίες [5].

Η ραγδαία ανάπτυξη της μηχανικής και βαθιάς μάθησης σε συνδυασμό με την αύξηση των υπολογιστικών δυνατοτήτων και της καλύτερης διαχείρισης μεγάλου όγκου δεδομένων, οδηγούν τους ερευνητές στον χώρο της ασφάλειας στο να εξερευνήσουν την χρήση τέτοιων τεχνικών σε συστήματα ανίχνευσης εισβολής. Προς αυτό συνηγορεί η εμπειρικά αποδεδειγμένη ικανότητα μοντέλων να ανιχνεύουν ανωμαλίες αλλά και να κατηγοριοποιούν τα δεδομένα που δέχονται (π.χ. δικτυακή κίνηση). Πλέον, η αποτελεσματική χρήση τέτοιων τεχνικών για την ανάπτυξη συστημάτων ανίχνευσης εισβολής έχει αναχθεί σε πεδίο με μείζον ερευνητικό ενδιαφέρον, εκμεταλλευόμενοι τόσο επιβλεπόμενη όσο και μη επιβλεπόμενη μάθηση [6], ανάλογα με τις απαιτήσεις και την αρχιτεκτονική της εφαρμογής αλλά και την διαθεσιμότητα αξιόπιστων επισημασμένων δεδομένων.

Παρά την διάδοση τεχνικών μηχανικής (και βαθιάς) μάθησης για την υλοποίηση συστημάτων ανίχνευσης, υπάρχουν αρκετές προκλήσεις που πρέπει να αντιμετωπιστούν. Για παράδειγμα, παρατηρείται σχετική έλλειψη πρόσφατων δημόσιων dataset, μεταξύ άλλων, για λόγους διασφάλισης της ιδιωτικότητας των χρηστών που παρήγαγαν τα δεδομένα. Η ευρεία χρήση τέτοιων μοντέλων στον πραγματικό κόσμο δεν είναι εγγυημένο ότι θα είναι εξίσου αποδοτική, εν μέρει επειδή αναπτύσσονται με δεδομένα που μπορεί να μην αντιπροσωπεύουν πλήρως την ποικιλία των απειλών που υπάρχουν. Για παράδειγμα, ένας οργανισμός μπορεί να χρησιμοποιήσει πραγματικά ή συνθετικά δεδομένα εκπαίδευσης που δεν περιέχουν συγκεκριμένες κλάσεις κακόβουλων δεδομένων. Ακόμα, τέτοια συστήματα ασφαλείας θα πρέπει να είναι σε θέση να αποδίδουν αξιόπιστα και σε περιπτώσεις όπου τα δεδομένα είναι imbalanced, καθώς και να περιορίζουν τις περιπτώσεις ψευδώς θετικών ευρημάτων. Ένα τελευταίο παράδειγμα πρόκλησης είναι ο σχεδιασμός μοντέλων τα οποία πετυχαίνουν ιδανική ισορροπία επιδόσεων και αποδοτικότητας, ώστε να παραμένουν αποτελεσματικά χωρίς να απαιτούν ή να καταχρώνται υπερβολικούς πόρους [6, 7].

Για την αντιμετώπιση τέτοιου είδους προκλήσεων, επιχειρείται πρόσφατα η εφαρμογή μεθόδων federated learning για την ανάπτυξη συστημάτων ανίχνευσης εισβολής. Συνοπτικά, ο όρος federated learning περιγράφει μια τεχνική βαθιάς μάθησης που στηρίζεται σε κατανεμημένη, συνεργατική εκπαίδευσης ενός κοινού μοντέλου, χωρίς την ανάγκη οι συμμετέχοντες να μοιραστούν τα δεδομένα που διαθέτουν με τα υπόλοιπα μέλη της ομοσπονδίας, μειώνοντας έτσι τον κίνδυνο να παραβιαστεί η ιδιωτικότητά τους. Μεταξύ άλλων πλεονεκτημάτων που θα αναλυθούν εκτενέστερα στην συνέχεια, η χρήση federated learning για την ανάπτυξη IDS υποστηρίζεται από τους ερευνητές για την δυνατότητα προσαρμογής της σε πιθανή ετερογένεια των χαρακτηριστικών των συστημάτων που συμμετέχουν καθώς και για το γεγονός ότι είναι λιγότερο υπολογιστικά απαιτητική για αυτά, σε σχέση με πιο συγκεντρωτικές προσεγγίσεις [8].

1.1 Αντικείμενο της διπλωματικής εργασίας

Η παρούσα εργασία αφορά τον σχεδιασμό, την υλοποίηση και την αξιολόγηση ενός συστήματος ανίχνευσης εισβολής το οποίο βασίζεται σε υβριδική αρχιτεκτονική βαθιάς μάθη-

σης, με το μοντέλο να εκπαιδεύεται συνεργατικά στο πλαίσιο cross-silo federated learning.

Ο όρος *υβριδική* αρχιτεκτονική, αναφέρεται στο γεγονός ότι το μοντέλο που επιλέγεται και υλοποιείται αποτελείται από δύο διακριτά τμήματα, με το πρώτο να εκπαιδεύεται με μη επιβλεπόμενη μάθηση και το δεύτερο με επιβλεπόμενη. Η προσέγγιση αυτή καθίσταται αναγκαία από το γεγονός ότι στην πράξη η πλειοψηφία των διαθέσιμων πληροφοριών αφορά μη επισημασμένα δεδομένα, και η μη επιβλεπόμενη μάθηση είναι σε θέση να υπερκεράσει την έλλειψη ετικετών σε αυτά. Το συγκεκριμένο μοντέλο επιθυμούμε να είναι σε θέση να ανιχνεύει αξιόπιστα εισβολές (μοντελοποίηση ως binary classification μεταξύ ομαλής και κακόβουλης δικτυακής κίνησης), αλλά και να μπορεί να κατηγοριοποιεί τις απειλές (μοντελοποίηση ως multiclass classification).

Για την αξιολόγηση της δυνατότητας αυτού του baseline μοντέλου να μεταφερθεί επιτυχώς σε περιβάλλον cross-silo federated learning, εκτελούνται για το federated μοντέλο πειράματα και υπολογισμοί μετρικών όμοια με αυτά του baseline. Έτσι, εξετάζουμε κατά πόσο μειώνεται η επίδοσή του όταν επιχειρούμε να εκμεταλλευτούμε τα πλεονεκτήματα του federated learning. Ακόμα, γίνεται διερεύνηση της συμπεριφοράς του μοντέλου που έχει εκπαιδευτεί με federated learning όταν αίρεται η i.i.d. υπόθεση για τα δεδομένα, δηλαδή όταν δεν υποθέτουμε ότι προέρχονται στατιστικά από ανεξάρτητες και ταυτόσημες κατανεμημένες τυχαίες μεταβλητές.

1.2 Οργάνωση της διπλωματικής εργασίας

Η διπλωματική εργασία είναι οργανωμένη σε δέκα κεφάλαια:

- Στο παρόν κεφάλαιο γίνεται μια σύντομη εισαγωγή στον χώρο των συστημάτων ανίχνευσης εισβολής, καθώς και στο Federated learning, με αναφορά στα πλεονεκτήματα και τις προκλήσεις που προκύπτουν από την χρήση του. Στην συνέχεια, παρουσιάζεται συνοπτικά το αντικείμενο της εργασίας.

- Το κεφάλαιο 2 είναι το πρώτο του θεωρητικού μέρους της εργασίας. Σε αυτό εισάγονται ορισμοί και το γενικότερο θεωρητικό υπόβαθρο που είναι χρήσιμο για την κατανόηση του θέματος. Πιο συγκεκριμένα, επικεντρωνόμαστε στην ανάλυση δικτυακής κίνησης βασισμένης σε flows, στον ορισμό και τα είδη συστημάτων ανίχνευσης εισβολής, και τέλος στο federated learning, περιγράφοντας τις βασικότερες αρχές που το διέπουν.

- Το κεφάλαιο 3 είναι αφιερωμένο στην αναφορά εργασιών που έχουν προηγηθεί, σχετικά με την χρήση μηχανικής μάθησης για την ανάπτυξη συστημάτων ανίχνευσης εισβολής, αναφέροντας παραδείγματα χρήσης του dataset που επιλέχθηκε για την συγκεκριμένη εργασία, και δίνοντας περισσότερο έμφαση σε υλοποιήσεις που χρησιμοποιούν federated learning.

- Στο κεφάλαιο 4 γίνεται αναλυτική περιγραφή του θέματος της διπλωματικής εργασίας, ορίζοντας αρχικά τους στόχους που θέτουμε για το προτεινόμενο σύστημα ανίχνευσης εισβολής, και στην συνέχεια περιγράφεται η αρχιτεκτονική του υβριδικού μοντέλου που επιλέγεται και της διάταξης federated learning για την συνεργατική εκπαίδευσή του.

- Το κεφάλαιο 5 αποτελεί την αρχή το πρακτικού μέρους της εργασίας. Σε αυτό, παρουσιάζεται το dataset που επιλέχθηκε, και περιγράφεται η αναγκαία προεπεξεργασία των δεδομένων του για την χρήση τους στα πειράματα που πραγματοποιήθηκαν.

- Στο κεφάλαιο 6 παρουσιάζεται λεπτομερώς η υλοποίηση του υβριδικού μοντέλου, της διάταξης federated learning καθώς και του αλγορίθμου federated averaging. Στην συνέχεια, περιγράφονται τα πειράματα που εκτελέστηκαν για την αξιολόγηση του, καθώς και οι μετρικές που χρησιμοποιήθηκαν.

- Το κεφάλαιο 7 αφιερώνεται στην αναλυτική παρουσίαση των αποτελεσμάτων για όλα τα πειράματα που εκτελέστηκαν, με παρατηρήσεις σχετικά με το αποτέλεσμα που προέκυψε για το καθένα από αυτά.

- Έχοντας παρουσιάσει τα αποτελέσματα των πειραμάτων, στο κεφάλαιο 8 συγκρίνουμε την επίδοση του συνεργατικού μοντέλου σε σχέση με το αντίστοιχο που εκπαιδεύτηκε χωρίς federated learning, ενώ γίνεται συζήτηση και αξιολόγησή του βάσει των αποτελεσμάτων σχετικά με την επίδοσή του αλλά και της συμπεριφοράς του κατά την εκπαίδευση σε διαφορετικά σενάρια, όταν παύει να ισχύει η υπόθεση i.i.d. για τα δεδομένα.

- Το κεφάλαιο 9 είναι το πρώτο εκ των δύο του επιλόγου. Σε αυτό παρουσιάζονται τα τελικά συμπεράσματα που προκύπτουν από την χρήση του υβριδικού μοντέλου για την ανίχνευση εισβολής, καθώς και για την χρήση federated learning σε αυτό το πλαίσιο, στηριζόμενοι σε παρατηρήσεις και στα αποτελέσματα των πειραμάτων.

- Τέλος, στο κεφάλαιο 10 αναφερόμαστε σε μελλοντικό ερευνητικό έργο που μπορεί να πραγματοποιηθεί στα πλαίσια βελτιστοποίησης της επίδοσης και της αποδοτικότητας της αρχιτεκτονικής που υλοποιήθηκε, καθώς και στην χρήση μεθόδων που αποσκοπούν στην αύξηση της ασφάλειας και της ιδιωτικότητας που προσφέρει.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται συνοπτικά βασικές έννοιες και τεχνικές που σχετίζονται με το αντικείμενο της εργασίας, με σκοπό την πληρέστερη κατανόησή της. Έτσι, εισάγονται έννοιες όπως η ανάλυση δικτυακής κίνησης η οποία βασίζεται σε *flows* καθώς και τα συστήματα ανίχνευσης εισβολής (*Intrusion Detection Systems, IDS*), με μια κλασσική κατηγοριοποίηση τους. Δίνεται περαιτέρω έμφαση στις αρχές που διέπουν την ομοσπονδιακή μάθηση (*Federated Learning, FL*), παρουσιάζοντας και συγκρίνοντας αδρομερώς τα είδη που απαντώνται στην βιβλιογραφία, καθώς και την περιγραφή μιας τυπικής διάταξης συστημάτων που συνεργάζονται στα πλαίσια του federated learning με σκοπό την ανάπτυξη ενός κοινού μοντέλου. Τέλος, αναφέρονται ορισμένα από τα πλεονεκτήματα του federated learning, αλλά και οι προκλήσεις που καλούνται να αντιμετωπιστούν στα πλαίσια συνεχούς βελτίωσης των εφαρμογών του.

2.1 Ανάλυση δικτυακής κίνησης βασισμένη σε flows

2.1.1 Ορισμός flow

Η έννοια του *flow*, σε τομείς που σχετίζονται με την ανάλυση της δικτυακής κίνησης, είναι ευρέως διαδεδομένη και συχνά αλλάζει σημασία ανάλογα με το συγκεκριμένο θέμα που μελετάται. Για παράδειγμα ένα flow μπορεί να οριστεί ως ¨...μια ακολουθία πακέτων από μια συγκεκριμένη πηγή προς συγκεκριμένο unicast, anycast ή multicast προορισμό, η οποία ορίζεται όπως επιθυμεί η πηγή...¨ (RFC 3697 [9]), ή ως ¨...ένα τεχνητό, λογικό ισοδύναμο μιας κλήσης ή σύνδεσης...¨ (RFC 2722 [10]). Στην περίπτωση της παρούσης ανάλυσης, ένα flow ορίζεται ως *μια ακολουθία πακέτων που έχουν ίδιες τιμές για το σύνολο των εξής 5 χαρακτηριστικών* [11].

{IP προέλευσης, IP προορισμού, θύρα προέλευσης, θύρα προορισμού, πρωτόκολλο}

Ο παραπάνω ορισμός είναι σύμφωνος με αυτόν που χρησιμοποιούν οι παραγωγοί του dataset (βλ. ενότητα 5.1) που αξιοποιούμε για την εκπαίδευση και την ανάλυση της επίδοσης του υβριδικού μοντέλου στην παρούσα εργασία.

2.1.2 Δεδομένα για συστήματα ανίχνευσης εισβολής

Τα δεδομένα που χρησιμοποιούνται για την ανάπτυξη συστημάτων εισβολών, τα οποία βασίζονται στην δικτυακή κίνηση, μπορούν αδρομερώς να χωριστούν στις εξής τρεις κατηγορίες [12]:

- Δεδομένα βασισμένα στα πεδία των πακέτων της δικτυακής κίνησης. Αυτά τα datasets περιέχουν αναλυτικά την δικτυακή κίνηση, η οποία μπορεί να εξάγεται σε μορφή pcap αρχείων και περιέχουν τα δεδομένα κάθε πακέτου. Τα διαθέσιμα μεταδεδομένα και οι επικεφαλίδες των πακέτων εξαρτώνται από το δίκτυο καταγραφής και το πρωτόκολλο που χρησιμοποιείται (π.χ. TCP, UDP, ICMP). Συχνά, πέρα από την επικεφαλίδα του πρωτοκόλλου μεταφοράς, υπάρχει και επικεφαλίδα IP, με πληροφορίες όπως οι διευθύνσεις πηγής και προορισμού.

- Δεδομένα βασισμένα σε flows. Αυτή η μορφή dataset περιέχει κυρίως συμπυκνωμένες μεταπληροφορίες, που αφορούν συγκεντρωτικά την ανταλλαγή πακέτων μέσα στο δίκτυο, συχνά χωρίς να συμπεριλαμβάνουν τα δεδομένα του. Ο συνήθης ορισμός του flow σε αυτά τα datasets είναι αυτός που χρησιμοποιείται και στην παρούσα εργασία. Αυτά τα flows μπορεί να είναι μονόδρομα (unidirectional flows) ή αμφίδρομα (bidirectional flows), δηλαδή να λαμβάνουν υπόψιν συνολικά την επικοινωνία μεταξύ δύο συστημάτων ή χρηστών μέσα στο δίκτυο. Τυπικά πρωτόκολλα για την εξαγωγή flows είναι τα NetFlow [13], IPFIX [14], sFlow [15] και OpenFlow [16], ενώ υπάρχει πληθώρα διαθέσιμων εργαλείων, όπως το nfdump [17] και το YAF [18], τα οποία μπορούν να μετατρέψουν δεδομένα που βασίζονται σε πακέτα, σε αντίστοιχα βασισμένα σε flows. Ένα από αυτά τα εργαλεία είναι και το CICFlowMeter-V3 το οποίο χρησιμοποιήθηκε για την παραγωγή χαρακτηριστικών που βασίζονται σε bidirectional flows (βλ. κεφάλαιο 5)

- Δεδομένα στα οποία η παραπάνω διάκριση δεν είναι αποκλειστική. Ένα παράδειγμα dataset αυτής της κατηγορίας μπορεί να περιέχει δεδομένα που βασίζονται σε flows αλλά να έχει εμπλουτιστεί και με δεδομένα από τα ίδια τα πακέτα ή από αρχεία καταγραφής των συστημάτων στο δίκτυο. Αντιπροσωπευτικό είναι το dataset KDD CUP 1999 [19], το οποίο πέρα από αθροιστικά χαρακτηριστικά (όπως το σύνολο των bytes που έστειλε ο κόμβος προέλευσης ή τα TCP flags) περιέχει και τον αριθμό αποτυχημένων προσπαθειών εισόδου. Η γενική δομή αυτών των dataset διαφέρει και εξαρτάται από την ανάλυση που πραγματοποιείται.

2.2 Συστήματα ανίχνευσης εισβολής

2.2.1 Ορισμός

Ανίχνευση εισβολής ονομάζεται η διαδικασία παρακολούθησης γεγονότων ενός υπολογιστικού συστήματος ή δικτύου, και η ανάλυσή τους για σημάδια πιθανών συμβάντων (incidents), δηλαδή παραβιάσεων ή άμεσων απειλών παραβίασης των πολιτικών ασφάλειας υπολογιστών, των αποδεκτών όρων χρήσης ή των καθιερωμένων πρακτικών ασφαλείας. Τα

συμβάντα αυτά μπορεί να προέρχονται από κακόβουλο λογισμικό, χρήστες με μη εξουσιοδοτημένη πρόσβαση στα συστήματα, ή εξουσιοδοτημένους χρήστες που καταχρώνται τα δικαιώματα τους ή επιχειρούν να αποκτήσουν περαιτέρω δικαιώματα που δεν τους αναλογούν. Αν και πολλά από αυτά τα συμβάντα είναι κακόβουλης φύσεως, κάποια μπορεί να οφείλονται σε ακούσιο λάθος [4].

Σύστημα ανίχνευσης εισβολής (intrusion detection system, IDS) ονομάζεται το λογισμικό που αυτοματοποιεί την διαδικασία ανίχνευσης εισβολής, μειώνοντας συχνά έτσι και την ανάγκη ανθρώπινης παρέμβασης. Παρόμοια, *σύστημα πρόληψης εισβολής (intrusion prevention system, IPS)*, καλείται το λογισμικό που διαθέτει τις λειτουργίες ενός IDS και επιπλέον έχει την δυνατότητα να προλάβει πιθανά συμβάντα. Συχνά, για λόγους συντομίας, χρησιμοποιείται ο όρος *σύστημα ανίχνευσης και πρόληψης εισβολής (intrusion detection and prevention systems, IDPS)* για να περιγράψει τόσο IDS όσο και IPS τεχνολογίες [4]. Στην παρούσα εργασία, θα χρησιμοποιούμε στο εξής για λόγους λιτότητας, μόνο τον όρο *σύστημα ανίχνευσης εισβολής* (ή IDS), έχοντας όμως υπόψιν ότι πολλά τέτοια συστήματα στοχεύουν και στην λήψη ενεργειών για πρόληψη.

2.2.2 Είδη συστημάτων ανίχνευσης εισβολής

Καθώς η χρήση των συστημάτων ανίχνευσης εισβολής έχει υιοθετηθεί ευρύτατα, το ενδιαφέρον σχετικά με την αρχιτεκτονική, τον τρόπο λειτουργίας και την επίδοση τους έχει παράξει πληθώρα διαφορετικών τέτοιων συστημάτων. Για την ταξινόμηση τους, μπορούν να χρησιμοποιηθούν χαρακτηριστικά που διαφοροποιούν αυτά τα συστήματα μεταξύ τους, παρά το γεγονός ότι πολλές τεχνολογίες IDS χρησιμοποιούν πολλαπλές μεθοδολογίες [4]. Για παράδειγμα, τα IDS μπορούν να διαχωριστούν με βάση τα παρακάτω κριτήρια [5]:

- Το είδος του συστήματος που καλούνται να προστατέψουν και την τοποθεσία των δεδομένων που χρησιμοποιούν για να γίνει ο έλεγχος. Αρχικά τα IDS ήταν *host-based*, με σκοπό την προστασία ενός κεντρικού mainframe, στον οποίο είχαν πρόσβαση πολλαπλοί τοπικοί χρήστες. Σε τέτοιου είδους περιβάλλον, τα δεδομένα μπορεί να προέρχονται από audit trails ή αρχεία καταγραφής συστήματος (system logs). Η ανάπτυξη των δικτύων και η δημιουργία πιο σύνθετων κατανεμημένων συστημάτων οδήγησαν στην χρήση *network-based* IDS τα οποία είναι σε θέση να προστατέψουν τα επιμέρους υποσυστήματα που επικοινωνούν. Η λειτουργία των network-based IDS βασίζεται κυρίως στην δικτυακή κίνηση για να ανιχνεύσουν πιθανές κακόβουλες ενέργειες, για παράδειγμα εξετάζοντας τα πακέτα που μεταφέρονται.

- Χρόνος και συχνότητα λειτουργίας. Κάποια συστήματα ανιχνεύουν σε πραγματικό χρόνο (real-time intrusion detection) βάσει γεγονότων και πληροφοριών από το περιβάλλον στο οποίο βρίσκονται. Το πλεονέκτημα όμως της συνεχούς αυτής προστασίας, συνοδεύεται συχνά από επιβάρυνση του συστήματος λόγω της ανάλυσης των δεδομένων, που απαιτεί δέσμευση υπολογιστικών πόρων. Αντίθετα, πιο στατικά IDS εκτελούν περιοδικά προγραμματισμένες ανιχνεύσεις, οι οποίες αναλύουν την τρέχουσα κατάσταση του περιβάλλοντος.

- Συμπεριφορά μετά την ανίχνευση. Περιγράφει την αντίδραση του συστήματος όταν

ανιχνευτεί εισβολή. Το IDS χαρακτηρίζεται ως *ενεργό (active)* όταν εφαρμόζει μέτρα για την αντιμετώπιση της εισβολής, τα οποία μπορεί να είναι είτε διορθωτικά (π.χ. διόρθωση ευπαθούς σημείου του συστήματος που προστατεύει), είτε προληπτικά (π.χ. αποσύνδεση χρηστών που πιθανά ευθύνονται για την εισβολή, ή τερματισμός υπηρεσιών). Αντίθετα, ένα IDS το οποίο σημαίνει συναγερμό όταν εντοπίσει πιθανή εισβολή, χωρίς περαιτέρω σύνθετες ενέργειες, ονομάζεται *παθητικό (passive)*.

- Μέθοδος ανίχνευσης, η οποία αναλύεται εκτενέστερα παρακάτω.

Για την ταξινόμηση των IDS με βάση την μέθοδο ανίχνευσης εισβολών, έχουν προταθεί πολλές κατηγοριοποιήσεις με μεταβλητό βαθμό λεπτομέρειας (βλ. [1, 5, 20, 21, 22, 23]). Παρακάτω φαίνεται ο βασικός διαχωρισμός των IDS βάσει των μεθόδων ανίχνευσης, όπως παρουσιάζεται από το National Institute of Standards and Technology (NIST) [4] και το [1]. Παρόλαυτα, επισημαίνεται πως συχνά χρησιμοποιούνται συστήματα που συνδυάζουν παραπάνω από μία μεθόδους, με σκοπό την βελτίωση της ικανότητας ανίχνευσης εισβολής:

Βασισμένα σε υπογραφές (Signature-based)

Υπογραφή (signature) ονομάζεται ένα μοτίβο / πρότυπο το οποίο αντιστοιχεί σε κάποια γνωστή απειλή για το σύστημα. Ένα παράδειγμα υπογραφής θα μπορούσε να ορίζεται για πακέτα δικτυακής κίνησης που προέρχονται από έναν άγνωστο εξωτερικό αποστολέα με προορισμό μια θύρα ενός υπολογιστή η οποία έχει δεσμευτεί αυστηρά για εσωτερική επικοινωνία μεταξύ των υπολογιστών του τοπικού δικτύου. Ένα τέτοιο συμβάν θα χαρακτηριζόταν αυτόματα ύποπτο από το IDS, καθώς η υπογραφή του (εξωτερική προέλευση και δεσμευμένη για εσωτερική χρήση θύρα ως προορισμός) θα αποτελούσε δυνητικά απειλή για το σύστημα. Αυτού του είδους τα συστήματα ονομάζονται και *βασισμένα σε γνώση (knowledge-based)* [24]. Η συγκεκριμένη μέθοδος αναγνώρισης είναι και η πιο απλή, καθώς βασίζεται στην σύγκριση μιας ενέργειας που αντιλαμβάνεται με προϋπάρχουσες υπογραφές, και είναι πολύ αποδοτική στο να ανιχνεύει γνωστές απειλές. Παρόλα αυτά η αποδοτικότητα της μειώνεται δραστικά όταν κληθεί να αντιμετωπίσει άγνωστες (ή παραλλαγμένες) απειλές, ή αν αυτές χρησιμοποιούν τεχνικές αποφυγής ανίχνευσης.

Βασισμένα σε ανίχνευση ανωμαλιών (Anomaly-based)

Αυτή η μέθοδος βασίζεται στην αναγνώριση ενεργειών που παρουσιάζουν σημαντική απόκλιση από αυτές που θεωρούνται φυσιολογικές κατά την διάρκεια λειτουργίας ενός συστήματος. Ενέργειες με τέτοιου είδους αποκλίνουσα συμπεριφορά θεωρούνται ως *ανωμαλίες (anomalies)* και μπορούν να ερμηνευτούν ως εισβολές. Για την μοντελοποίηση των φυσιολογικών ενεργειών μέσα στο σύστημα που προστατεύεται, είναι διαδεδομένη η χρήση της έννοιας του *προφίλ (profile)* η οποία αντιπροσωπεύει την συμπεριφορά φυσικών χρηστών, υπολογιστών, συνδέσεων ή υπηρεσιών στο σύστημα [4, 21]. Στην βιβλιογραφία, αυτή η κατηγορία συστημάτων ανίχνευσης εισβολής που βασίζονται σε εύρεση ανωμαλιών απαντάται και ως *βασισμένη σε συμπεριφορά (behaviour-based)* [5]. Για την λειτουργία των anomaly-based IDS, απαιτείται αρχικά μια φάση εκπαίδευσης (με την ευρύτερη έννοια του όρου, όχι υποχρεωτικά στα πλαίσια μηχανικής μάθησης), κατά την οποία το IDS αναπτύσσει τα προφίλ

και μαθαίνει να αναγνωρίζει την φυσιολογική συμπεριφορά εντός του συστήματος [4, 21]. Ανάλογα με τις διαδικασίες που απαιτούνται για την μοντελοποίηση της συμπεριφοράς, τα συστήματα αυτά μπορούν να διαχωριστούν περαιτέρω, με ένα παράδειγμα κατηγορίας να είναι αυτά που χρησιμοποιούν μηχανική μάθηση [20].

Ένα σημαντικό πλεονέκτημα των anomaly-based IDS, είναι ότι έχουν την δυνατότητα να ανιχνεύουν και νέου είδους επιθέσεις, που δεν είχαν παρουσιαστεί κατά την εκπαίδευση, καθώς δεν βασίζονται στο να υπάρχει ήδη κάποια υπογραφή στο σύστημα η οποία να τις περιγράφει. Ακόμα, επειδή βασίζονται σε προφίλ, είναι ικανά να εντοπίζουν και απειλές εντός του συστήματος που προστατεύουν, όταν παρατηρηθεί απόκλιση από την αναμενόμενη συμπεριφορά κάποιου χρήστη ή εφαρμογής, ενώ και για τον επιτιθέμενο είναι συχνά δύσκολο να γνωρίζει εκ των προτέρων ποιες ενέργειες θεωρούνται φυσιολογικές, ειδικά αν δεν έχει πρότερη γνώση της λειτουργίας του συστήματος. Ασφαλώς, τα anomaly-based IDS έχουν και μειονεκτήματα. Παραδείγματος χάρη, πολλές φορές παρουσιάζουν ψευδώς θετικά αποτελέσματα, καθώς μπορεί να χαρακτηρίσουν ως κακόβουλες και κάποιες φυσιολογικές συμπεριφορές που αποκλίνουν αισθητά από τα προφίλ στα οποία βασίζονται, ειδικά σε ένα σύνθετο ή έντονα δυναμικό περιβάλλον λειτουργίας, οι συμμετέχοντες του οποίου παρουσιάζουν ετερογένεια ή είναι παροδικοί. Για αυτόν τον λόγο, τα προφίλ αυτά μπορεί να είναι δυναμικά, και να ανανεώνονται περιοδικά ή κατά βούληση. Παρόμοιο πρόβλημα προκύπτει όταν κακόβουλη συμπεριφορά συμπεριληφθεί λανθασμένα κατά την εκπαίδευση του συστήματος ανίχνευσης εισβολής ως μέρος ενός φυσιολογικού προφίλ, με αποτέλεσμα να αγνοηθεί όταν προκύψει. Τέλος, μπορεί να είναι δύσκολο για έναν αναλυτή να καταλάβει τον λόγο για τον οποίο το σύστημα ανίχνευσης χαρακτήρισε μια συγκεκριμένη συμπεριφορά ως ανωμαλία, δηλαδή δεν εγγυώνται επεξηγηματικότητα [4, 21, 20].

Ανάλυσης πρωτοκόλλου με διατήρηση κατάστασης (Stateful Protocol Analysis, SPA)

Η συγκεκριμένη μέθοδος αναφέρεται στην διαδικασία σύγκρισης προκαθορισμένων προφίλ συμπεριφορών που θεωρούνται φυσιολογικές, με εκείνες που παρατηρούνται στο σύστημα, σκοπεύοντας στην ανίχνευση αποκλίσεων από αυτές. Σε αντίθεση όμως με τις μεθόδους ανίχνευσης ανωμαλιών που παρουσιάστηκαν παραπάνω, αυτή η μέθοδος δεν μοντελοποιεί την συμπεριφορά στο δίκτυο ή σε μηχανήματα στο σύστημα, αλλά βασίζεται σε προφίλ που έχουν αναπτυχθεί από τους σχεδιαστές των συστημάτων ανίχνευσης και περιγράφουν πώς ακολουθείται σωστά ένα πρωτόκολλο επικοινωνίας ή λειτουργίας, διατηρώντας μια έννοια κατάστασης (state) στην οποία βρίσκεται κάθε φορά η λειτουργία που παρακολουθείται. Για παράδειγμα, ένα τέτοιο IDS μπορεί να είναι σε θέση να ταιριάζει responses με τα αντίστοιχα requests που τα προκάλεσαν. Η διατήρηση κατάστασης επιτρέπει σε αυτά τα συστήματα ανίχνευσης εισβολής να αποθηκεύουν περαιτέρω πληροφορίες σχετικά με ύποπτη συμπεριφορά, που βοηθούν κατά την διερεύνηση συμβάντων σε δεύτερο χρόνο. Η ανάλυση πρωτοκόλλου που πραγματοποιεί βασίζεται σε προδιαγραφές κατασκευαστών λογισμικού, είτε οργανισμών όπως το Internet Engineering Task Force, (IETF). Κάποιοι κατασκευαστές τέτοιων συστημάτων IDS χρησιμοποιούν τον όρο *deep packet inspection* για συστήματα με συγγενή λειτουργικότητα, αν και ο συγκεκριμένος όρος πιο ορθά αναφέρεται στην παρακολούθηση δικτυακής κίνησης εξετάζοντας και το payload των πακέτων που ανταλλάσσονται. Η

ανάλυση πρωτοκόλλου με διατήρηση κατάστασης είναι γνωστή και ως *ανίχνευση βασισμένη σε προδιαγραφή (specification-based detection)* [1, 4].

Παρά τα σημαντικά πλεονεκτήματα που προσφέρει η ανάλυση πρωτοκόλλων με επίγνωση κατάστασης για το σύστημα, ένα τέτοιο IDS έχει και αρκετά μειονεκτήματα. Το βασικότερο είναι ότι η αυτή η ανάλυση για την διατήρηση κατάστασης τα κάνει πιο απαιτητικά σε πόρους, ειδικά όταν παρακολουθούν πολλαπλά sessions. Ακόμα, δεν μπορούν να ανιχνεύσουν επιθέσεις που, παρά το γεγονός ότι ακολουθούν μια συμπεριφορά συμβατή με το εκάστοτε πρωτόκολλο, ο όγκος των ενεργειών στα πλαίσια αυτής προκαλεί άρνηση υπηρεσιών. Τέλος, ενώ βασίζονται στον ορισμό του εκάστοτε πρωτοκόλλου, η μοντελοποίηση του στα πλαίσια του IDS μπορεί να απέχει από την υλοποίηση του πρωτοκόλλου μεταξύ διαφορετικών εκδόσεων εφαρμογών, λειτουργικών συστημάτων, ή υπηρεσιών με τις οποίες το σύστημα επικοινωνεί [4].

2.3 Ομοσπονδιακή Μάθηση (Federated Learning)

Ομοσπουδιακή μάθηση (Federated Learning, FL) ονομάζεται μια τεχνική μηχανικής μάθησης στην οποία πολλαπλοί *clients* συνεργάζονται με σκοπό την εκπαίδευση ενός κοινού μοντέλου, υπό τον συντονισμό ενός κεντρικού *server*. Τα δεδομένα τα οποία χρησιμοποιούνται για την εκπαίδευση των clients παραμένουν σε αυτούς, και δεν απαιτείται διαμοιρασμός τους, ούτε μεταξύ των clients, ούτε με τον server [25]. Σαν όρος χρησιμοποιήθηκε για πρώτη φορά στο [26] για την εκπαίδευση βαθιών νευρωνικών δικτύων.

2.3.1 Περιγραφή συστήματος Federated Learning

Στην κλασσική του μορφή, ένα σύστημα federated learning, αποτελείται από πολλαπλούς (πιθανώς ετερογενείς) clients και έναν server. Ενώ απαιτείται δυνατότητα (έστω και παροδικής) αμφίδρομης επικοινωνίας του κάθε client με τον server για την αποστολή βαρών και άλλων παραμέτρων του μοντέλου, οι clients μεταξύ τους δεν έχουν κάποια επαφή. Τα δεδομένα εκπαίδευσης του κάθε client βρίσκονται μόνο σε αυτόν και σε καμία στιγμή, καθ' όλη την διάρκεια της μάθησης, δεν αποστέλλονται στο δίκτυο.

Παρακάτω παρουσιάζεται μια τυπική διαδικασία μάθησης με χρήση federated learning με τον κεντρικό server να δρα ως συντονιστής της εκπαίδευσης. Το μοντέλο μάθησης παραμετροποιείται από βάρη, όπως στα βαθιά νευρωνικά δίκτυα. Η διαδικασία επαναλαμβάνεται, έως ότου διακοπεί από τον επιβλέποντα της εκπαίδευσης [25], με την κάθε επανάληψη να ονομάζεται *γύρος (round)* (παρουσιάζεται γραφικά στο σχήμα 2.1) και να περιέχει τα εξής βήματα:

1. **Επιλογή των client (Client selection)**: Ο server επιλέγει ένα υποσύνολο των client για τον συγκεκριμένο γύρο. Η επιλογή αυτή μπορεί να γίνεται βάσει ορισμένης πολιτικής, ή συγκεκριμένων προϋποθέσεων που πρέπει να πληρούν οι clients που θα συμμετάσχουν στον γύρο, όπως η διαθεσιμότητά τους. Για παράδειγμα, αν οι clients είναι κινητά τηλέφωνα, τότε θα επιλεχθούν εκείνα που είναι αδρανή, που βρίσκονται σε φόρτιση κ.τ.λπ. Σε ορισμένες περιπτώσεις, η επιλογή μπορεί να γίνεται και τυχαία στηριζόμενη σε κάποια συγκεκριμένη κατανομή.

2. **Μετάδοση (Broadcast)**. Τα τρέχοντα βάρη του κοινού μοντέλου μεταδίδονται σε όλους τους clients που έχουν επιλεχθεί να συμμετάσχουν στον γύρο. Η αρχικοποίηση των βαρών που αποστέλλονται στους clients κατά τον πρώτο γύρο εκπαίδευσης είναι συνήθως ευθύνη του κεντρικού server. Μαζί με τα βάρη, μεταδίδεται και ένα πλάνο εκπαίδευσης που θα ακολουθήσουν οι clients.

3. **Εκπαίδευση τοπικών μοντέλων στους clients**. Σε αυτό το βήμα, ο κάθε client εκπαιδεύει ένα τοπικό αντίγραφο του συνολικού μοντέλου, όπως το δέχτηκε από τον server στο προηγούμενο βήμα. Για την εκπαίδευση του τοπικού του μοντέλου, ο εκάστοτε client χρησιμοποιεί μόνο τα δεδομένα που έχει τοπικά διαθέσιμα. Στο τέλος αυτού του βήματος, κάθε client έχει πλέον στο τοπικό του μοντέλο ανανεωμένα βάρη (ή υπολογισμένες ανανεώσεις βαρών, weight updates) τα οποία έχουν προκύψει βάσει του αλγορίθμου βελτιστοποίησης που έχει επιλεχθεί.

4. **Συλλογή βαρών**. Ο server συλλέγει τα βάρη (ή τις ανανεώσεις βαρών) από όλους τους clients που συμμετείχαν στον γύρο, ή από ένα υποσύνολο του, το μέγεθος του οποίου κρίνεται ικανοποιητικό.

5. **Συμψηφισμός (Aggregation)**. Ο server συμψηφίζει τα βάρη βάσει συγκεκριμένου αλγορίθμου. Τα βήματα συλλογής και συμψηφισμού των ανανεώσεων βαρών επιδέχονται πολλές παραλλαγές και βελτιστοποιήσεις στην πράξη, καθώς μπορούν να εφαρμοστούν διαφορετικοί μέθοδοι με σκοπό, μεταξύ άλλων, την βελτίωση της ασφάλειας και της ιδιωτικότητας των clients, την επιτάχυνση της εκπαίδευσης, ή την μείωση των δικτυακών απαιτήσεων.

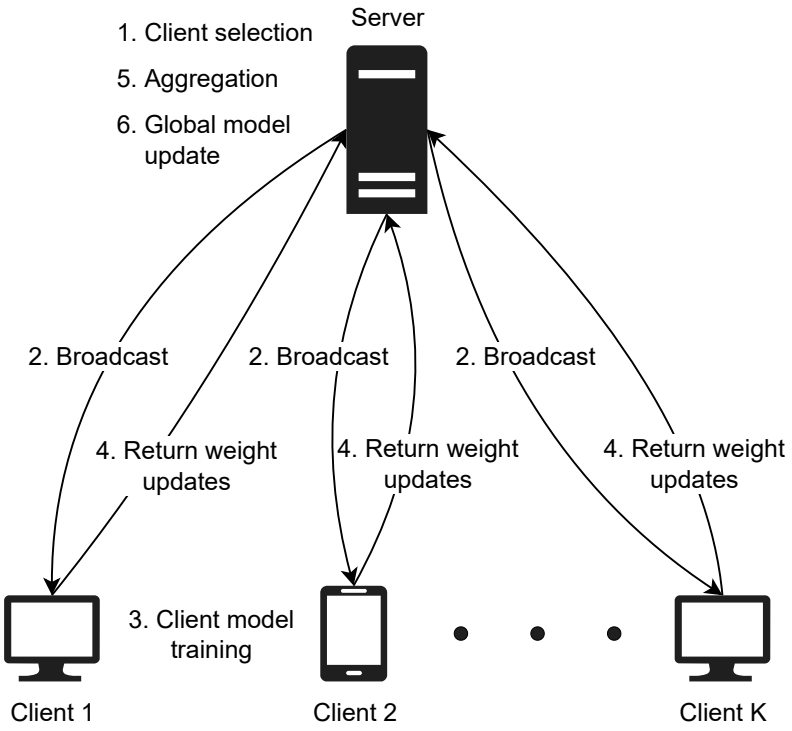
6. **Ανανέωση του μοντέλου (Model update)**. Με βάση το αποτέλεσμα του συμψηφισμού που προέκυψε στο προηγούμενο βήμα, ο server ανανεώνει το κοινό συνεργατικό μοντέλο.

2.3.2 Ο αλγόριθμος Federated Averaging

Η βελτιστοποίηση του κοινού μοντέλου, όπως ορίζεται στα πλαίσια του federated learning, ονομάζεται *ομοσπουδιακή βελτιστοποίηση (federated optimization)*. Αν και συνδέεται με το τυπικό πρόβλημα κατανεμημένης βελτιστοποίησης (distributed optimization), δηλαδή την ελαχιστοποίηση μιας κοινής αντικειμενικής συνάρτησης που είναι το άθροισμα των επιμέρους αντικειμενικών συναρτήσεων σε ένα σύνολο συνεργαζόμενων κόμβων [27], έχει ορισμένες διαφορές [26]:

- Δεν προϋποθέτει υποχρεωτικά ανεξάρτητα και ταυτόσημα κατανεμημένα (independent and identically distributed, i.i.d.) δεδομένα. Λόγω της φύσης του προβλήματος, τα τοπικά δεδομένα κάποιου client μπορεί να μην είναι αντιπροσωπευτικά της κατανομής των συνολικών δεδομένων.

- Δεν προϋποθέτει ισορροπημένα, ως προς το πλήθος τους, δεδομένα. Έτσι, ορισμένοι clients μπορεί να διαθέτουν περισσότερα ή λιγότερα δεδομένα από τους υπόλοιπους.



Σχήμα 2.1: *Μια τυπική διαδικασία federated learning*

- Το πλήθος των clients που συμμετέχουν στην βελτιστοποίηση δύναται να είναι μεγαλύτερο από το μέσο πλήθος δειγμάτων του εκάστοτε client.
- Μπορεί να υπάρχει περιορισμένη δυνατότητα επικοινωνίας μεταξύ των clients και του server (π.χ. επειδή κάποιοι clients είναι μη διαθέσιμοι ή λόγω υψηλού κόστους σύνδεσης).

Στο federated learning, ο κλασσικός επαναληπτικός αλγόριθμος βελτιστοποίησης που χρησιμοποιείται για την ελαχιστοποίηση της συνάρτησης απώλειας (loss function) ονομάζεται *Federated Averaging (FedAvg)* [26], και βασίζεται στον αλγόριθμο στοχαστικής κατάβασης κλίσης (Stochastic Gradient Descent, SGD). Για την περιγραφή του αλγορίθμου, υιοθετείται ο παρακάτω συμβολισμός:

- K: Το συνολικό πλήθος των clients.
- C: Το κλάσμα των clients που επιλέγονται για να συμμετάσχουν σε κάθε γύρο.
- E: Το πλήθος των εποχών τοπικής εκπαίδευσης σε κάθε client, ανά γύρο.
- B: Το μέγεθος των minibatch των τοπικών δεδομένων του κάθε client.
- n: Ο συνολικός αριθμός δειγματικών στοιχείων (σε όλους τους client).
- n_k: Ο αριθμός δειγματικών στοιχείων στον k-οστό client, με P_k τους δείκτες τους.
- l: Η συνάρτηση απώλειας που χρησιμοποιεί ο κάθε client.

- η: Το learning rate που χρησιμοποιούν οι clients για την ανανέωση των βαρών στα πλαίσια του SGD που εκτελούν.
- w_t: Τα βάρη του κοινού συνεργατικού μοντέλου στον γύρο t.
- w_t^k: Τα βάρη του τοπικού μοντέλου που επέστρεψε ο client k στον server στον γύρο t.

Με βάση τα παραπάνω, παρουσιάζεται στο αλγόριθμο 2.1 σε μορφή ψευδοκώδικα ο αλγόριθμος Federated Averaging, όπως ορίστηκε στην αρχική του μορφή [26]. Ισοδύναμα, οι clients μπορούν να αποστέλλουν τις ανανεώσεις, ή διαφορές των βαρών, όπως προκύπτουν στο τοπικό τους μοντέλο και στην συνέχεια ο server να τις συνοψίζει στο κοινό μοντέλο, ακόμα και με χρήση δικού του learning rate, όχι υποχρεωτικά ίδιο με των clients [28]. Η ειδική περίπτωση κατά την οποία ο κάθε client χρησιμοποιεί αυτούσιο όλο το τοπικό του dataset (δηλαδή ως ένα minibatch), μία μόνο φορά πριν ανανεώσει τα βάρη του τοπικού μοντέλου και τα στείλει στον server (δηλαδή $E = 1$) ονομάζεται Federated SGD, (FedSGD) [26].

Αλγοριθμος 2.1: *Ο αλγόριθμος Federated Averaging*

Server executes:
  Initialize w_0
  for each round $t = 1, 2, \ldots$ **do**
    $m \leftarrow \max(C \times K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **do in parallel**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    end for
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$
  end for

ClientUpdate(k, w)**:** ▷ Runs on client k
  $\mathcal{B} \leftarrow$ Split P_k in minibatches of size B
  for each local epoch i from 1 to E **do**
    for each batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \times \nabla l(w; b)$
    end for
  end for
  Return w to server

2.3.3 Κατηγοριοποίηση συστημάτων Federated Learning

Στην βιβλιογραφία υπάρχουν πολλαπλοί τρόποι κατηγοριοποίησης των διαφορετικών συστημάτων federated learning, ο καθένας με σκοπό την ανάδειξη συγκεκριμένων χαρακτηριστικών διαφορών. Ένας διαδεδομένος τρόπος κατηγοριοποίησης αυτών των συστημάτων, βασίζεται στον τρόπο κατανομής των δεδομένων και των χαρακτηριστικών τους στους διαφορετικούς clients που συμμετέχουν, με διαφορετικές έννοιες ασφάλειας για την κάθε κατηγορία [29, 30]:

- **Οριζόντιο (Horizontal) Federated Learning**. Στην κατηγορία αυτή εμπίπτουν τα συστήματα στα οποία τα δεδομένα που βρίσκονται στον κάθε client, ενώ μπορεί να

προέρχονται από διαφορετικές πηγές, ανήκουν στο ίδιο feature space, δηλαδή περιγράφονται από το ίδιο σύνολο χαρακτηριστικών και ετικετών. Στο τέλος της εκπαίδευσης ο server μοιράζεται το μοντέλο με τους clients οι οποίοι μπορούν να το χρησιμοποιήσουν ατομικά. Για παράδειγμα, διαφορετικά νοσοκομεία τα οποία θέλουν να μοντελοποιήσουν την πορεία της νόσησης των ασθενών τους κατά την διάρκεια περίθαλψής τους, μπορούν να χρησιμοποιήσουν κοινά χαρακτηριστικά και μετρικές για να την περιγράψουν, αν και το σύνολο των κοινών ασθενών (άρα και δεδομένων) μεταξύ τους να είναι μικρό. Σχετικά με την έννοια της ασφάλειας, για την σωστή λειτουργία τους, τα συστήματα αυτά τυπικά υποθέτουν ότι οι συμμετέχοντες είναι ειλικρινείς (honest) και δεν πράττουν κακόβουλα, ενώ θεωρητικά μόνο ο server δύναται να είναι honest-but-curious, δηλαδή ενώ είναι έγκυρα συμμετέχων και συμπεριφέρεται εντός πρωτοκόλλου, θα προσπαθήσει να μάθει πληροφορίες μέσω των μηνυμάτων που φτάνουν σε αυτόν και να θέσει σε κίνδυνο την ιδιωτικότητά των clients [29]. Ως υποσημείωση, αναφέρεται ότι σε αυτήν την κατηγορία εμπίπτει και το σύστημα που υλοποιείται στην παρούσα εργασία, καθώς τα δεδομένα που βρίσκονται στον κάθε client για την προσομοίωση προέρχονται από το ίδιο dataset.

- **Κατακόρυφο (Vertical) Federated Learning**. Αυτή η κατηγορία χρησιμοποιείται σε περιπτώσεις που τα δεδομένα είναι διαχωρισμένα κατακόρυφα, δηλαδή ενώ αφορούν κυρίως τις ίδιες πηγές (π.χ. χρήστες ή παρατηρήσεις συμβάντων), τα δείγματα του κάθε client διαθέτουν διαφορετικά features ή και ετικέτες. Για παράδειγμα, ένα νοσοκομείο και ένα τοπικό φαρμακείο που μοιράζονται συχνά κοινούς ασθενείς και πελάτες, αλλά τους περιγράφουν με διαφορετικά χαρακτηριστικά λόγω της φύσης των δεδομένων που συλλέγουν, μπορούν να συνεργαστούν ώστε να αναπτύξουν ένα μοντέλο πρόβλεψης των πιο συχνών παθήσεων. Τέτοια μοντέλα επιχειρούν να συμψηφίζουν τα διαφορετικά χαρακτηριστικά κατά την μάθηση, διατηρώντας την ιδιωτικότητα των συμμετεχόντων και των δεδομένων τους. Όσον αφορά τον ορισμό ασφάλειας, τυπικά υποθέτουν honest-but-curious συμμετέχοντες. Ένας κακόβουλος παράγοντας μπορεί να μάθει πράγματα μόνο από τον client στον οποίο έχει αποκτήσει πρόσβαση. Μετά την ολοκλήρωση της εκπαίδευσης, ο κάθε client διαθέτει μόνο τις παραμέτρους του μοντέλου που αφορούν τα features των δεδομένων του, συνεπώς κατά το inference χρειάζεται πάλι να συνεργαστούν [29].

- **Ομοσπονδιακή Μεταφοράς Μάθησης (Federated Transfer Learning, FTL)**. Στο federated transfer learning [31] εμπίπτουν περιπτώσεις ομοσπονδιακής μάθησης που αφορούν clients, τα δεδομένα των οποίων εν γένει διαφέρουν τόσο σε χαρακτηριστικά (ή και ετικέτες), όσο και στις οντότητες από τις οποίες προέρχονται. Για παράδειγμα, δύο οργανισμοί που μπορεί να μοιράζονται μόνο λίγους κοινούς χρήστες ή χαρακτηριστικά που τους περιγράφουν, μπορούν να χρησιμοποιήσουν τεχνικές που προέρχονται από το transfer learning [32], ώστε να συνεργαστούν χρησιμοποιώντας το σύνολο των δεδομένων τους και όλο το feature space αυτών. Πιο συγκεκριμένα, κατά την εκπαίδευση μαθαίνεται μια κοινή αναπαράσταση του feature space χρησιμοποιώντας περιορισμένα κοινά στοιχεία και στην συνέχεια αυτή μεταφέρεται για να εφαρμοστεί στα δεδομένα του εκάστοτε client. Συνήθως το federated transfer learning υλοποιείται

μεταξύ δύο συμμετεχόντων και χρησιμοποιεί παρεμφερή πρωτόκολλα και διαδικασίες με το vertical federated learning, με τον ορισμό ασφάλειάς του να αποτελεί επέκταση του αντίστοιχου στο vertical federated learning [29].

Ένας δεύτερος σημαντικός τρόπος κατηγοριοποίησης είναι το μέγεθος της ομοσπονδίας και το είδος των clients που συμμετέχουν σε αυτή. Έτσι, γίνεται διαχωρισμός μεταξύ *cross-device* και *cross-silo* federated learning, με τις δύο μεθόδους να εμφανίζουν αρκετές διαφορές. Στον πίνακα 2.1 παρουσιάζονται συνοπτικά οι βασικότερες διαφορές και ομοιότητες για τις τυπικές μορφές τους [25, 33, 34].

Ιδιότητα	**cross-device**	**cross-silo**
Είδος client	Κινητά τηλέφωνα, συσκευές IoT	Οργανισμοί, data centers
Πλήθος clients	Μεγάλο πλήθος, π.χ. 10^{10}	~ 2 - 100
Διαθεσιμότητα clients	Ένα κλάσμα των client είναι διαθέσιμο κάθε φορά	Πάντα ή σχεδόν πάντα διαθέσιμοι clients
Αναγνώριση clients	Δεν γίνεται χρήση αναγνωριστικών	Μοναδική, σταθερή ταυτότητα για κάθε client
Δυνατότητα stateful clients	Όχι	Ναι
Ενορχήστρωση διαδικασίας	Από server	Από server
Υπολογιστική ισχύς	Συνήθως περιορισμένη	Συνήθως μεγάλη
Πλήθος δεδομένων	Συνήθως σχετικά μικρότερο	Μπορεί να είναι πολύ μεγάλο
Κατανομή δεδομένων	Τοπικά στον κάθε agent Όχι υποχρεωτικά i.i.d.	Τοπικά στον κάθε agent Όχι υποχρεωτικά i.i.d.
Διαίρεση δεδομένων	Οριζόντια	Οριζόντια ή κατακόρυφα
Βασικός παράγοντας επιβράδυνσης	Συνήθως η επικοινωνία, μη αξιόπιστο δίκτυο	Είτε οι υπολογισμοί είτε η επικοινωνία

Πίνακας 2.1: *Σύγκριση τυπικών συστημάτων cross-device και cross-silo federated learning*

2.3.4 Πλεονεκτήματα και προκλήσεις

Ένα από τα σημαντικότερα πλεονεκτήματα που προσφέρει η χρήση federated learning είναι η δυνατότητα αξιοποίησης μεγάλου όγκου, πιθανά ευαίσθητων, δεδομένων για την εκπαίδευση, που προέρχονται από διαφορετικές πηγές, χωρίς όμως οι συμμετέχοντες να χρειάζεται να τα μοιραστούν, καθώς η ιδιωτικότητα και ασφάλεια των clients είναι από τους βασικούς στόχους της μεθόδου. Σε άλλες περιπτώσεις, η συγκέντρωση ακόμα και ανωνυμοποιημένων δεδομένων μπορεί να θέσει σε κίνδυνο την ιδιωτικότητα των χρηστών συνδυάζοντάς τα με άλλες πληροφορίες (joins), όπως συμβαίνει σε αρκετές περιπτώσεις (π.χ. σε δημογραφικές μελέτες [35]). Μερικά ακόμα από τα βασικότερα πλεονεκτήματα του federated learning, πέρα από τα πλεονεκτήματα που προσφέρει η κατανομή των απαραίτητων υπολογισμών, είναι και οι διαφορές του από το συστήματα που επιλύουν το τυπικό πρόβλημα κατανεμημένης βελτιστοποίησης. Δηλαδή, δεν θέτει προϋποθέσεις σχετικά με την i.i.d. ιδιότητα των δεδομένων, λειτουργεί ακόμα και όταν τα δεδομένα είναι μη ισορροπημένα (ως προς το πλήθος τους) στους clients, επιτρέπει το πλήθος των συμμετεχόντων κόμβων να είναι μεγαλύτερο από το μέσο πλήθος δειγμάτων, ενώ αντιμετωπίζει περιπτώσεις περιορισμένης δυνατότητας επικοινωνίας.

Πέρα από τα εγγενή του πλεονεκτήματα, η εφαρμογή του federated learning, τόσο σε

ερευνητικό περιβάλλον, όσο και στην ευρύτερη χρήση του στον χώρο της τεχνολογίας, ενέχει πλήθος προκλήσεων. Λόγω της διαφορετικής φύσεώς τους, οι πιθανές μέθοδοι αντιμετώπισης διαφέρουν, τόσο ως προς την δυσκολία που επιχειρούν να υπερκεράσουν όσο και στην προσέγγιση που ακολουθούν. Παρακάτω παρουσιάζεται μια κατηγοριοποίηση των προκλήσεων του federated learning, οι οποίες το διαφοροποιούν από τα παραδοσιακά προβλήματα (κατανεμημένης) μάθησης [36].

Ετερογένεια συστημάτων

Σε ένα σύστημα federated learning, ειδικά στην περίπτωση cross-device, υπάρχουν σημαντικές διαφορές στα χαρακτηριστικά και τις δυνατότητες των συστημάτων που συμμετέχουν. Αυτές οι διαφορές μπορεί να οφείλονται στο hardware και την κατασκευή των συστημάτων (π.χ. διαθέσιμη μνήμη, επεξεργαστές και επιταχυντές που χρησιμοποιούν), αλλά και στην υποδομή που τα περιβάλλει, όπως για παράδειγμα το δίκτυο με το οποίο είναι συνδεδεμένα (το είδος, η ταχύτητα και η χωρητικότητά του). Επιπλέον, τα συστήματα μηχανικής μάθησης που βασίζονται σε ετερογενή υποσυστήματα, χρειάζεται να είναι ανεκτικά σε σφάλματα (fault tolerant) [36].

Ένας τρόπος για την αντιμετώπιση αυτού του είδους της ετερογένειας ονομάζεται *ενεργή δειγματοληψία (active sampling)* [36] των clients που θα συμμετάσχουν σε κάθε γύρο με βάση συγκεκριμένη πολιτική για την επιλογή τους. Για παράδειγμα, ο server μπορεί να λαμβάνει υπόψιν τους διαθέσιμους πόρους των clients [37]. Η επιλογή των clients, πέρα από καθοριστική για την ταχύτητα της εκπαίδευσης, είναι σημαντική και για την τελική επίδοση του μοντέλου που εκπαιδεύεται [38].

Στατιστική ετερογένεια

Ένα από τα χαρακτηριστικά του federated learning, όπως παρουσιάστηκαν στην ενότητα 2.3.2, είναι το γεγονός ότι δεν προϋποθέτει τα δεδομένα να είναι i.i.d. Η χαλάρωση αυτής της απαίτησης που υπάρχει στις περισσότερες κλασσικές μεθόδους μηχανικής μάθησης, επιφέρει προκλήσεις σχετικά με την σύγκλιση του αλγορίθμου καθώς και με την τελική επίδοση του μοντέλου. Ταυτόχρονα ένα σύστημα federated learning, χρειάζεται πολλές φορές να είναι δίκαιο ως προς τους clients, δηλαδή να μην παρουσιάζει μεροληψία το τελικό κοινό μοντέλο, επηρεαζόμενο από εκείνους που έχουν μεγαλύτερο όγκο δεδομένων εκπαίδευσης, στην περίπτωση που αυτά είναι ανισομερώς κατανεμημένα [36].

Το ερευνητικό ενδιαφέρον γύρω από το federated learning χρησιμοποιώντας μη i.i.d. δεδομένα είναι μεγάλο. Για παράδειγμα, στο [39] οι συγγραφείς δείχνουν ότι σε ακραίες περιπτώσεις ανισοκατανομής των κλάσεων, όταν ο κάθε client εκπαιδεύεται μόνο σε μια συγκεκριμένη κλάση, η χρήση του FedAvg οδηγεί σε σημαντική μείωση της ακρίβειας του τελικού μοντέλου, ενώ προτείνουν στρατηγική για τον περιορισμό του προβλήματος. Ακόμα έχουν προταθεί τεχνικές ώστε να αυξάνεται η αποδοτικότητα του federeated learning ακόμα και στην περίπτωση μη i.i.d. δεδομένων. Για παράδειγμα, στο [40] οι συγγραφείς προτείνουν μια νέα μέθοδο συμπίεσης που είναι ανθεκτική και σε μη i.i.d. δεδομένα, ενώ στο [41] προτείνεται ένα σύστημα το οποίο επιλέγει κατάλληλα ένα υποσύνολο των clients, ώστε να εξισορροπεί την μεροληψία που εισάγεται από τα μη i.i.d. δεδομένα.

Κόστος επικοινωνίας

Το κόστος της επικοινωνίας μεταξύ των συμμετεχόντων στην μάθηση, τόσο βάσει του χρόνου, όσο και βάσει του όγκου της πληροφορίας που ανταλλάσσεται (latency, round-trip time και bandwidth) αποτελεί σημαντικό κομμάτι της απόδοσης ενός συστήματος federated learning, ενώ μπορεί να αποτελέσει και τον κύριο παράγοντα καθυστερήσεων. Το γεγονός ότι μεταφέρονται στο δίκτυο μόνο οι ανανεώσεις των βαρών του μοντέλου βοηθάει στην ταχύτητα, όμως δεν είναι πάντα αρκετό. Στην βιβλιογραφία έχουν προταθεί λύσεις για την αντιμετώπιση αυτού του προβλήματος, όπως:

- Αύξηση του αριθμού των εποχών εκπαίδευσης σε κάθε client ανά γύρο [36], έτσι ώστε ένα σημαντικό υποσύνολο των υπολογισμών να πραγματοποιείται σε αυτούς, και να μειώνονται οι απαιτήσεις για συχνή επικοινωνία με τον server.

- Αποκεντροποιημένη μάθηση, πχ με την χρήση blockchain [42] ή με consensus μεταξύ των clients [43].

- Με χρήση μεθόδων συμπίεσης των δεδομένων που ανταλλάσσονται μεταξύ clients και server. Για παράδειγμα στο [28], οι συγγραφείς αναφέρουν δύο βασικές προσεγγίσεις που σχετίζονται με τις ανανεώσεις βαρών. Η πρώτη είναι η χρήση *structured updates*, όπου σκοπός είναι η μάθηση των ανανεώσεων από έναν πιο περιορισμένο χώρο που μπορεί να περιγραφεί από λιγότερες μεταβλητές. Σε αυτή την προσέγγιση εντάσσουν μεθόδους που περιορίζουν τον βαθμό του πίνακα που περιέχει τις ανανεώσεις καθώς και την απόκρυψη και μη αποστολή (*mask*) ενός τυχαίου (αλλά γνωστού) υποσυνόλου των τιμών του ώστε να είναι αραιός. Η δεύτερη προσέγγιση αφορά την χρήση *sketched updates*, όπου σκοπός είναι η συμπίεση της πληροφορίας που αποστέλλουν οι clients στους server. Σε αυτή την κατηγορία εμπίπτουν μέθοδοι όπως η υποδειγματοληψία κανονικοποιημένων τιμών του πίνακα ανανεώσεων (*subsampling*), ή η πιθανοκρατική κβάντωση (*probabilistic quantization*) με μείωση της ακρίβειας αναπαράστασης των βαρών του μοντέλου.

Ασφάλεια και ιδιωτικότητα

Στα πλαίσια του federated learning, οι όροι ασφάλεια (security) και ιδιωτικότητα (privacy) συχνά χρησιμοποιούνται από κοινού, παρόλαυτα αναφέρονται σε διαφορετικές ιδιότητες που πρέπει να έχει το σύστημα μάθησης. Ο όρος ασφάλεια αναφέρεται σε ανθεκτικότητα ενάντια σε μην εξουσιοδοτημένη ή κακόβουλη πρόσβαση στο σύστημα, αλλαγή των δεδομένων ή άρνηση πρόσβασης σε αυτά. Αντίθετα ο όρος εμπιστευτικότητα αναφέρεται στην ικανότητα προφύλαξης προσωπικών πληροφοριών από τα δεδομένα (π.χ. ιατρικά, ανωνυμοποιημένα δεδομένα από κάποιο dataset ανοιχτής πρόσβασης) και το σύνολο της διαδικασίας μάθησης. Κακόβουλες ενέργειες μπορεί να έχουν ως στόχο τόσο τον server, όσο και κάποιους από τους clients [44]. Επιπλέον μπορεί να γίνει μοντελοποίηση των επιθέσεων ενάντια σε federated learning συστήματα, ώστε να διαφοροποιούνται οι περιπτώσεις όπου οι απειλές προέρχονται μέσα από το ίδιο το σύστημα (*insider attacks*), ή από παράγοντα εκτός αυτού (*outsider attacks*). Τα insider attacks συνήθως εκτελούνται από client που ενεργεί κακόβουλα, ή από

τον ίδιο τον server, και είναι δυσκολότερο να αντιμετωπιστούν. Παράδειγμα outsider attack είναι το να κρυφακούει (eavesdropping) κάποιος την επικοινωνία μεταξύ clients και servers ώστε να υποκλέψει πληροφορίες [45].

Σχετικά με το θέμα της ασφάλειας, ένα σύστημα federated learning πρέπει να είναι σε θέση να εγγυάται εμπιστευτικότητα, ακεραιότητα και διαθεσιμότητα. Ένα παράδειγμα είδους επίθεσης που είναι πολύ πιθανό να πλήξει συστήματα federated learning ονομάζεται poisoning, και ως προς αυτό διακρίνονται δύο ευρύτερες κατηγορίες: data poisoning και model poisoning. Συνήθης στόχος του επιτιθέμενου είναι είτε να μειωθεί η ακρίβεια του μοντέλου ή να επηρεαστεί η πιθανότητα αυτό να προβλέψει κάποια συγκεκριμένη κλάση. Στο data poisoning γίνεται απόπειρα να προστεθούν στα δεδομένα δείγματα τα οποία θα επηρεάσουν αρνητικά την εκπαίδευση των τοπικών μοντέλων, με σκοπό να τροποποιήσουν τις ανανεώσεις των βαρών που αποστέλλονται στον server και συνεπώς και το τελικό μοντέλο. Στην ευρύτερη κατηγορία του data poisoning εντάσσονται και επιθέσεις στις οποίες τροποποιούνται τα υπάρχοντα δεδομένα. Αντίθετα με το data poisoning, στο οποίο ο επιτιθέμενος επιχειρεί να επηρεάσει το τελικό μοντέλο μέσω των δεδομένων, στο model poisoning η κακόβουλη ενέργεια το αφορά πιο άμεσα καθώς ο επιτιθέμενος στοχεύει στο να τροποποιήσει τις ανανεώσεις βαρών που αποστέλλουν οι clients στον server [33, 44, 45]. Στο [46] οι συγγραφείς δείξανε ότι στο πλαίσιο της ομοσπονδιακής μάθησης το model poisoning είναι πιο αποτελεσματικό από το data poisoning.

Όσον αφορά την ιδιωτικότητα των συστημάτων federated learning, πέρα από τον περιορισμό της πρόσβασης στα δεδομένα μόνο με την κατάλληλη εξουσιοδότηση, έχει αποδειχτεί ότι και οι ανανεώσεις βαρών, ακόμα και στα πλαίσια του federated learning που ανταλλάσσονται μεταξύ clients και server, μπορεί να την προσβάλουν και να διαρρεύσουν πληροφορίες (π.χ. [47, 48]). Επιθέσεις οι οποίες επιχειρούν να συμπεράνουν πληροφορίες για τα δεδομένα (π.χ. χαρακτηριστικά και κλάσεις δεδομένων ή αν ανήκει κάποιο παράδειγμα στα δεδομένα εκπαίδευσης) ονομάζονται *inference attacks* και αποτελούν σημαντικό κίνδυνο για την ιδιωτικότητα των συστημάτων federated learning [33, 45]. Μεταξύ άλλων, μέθοδοι για την διατήρηση της ιδιωτικότητας είναι η προσπάθεια επίτευξης *differential privacy* [49, 50, 51, 52] (π.χ. στα πλαίσια του federated learning γίνεται προσθήκη τεχνητού θορύβου στις παραμέτρους, πριν τον συμψηφισμό τους), η εφαρμογή *ομομορφικής κρυπτογράφησης (homomorphic encryption)* [50, 53], κατά την οποία οι πληροφορίες που μεταφέρονται μεταξύ των clients και του server κρυπτογραφούνται, και η χρήση *secure multiparty computation (SMC)* [51], που βασίζεται σε από κοινού υπολογισμούς με χρήση κρυπτογραφικών μεθόδων, χωρίς οι επιμέρους συμμετέχοντες να μοιράζονται ευαίσθητα δεδομένα.

Κεφάλαιο 3

Σχετικές εργασίες

Στο κεφάλαιο αυτό γίνεται σύντομη αναφορά σε ένα υποσύνολο εργασιών που αφορούν συστήματα ανίχνευσης εισβολής που αξιοποιούν (βαθιά) μηχανική μάθηση στην βιβλιογραφία, δίνοντας έμφαση σε εκείνα τα οποία χρησιμοποιούν τεχνικές federated learning. Τονίζεται ότι η ανασκόπηση αυτή δεν είναι εξαντλητική, αλλά γίνεται αναφορά συγκεκριμένων εργασιών, με σκοπό την ανάδειξη της πορείας των ερευνών επί του θέματος, δίνοντας περισσότερη προσοχή σε αυτές που στηρίζονται στην μηχανική μάθηση.

3.1 Συστήματα ανίχνευσης εισβολής με μηχανική μάθηση

Μια θεμελιώδης εργασία η οποία μοντελοποιεί την ανίχνευση εισβολής, εισάγοντας όρους (π.χ. προφίλ) για την περιγραφή της είναι η [3], στην οποία γίνεται η βασική υπόθεση ότι οι παραβιάσεις ασφαλείας μπορούν να ανιχνευτούν παρακολουθώντας τα αρχεία καταγραφής ενός συστήματος για ανώμαλα μοτίβα συμπεριφοράς.

Στην βιβλιογραφία υπάρχουν αρκετές δημοσιεύσεις ανασκοπικού χαρακτήρα, που στοχεύουν στην σύνοψη τεχνολογιών αλλά και στην ταξινόμησή τους. Μια πολύ γνωστή από αυτές είναι η [1] στην οποία γίνεται αναλυτική ανασκόπηση των μέχρι τότε εργασιών στα συστήματα ανίχνευσης εισβολής, ενώ ταυτόχρονα οι συγγραφείς προτείνουν μια συστηματική κατηγοριοποίηση και σύγκρισή τους βάσει της προσέγγισης και των τεχνολογιών που ακολουθούν, αναδεικνύοντας πλεονεκτήματα και μειονεκτήματά τους. Μια δεύτερη, στην οποία γίνεται εκτενής αναφορά και στην χρήση τεχνικών μηχανικής μάθησης είναι η [21].

Εφαρμογές μηχανικής μάθησης έχουν αξιοποιηθεί ευρέως στο παρελθόν στα πλαίσια της ασφάλειας υπολογιστικών συστημάτων και δικτύων [6, 54] μέσω ανίχνευσης εισβολών. Μεταξύ άλλων, έχουν χρησιμοποιηθεί Support Vector Machines [55, 56, 57], k-Nearest Neighbours [58, 59], decision trees / random forest [60, 61], αλλά και μη-επιβλεπόμενες μέθοδοι, όπως k-means clustering [62, 63].

Με την ανάπτυξη της βαθιάς μηχανικής μάθησης, κινήθηκε και προς αυτή το ερευνητικό ενδιαφέρον. Έτσι, η χρήση μοντέλων βαθιών μάθησης για τον σχεδιασμό και την υλοποίηση συστημάτων ανίχνευσης (και πρόληψης) εισβολής ή ανωμαλιών είναι ένα ενεργά αναπτυσσόμενο πεδίο, με δημοσιεύσεις που συνοψίζουν τις κατευθύνσεις προς τις οποίες κινείται [6, 64, 65]. Από αυτές προκύπτει ότι δοκιμάζονται διαφορετικά είδη μοντέλων, στα πλαίσια επιβλεπόμενης, μη επιβλεπόμενης μάθησης, καθώς και συνδυασμού τους. Παρακάτω παραθέτονται ορισμένα παραδείγματα:

- Deep (feed-forward) Neural Network, DNN: Στο [66] οι συγγραφείς εξετάζουν την αποτελεσματικότητα των DNN για διαφορετικά datasets, συγκρίνοντάς τα με άλλες μεθόδους μηχανικής και βαθιάς μάθησης και εν τέλει προτείνουν ένα ολοκληρωμένο IDS βασισμένο σε DNN. Στο [67] αξιολογείται ένα μοντέλο DNN βάσει του NSL-KDD [68], και στην συνέχεια συγκρίνεται η απόδοση του σε τεχνητά ανταγωνιστικά (adversarial) δεδομένα που παράγονται αλγοριθμικά.

- Recurrent Neural Networks, RNN: Ένα σημαντικό χαρακτηριστικό των RNN, είναι ότι διατηρούν εσωτερική κατάσταση, που χρησιμοποιείται κατά την εκπαίδευση και μπορούν να αναλύουν και χρονικές εξαρτήσεις μεταξύ των δεδομένων, επιτρέποντας έτσι να χρησιμοποιηθούν και για συστήματα τα οποία μοντελοποιούν την δικτυακή κίνηση ως χρονοσειρές. Μια από τις πιο γνωστές εργασίες που υλοποιούν IDS με την χρήση RNN είναι η [69] για binary και multiclass classification. Το ενδιαφέρον για την χρήση RNN κινήθηκε και προς πιο εξειδικευμένες μορφές τους, όπως τα Long Short Term Memory (LSTM), τα οποία αποτέλεσαν βάση για πολλά συστήματα (π.χ. [70, 71, 72]), αλλά και τα Gated Recurent Units, (GRU) (π.χ. [73]).

- Convolutional Neural Networks, (CNN). Αν και συνήθως τα CNN χρησιμοποιούνταν για την ανάλυση εικόνων και την όραση υπολογιστών, για την ικανότητα που έχουν να εξάγουν τοπικές συσχετίσεις, έχουν βρει εφαρμογές σε πολλούς τομείς, μεταξύ αυτών και τα συστήματα ανίχνευσης εισβολής. Ορισμένα παραδείγματα αποτελούν τα [74], [75] (στο οποίο οι συγγραφείς εκπαίδευσαν ένα βαθύ CNN, και στην συνέχεια χρησιμοποίησαν τις εξόδους από τα συνελικτικά επίπεδα ως εξαγόμενα χαρακτηριστικά, για να τα τροφοδοτήσουν σε πιο απλούς ταξινομητές SVM και 1-NN), καθώς και το [76] (στο οποίο χρησιμοποιείται και το CSE-CIC-IDS2018).

- Autoencoders, (AE): Οι βαθιοί autoencoders είναι πολύ διαδεδομένες αρχιτεκτονικές μη επιβλεπόμενης μάθησης που μεταξύ άλλων χρησιμοποιούνται για μείωση της διαστατικότητας των δεδομένων, εξαγωγή χαρακτηριστικών, ή μείωση θορύβου από τα δεδομένα [64, 65], ενώ υπάρχουν αρκετές παραλλαγές τους (π.χ. stacked AE, de-noising AE, sparse AE, variational AE). Παρεμφερή εφαρμογή βρίσκουν και στα συστήματα ανίχνευσης εισβολής ή ανωμαλιών, όπου χρησιμοποιούνται συνήθως σε συνδυασμό με μοντέλο επιβλεπόμενης μάθησης. Μπορούν να χρησιμοποιηθούν είτε αυτοτελώς με κάποιο softmax επίπεδο ώστε να παρέχουν προβλέψεις κλάσεων (π.χ. [77, 78]), είτε σε συνδυασμό με κάποιο ταξινομητή. Σημαντικό παράδειγμα αποτελεί η δημοσίευση [79], στην οποία εισάγεται η αρχιτεκτονική stacked nonsymmetric deep autoencoder (NDAE) (της οποίας γίνεται χρήση και στην παρούσα εργασία) και χρησιμοποιεί random forest ως ταξινομητή. Ακόμα, στο [80] οι συγγραφείς υλοποιούν και συγκρίνουν τέσσερις αρχιτεκτονικές, μια από τις οποίες αποτελείται από stacked symmetrical autoencoder που ακολουθείται από ένα βαθύ feed forward νευρωνικό δίκτυο για την ταξινόμηση. Δύο ακόμα παραδείγματα χρήσης autoencoder είναι το [81], το οποίο χρησιμοποιεί stacked sparse autoencoder για feature extraction και στην συνέχεια decision tree, SVM, ANN και C4.5 για επιλογή χαρακτηριστικών, καθώς και το [82] το οποίο παρουσιάζει μοντέλο Xgboost βασισμένο σε stacked sparse autoencoder για

την μάθηση λανθανόντων χαρακτηριστικών.

- Deep belief networks, (DBN): Τα DBN μπορούν να κατασκευαστούν ως stacked restricted Boltzmann machines και μπορούν να θεωρηθούν ως ένα είδος παραγωγικού αλγορίθμου βασιζόμενο σε μη επιβλεπόμενη μάθηση, που δύναται να χρησιμοποιηθεί για εξαγωγή χαρακτηριστικών με μείωση της διαστατικότητας αλλά και για ταξινόμηση όταν στο τελευταίο επίπεδο χρησιμοποιηθεί ταξινομητής [64, 65]. Δύο παραδείγματα ερευνών στα οποία έχει χρησιμοποιηθεί DBN είναι το [83], στο οποίο γίνεται σύγκριση με άλλα μοντέλα που χρησιμοποιούν DBN, SVM, και συνδυασμό τους, καθώς και το [84], στο οποίο προτείνεται μια μέθοδος που συνδυάζει την χρήση γενετικού αλγορίθμου με DBN.

3.2 Συστήματα ανίχνευσης εισβολής και CSE-CIC-IDS2018

Παρά το γεγονός ότι το CSE-CIC-IDS2018 αποτελεί ένα από τα πιο πρόσφατα datasets σχετικά με την ανίχνευση εισβολής, τα πλεονεκτήματα που προσφέρει (π.χ. δημόσια διαθέσιμο, πλήθος δεδομένων, ποικιλία κλάσεων) έναντι των προηγούμενων, έχουν οδηγήσει στο να χρησιμοποιηθεί ήδη σε πολλές ερευνητικές εργασίες, όπως γίνεται φανερό και από ανασκοπικές δημοσιεύσεις [85, 86]. Μερικές από τις εργασίες που το χρησιμοποιούν, με αναφορά στο μοντέλο που επιλέχθηκε, είναι οι εξής: [87] (χρήση LSTM με attention mechanism), [88] (χρήση convolutional autoencoder), [89] (χρήση CNN και σύγκριση του με RNN), [90] (μοντέλο δύο επιπέδων βασισμένο σε autoencoders), [91] (χρήση denoising autoencoder και ευριστικής μεθόδου για διαχωρισμό των κλάσεων), [80] (deep feed forward neural network, μεταξύ άλλων, καθώς και χρήση autoencoder πριν από τον ταξινομητή).

3.3 Συστήματα ανίχνευσης εισβολής και federated learning

Η ανάδειξη του federated learning ως μια βιώσιμη, πιο αποκεντρωμένη εναλλακτική μέθοδο μηχανικής μάθησης, η οποία μπορεί να φέρει πλεονεκτήματα ασφάλειας και ιδιωτικότητας για τους συμμετέχοντες, διατηρώντας υψηλές επιδόσεις, συνέβαλε τα τελευταία χρόνια στην έρευνα και υλοποίηση συστημάτων ανίχνευσης εισβολής που βασίζονται σε τεχνικές ομοσπονδιακής μάθησης. Αυτό φαίνεται και από πρόσφατες ανασκοπικές εργασίες που παρουσιάζουν τέτοιου είδους εφαρμογές προσπαθώντας να τις συνοψίσουν και να τις κατηγοριοποιήσουν (για μια πιο εξαντλητική ανασκόπηση βλ. [8, 92, 93]). Παρακάτω αναφέρονται ορισμένες από τις εργασίες που σκοπό έχουν την ανάπτυξη IDS με χρήση μεθόδων federated learning.

Ένα κομμάτι της βιβλιογραφίας αφορά federated learning IDS για περιβάλλον Internet of Things, (IoT), για παράδειγμα: [94] (προτείνεται ένα αποκεντρωμένο σύστημα fedearted learning IDS, με χρήση τεχνολογιών blockchain για έλεγχο πρόσβασης), [95] (χρήση μοντέλου CNN, με τροποποιημένο FedAvg: FedACNN), [96] (ανάπτυξη IDS για IoT στον αγροτικό τομέα, με χρήση μοντέλων DNN, CNN, LSTM, σε τρία διαφορετικά dataset, μεταξύ των οποίων και το CSE-CIC-IDS2018), Στο [97] οι συγγραφείς προτείνουν σύστημα differential privacy που βασίζεται σε ομάδες (cohorts) από clients που έχουν ετερογενείς απαιτήσεις

ιδιωτικότητας, κάνοντας και χρήση μεθόδων συνεχούς μάθησης (continual learning). Η αξιολόγηση του γίνεται με την χρήση DNN στο CSE-CIC-IDS2018). Υπάρχουν εργασίες που επικεντρώνονται στην ανίχνευση κατανεμημένων επιθέσεων άρνησης υπηρεσιών όπως για παράδειγμα το [98] στο οποίο προτείνονται μέθοδοι hierarchical aggregation και resampling για την αντιμετώπιση του class imbalance, το [99] στο οποίο γίνεται χρήση μη i.i.d. δεδομένων, καθώς και το [100] στο οποίο παρουσιάζεται federated learning μοντέλο το οποίο μπορεί να εκπαιδεύεται κατ' επανάληψη με δεδομένα νέων επιθέσεων.

Ασφαλώς έρευνα πραγματοποιείται και για διαφορετικούς τομείς. Στο [101] οι συγγραφείς επικεντρώνονται στην υλοποίηση συστήματος FL IDS με χρήση DNN για έξυπνο 5G metering δίκτυο. Στο [102], περιγράφεται το DeepFed, ένα σύστημα FL IDS που χρησιμοποιεί CNN και GRU, για την προστασία βιομηχανικών cyber-physical συστημάτων. Στο [103] οι συγγραφείς, δίνοντας έμφαση στην επεξηγηματικότητα του μοντέλου, υλοποιούν και αξιολογούν σε διαφορετικά datasets το FEDFOREST, ένα FL IDS που βασίζεται σε Gradient Boosting Decision Trees. Στο [104] η μελέτη επικεντρώνεται στην προστασία ασύρματων δικτύων με χρήση federated learning βασισμένο σε stacked autoencoders και το μοντέλο αξιολογείται βάση του Agean Wi-Fi Intrusion Dataset (AWID) [105]. Τέλος, στο [106] οι συγγραφείς επικεντρώνονται στον χώρο της υγείας και τα eHealth δίκτυα, προτείνοντας ένα σύστημα για εύρεση ανωμαλιών, το οποίο προσφέρει ισχυρότερη ιδιωτικότητα μέσω federated learning και differential privacy.

Κεφάλαιο 4

Περιγραφή θέματος

Στο κεφάλαιο αυτό αρχικά παρουσιάζονται οι στόχοι τους οποίους επιχειρούμε να πετύχουμε με την ανάπτυξη ενός υβριδικού συστήματος ανίχνευσης εισβολής που εκπαιδεύεται με μεθόδους federated learning σε cross-silo περιβάλλον. Στην συνέχεια περιγράφεται η αρχιτεκτονική του συστήματος που υλοποιήθηκε, αναφέροντας τα πλεονεκτήματα τα οποία μπορεί να προσφέρει, καθώς και το σκεπτικό πίσω από βασικές αρχιτεκτονικές επιλογές.

4.1 Στόχοι του προτεινόμενου federated IDS

Εφοδιασμένοι με την κατανόηση βασικών εννοιών και προκλήσεων των συστημάτων ανίχνευσης εισβολής και του federated learning, γίνονται πλέον αντιληπτά τόσο τα πλεονεκτήματα που μπορεί να προσφέρει ο συνδυασμός τους στα πλαίσια συνεργατικής μάθησης, όσο και ορισμένες δυνατότητες που πρέπει να έχει η υλοποίηση ενός τέτοιου συστήματος.

Μια από τις πιο σημαντικές ιδιότητες που πρέπει να πληροί το συγκεκριμένο σύστημα είναι η δυνατότητα απομόνωσης των δεδομένων στους client, χωρίς να γίνεται μεταφορά αυτών, ούτε μεταξύ των clients, ούτε με τον server. Ο συγκεκριμένος στόχος δικαιολογείται μεταξύ άλλων, τόσο λόγω της επιβάρυνσης που θα προέκυπτε στο δίκτυο μια τέτοια μεταφορά μεγάλου όγκου δεδομένων, όσο και για την δυνατότητα των clients να μην μοιραστούν τα δεδομένα τα οποία διαθέτουν για λόγους ιδιωτικότητας, καταφέρνοντας ωστόσο να συμμετάσχουν ενεργά στην διαδικασία μάθησης, συνεισφέροντας με την χρήση των δεδομένων τους.

Επιπλέον, μας ενδιαφέρει οι clients να μπορούν να εκτελούν το κομμάτι της εκπαίδευσης του τοπικού τους μοντέλου παράλληλα, χωρίς να εξαρτάται η πορεία της εκπαίδευσης σε κάθε γύρο από την πρόοδο των υπολοίπων. Κάτι τέτοιο είναι αναγκαίο καθώς διαφορετικοί clients είναι πιθανό να έχουν διαφορετική υπολογιστική ισχύ, καθώς και προτεραιοποίηση της εκπαίδευσης που εκτελούν, σε σχέση με άλλες λειτουργίες τους.

Σημαντική είναι ακόμα η δυνατότητα του τελικού μοντέλου να μπορεί να εντοπίζει κακόβουλη κίνηση όταν αυτή είναι πιο σπάνια σε σχέση με την φυσιολογική κίνηση στο δίκτυο. Αυτό αποτελεί φυσική συνέπεια του γεγονότος ότι η πλειοψηφία των ενεργειών στο δίκτυα και τα συστήματα που προστατεύει το IDS προέρχεται, υπό φυσιολογικές συνθήκες, από ενέργειες έννομων χρηστών ή εφαρμογών. Έτσι, πέρα από την δυνατότητα του federated IDS να εκπαιδεύεται σωστά και με την χρήση μη ισορροπημένου συνόλου δεδομένων, ιδανικά θα

μπορεί να μειώσει στο ελάχιστο περιπτώσεις εσφαλμένου συναγερμού, λόγω ψευδώς θετικών αποτελεσμάτων, που δύναται να προκύψουν από έγκυρη συμπεριφορά που αποκλίνει όμως από την πλειοψηφία της κίνησης στο δίκτυο.

Πέρα από την δυνατότητα ανίχνευσης επίθεσης και δυϊκού διαχωρισμού μεταξύ φυσιολογικής και κακόβουλης κίνησης, είναι χρήσιμο για τους διαχειριστές των συστημάτων και των δικτύων που προστατεύονται, να μπορεί το IDS να ταξινομεί και το είδος της επίθεσης (binary vs multiclass classification). Αυτό συμβαίνει, καθώς ο τρόπος αντίδρασης και προστασίας έναντι κακόβουλων ενεργειών είναι άρρηκτα συνδεδεμένος με το είδος τους, λόγω διαφορετικών στόχων (άρνηση υπηρεσιών, παρακολούθηση, υποκλοπή δεδομένων κ.τ.λ.), και επειδή ανάλογα με το είδος και την συχνότητά της εκάστοτε επίθεσης δύναται να ληφθούν διαφορετικά μέτρα ενεργούς πρόληψης.

Τέλος, είναι σημαντικό για την σωστή λειτουργιά του federated συστήματος ανίχνευσης εισβολής, να επαληθευθεί ότι λειτουργεί αξιόπιστα ακόμα και σε περιπτώσεις που η i.i.d. υπόθεση για τα δεδομένα δεν ισχύει. Καθώς το τελικό μοντέλο προκύπτει από την συνεργασία πολλαπλών clients, είναι πιθανό η κίνηση και οι επιθέσεις που έχει καταγράψει ο καθένας στα δεδομένα του να μην ακολουθούν κοινή κατανομή. Για παράδειγμα, κάποιο υποσύνολο των clients μπορεί να δέχεται συγκεκριμένες κατηγορίες επιθέσεων πολύ συχνότερα από κάποιο άλλο, με συνέπεια τα δεδομένα που διαθέτουν να είναι μη ισορροπημένα, ή ορισμένοι clients οι οποίοι συσχετίζονται μεταξύ τους (π.χ. συμμετέχουν σε κάποιο κοινό δίκτυο, πέρα από το περιβάλλον του federated learning) να παρουσιάζουν συσχέτιση και στο είδος επιθέσεων που δέχονται.

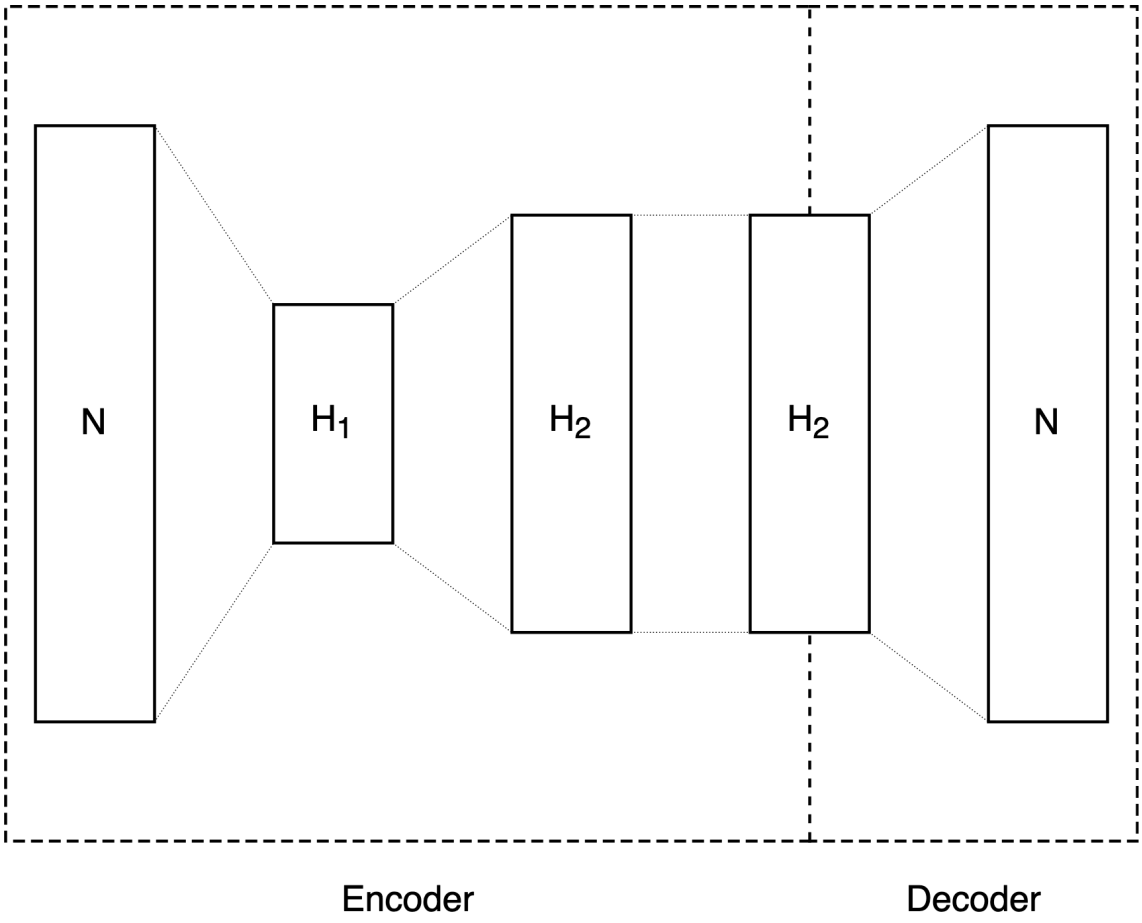
4.2 Περιγραφή αρχιτεκτονικής

4.2.1 Αρχιτεκτονική μοντέλου

Οι βασικές επιλογές για την αρχιτεκτονική του μοντέλου λαμβάνονται τόσο βάσει των στόχων που επιχειρεί να πετύχει, όσο και με επίγνωση των συνθηκών και των δυσκολιών που αντιμετωπίζονται συχνά σε ένα σύγχρονο κατανεμημένο περιβάλλον, το οποίο βρίσκεται υπό τον κίνδυνο επιθέσεων.

Μια πολύ σημαντική πρόκληση που προκύπτει κατά την σχεδίαση IDS τα οποία βασίζονται σε μηχανική μάθηση είναι η δυσκολία (ή ακόμα και ανικανότητα) να αποκτήσει κανείς έγκυρα επισημασμένα (labeled) δεδομένα. Αυτό συμβαίνει καθώς η διαδικασία ανάθεσης ετικετών είναι χρονοβόρα και μπορεί να απαιτεί ειδική γνώση και ανθρώπινη παρέμβαση, ενώ συχνά μόνο ένα ποσοστό των δεδομένων μπορεί να επισημανθεί επιτυχώς. Ακόμα, δεν είναι βέβαιο ότι σε περίπτωση πιθανής επίθεσης, εκείνη θα ταιριάζει με μια από τις κλάσεις στις οποίες έχει εκπαιδευτεί το μοντέλο [107].

Στην βιβλιογραφία (όπως φαίνεται και από το κεφάλαιο 3), πέρα της επιβλεπόμενης μάθησης, έχουν αναπτυχθεί IDS που βασίζονται σε μη επιβλεπόμενη μάθηση (π.χ. για anomaly detection), καθώς και υβριδικά, στην τελική αρχιτεκτονική των οποίων συνδυάζονται μοντέλα και από τις δύο κατηγορίες. Στηριζόμενοι τόσο στην πρακτική δυσκολία εύρεσης επισημασμένων δεδομένα στον πραγματικό κόσμο όσο και στην διάδοση της χρήσης μη επιβλεπόμενης μάθησης, επιλέγουμε ως μοντέλο βάσης (baseline) για τα πειράματα ένα υβρι-



Σχήμα 4.1: *Nonsymmetric deep autoencoder, NDAE*

δικό μοντέλο που συνδυάζει έναν βαθύ autoencoder, και ένα βαθύ feed-forward νευρωνικό δίκτυο. Οι autoencoder, πέρα από την δυνατότητα να εκπαιδεύονται με μη επισημασμένα δεδομένα, χρησιμοποιούνται συχνά για εξαγωγή χαρακτηριστικών (feature extration) και για μείωση της διαστατικότητας των δεδομένων (dimentionality reduction). Ακόμα, σε σχέση με το PCA (στην τυπική του τουλάχιστον μορφή), έχουν την δυνατότητα να εντοπίζουν και μη γραμμικές συσχετίσεις οδηγώντας σε πιο ισχυρές γενικεύσεις [79].

Πιο συγκεκριμένα, για το κομμάτι του μοντέλου που αφορά την μη επιβλεπόμενη μάθηση θα χρησιμοποιήσουμε τον βαθύ μη συμμετρικό autoencoder *(nonsymmetric deep autoencoder, NDAE)*, όπως περιγράφεται αναλυτικά στο [79]. Το συγκεκριμένο δίκτο autoencoder μπορεί να μάθει και να εξάγει μη τετριμμένα χαρακτηριστικά, κλιμακώνει καλά και για δεδομένα εισόδου περισσότερων διαστάσεων, ενώ απορρίπτοντας τον παραδοσιακό συμμετρικό χαρακτήρα μπορεί να μειωθεί το υπολογιστικό και χρονικό κόστος με ελάχιστες επιπτώσεις στην επίδοση του [79]. Στο 4.1 φαίνεται ένα παράδειγμα nonsymmetric deep autoencoder, με N να είναι η διάσταση των δεδομένων εισόδου (και ανακατασκευής), και H_1, H_2 να είναι οι διαστάσεις του πρώτου, του δεύτερου και τρίτου κρυφού επιπέδου αντίστοιχα.

Κατά την διαδικασία μάθησης, συγκρίνουμε την έξοδο του NDAE με την αρχική είσοδο, με σκοπό να είναι όσο το δυνατόν παραπλήσιες. Η διαφορά τους χρησιμοποιείται ως σφάλμα κατά το backpropagation, για την ανανέωση των βαρών του δικτύου. Στο τέλος της

μάθησης, απομονώνουμε τον encoder, έχοντας έτσι ένα δίκτυο το οποίο μπορεί να λάβει είσοδο μεγέθους N, και να την κωδικοποιήσει σε $H_2 < N$ διαστάσεις έχοντας κάνει εξαγωγή μη τετριμμένων χαρακτηριστικών και εντοπισμό μη γραμμικών συσχετίσεων. Στο [79] γίνεται stack δύο τέτοιων NDAEs, ώστε να προκύψει τελικά βαθύτερο μοντέλο. Δοκιμάζοντας παρόμοιες πιθανές βαθύτερες αρχιτεκτονικές στην παρούσα εργασία, δεν παρατηρήθηκε βελτίωση στα αποτελέσματα, ενώ όπως ήταν αναμενόμενο ο χρόνος εκπαίδευσης αυξήθηκε, συνεπώς γίνεται χρήση ενός NDAE. Γενικότερα όμως, διαφορετικά datasets θα μπορούσαν να ωφεληθούν από την χρήση βαθύτερης αρχιτεκτονικής με autoencoder, είτε αυτοί εκπαιδεύονταν συνολικά με backpropagation, είτε ο κάθε autoencoder ξεχωριστά.

Για το κομμάτι του classification θα χρησιμοποιήσουμε ένα feed-forward βαθύ νευρωνικό δίκτυο, το οποίο παρά την σχετική απλότητα του σε σύγκριση με πιο σύνθετες αρχιτεκτονικές (π.χ. CNN, RNN) είναι σε θέση να αποδώσει παρόμοια ή και καλύτερα σε εφαρμογές ανίχνευσης επίθεσης, απαιτώντας μικρότερο χρόνο μάθησης [80]. Περισσότερες πληροφορίες για τις υπερπαραμέτρους του δικτύου δίνονται στο κεφάλαιο που περιγράφει τις λεπτομέρειες υλοποίησης και τα πειράματα.

4.2.2 Διάταξη federated learning

Για την μοντελοποίηση του federated learning, θεωρούμε cross-silo περιβάλλον, το οποίο υποθέτει δικτυακά αξιόπιστους clients (με την έννοια ότι η επικοινωνία του server με αυτούς είναι σταθερή και υπάρχουν μηδαμινές διακοπές), οι οποίοι μπορούν να διαθέσουν επαρκή υπολογιστική ισχύ για την εκπαίδευση και χρήση του μοντέλου. Μια τέτοια υπόθεση είναι λογική για συστήματα ανίχνευσης εισβολής που βασίζονται σε cross-silo federated learning, καθώς μπορούν να εκπαιδευτούν συνεργατικά και να εφαρμοστούν μεταξύ υποδικτύων ή υποσυστημάτων ενός μεγαλύτερου οργανισμού (π.χ. τμήματα ή σχολές ενός πανεπιστημίου), αλλά και μεταξύ διαφορετικών οργανισμών, η επικοινωνία των οποίων στηρίζεται σε αξιόπιστο δίκτυο, όπως εθνικά δίκτυα υποδομών (π.χ. GRNET), Internet Exhange Points (IXes ή IXPs), ή νοσοκομεία και μονάδες υγείας.

Η τελική διαρρύθμιση που ακολουθείται για το federated learning, είναι όμοια με αυτή που φαίνεται στο σχήμα 2.1, στο οποίο φαίνονται και τα βήματα τα οποία ακολουθούνται κατά την εκπαίδευση. Μια διαφορά είναι ότι στην συγκεκριμένη περίπτωση οι clients δεν είναι κινητές συσκευές, αλλά servers που διαθέτουν αυξημένη υπολογιστική ισχύ και αξιόπιστη δικτυακή παρουσία.

Η εκπαίδευση των τοπικών μοντέλων στους clients γίνεται παράλληλα σε κάθε γύρο, παρέχοντας έτσι ανοχή στην ετερογένεια των δυνατοτήτων του καθενός, και ασύγχρονες αποστολές των ανανεώσεων βαρών στον server. Είναι σημαντικό να αναφέρουμε ότι θέλοντας να πετύχουμε πλήρως federated μάθηση, πέρα από το DNN, και το NDAE εκπαιδεύεται με ομοσπονδιακή μάθηση ώστε να προκύψει το τελικό κοινό μοντέλο το οποίο διανέμεται στους clients. Η μη επιβλεπόμενη federated εκπαίδευση η οποία αφορά τον autoencoder, γίνεται πριν και ανεξάρτητα από την federated εκπαίδευση του DNN. Αυτό επιτρέπει στο σύστημα να βελτιώνει συνεχώς το NDAE, με μη επισημασμένα δεδομένα που συγκεντρώνουν οι clients, και η εκπαίδευση του classifer να γίνεται σε δεύτερο χρόνο όταν έχουν συγκεντρωθεί όσα επισημασμένα δεδομένα κρίνονται αρκετά για την εκπαίδευση του.

Τέλος, στο αλγοριθμικό σκέλος, ακολουθούμε μια διαδεδομένη προσέγγιση που διαφέρει λίγο από το federated averaging όπως είχε οριστεί στην αρχική του μορφή, έτσι ώστε κατά τον συμψηφισμό και την ανανέωση των βαρών στο συνολικό μοντέλο, ο server να χρησιμοποιεί learning rate διάφορο της μονάδας στο gradient descent. Περισσότερες λεπτομέρειες σχετικά με τις υπερπαραμέτρους του συστήματος, δίνονται στην ενότητα που περιγράφει τις λεπτομέρειες υλοποίησης και τα πειράματα.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 5

Dataset και προεπεξεργασία δεδομένων

Στο κεφάλαιο αυτό παρουσιάζεται το dataset που χρησιμοποιήθηκε καθώς και η απαραίτητη προεπεξεργασία που χρειάστηκαν τα δεδομένα ώστε να έρθουν σε κατάλληλη μορφή για τα πειράματα.

5.1 Περιγραφή του CSE-CIC-IDS2018

Το dataset που επιλέχθηκε είναι το CSE-CIC-IDS2018, το οποίο αποτελεί προϊόν συνεργασίας των Communications Security Establishment (CSE) και Canadian Institute for Cybersecurity (CIC). Οι βασικότεροι λόγοι επιλογής του συγκεκριμένου dataset είναι ότι περιέχει μεγάλη ποικιλία δεδομένων που έχουν παραχθεί βάσει των χαρακτηριστικών ρεαλιστικής δικτυακής κίνησης, είναι δημόσια διαθέσιμο και παρέχονται λεπτομέρειες για την παραγωγή του, και ότι παρά το γεγονός ότι είναι πρόσφατο, είναι ήδη διαδεδομένη η χρήση του στην βιβλιογραφία. Σε αυτή την ενότητα θα περιγραφεί ο τρόπος παραγωγής του, η δομή του και στην συνέχεια η προεπεξεργασία που απαιτήθηκε για την χρήση του.

5.1.1 Παραγωγή και διάθεση dataset

Το CSE-CIC-IDS2018 διατίθεται δημοσίως μέσω του μητρώου Registry of Open Data του Amazon Web Services (AWS) στην ιστοσελίδα [108], με τίτλο A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018). Αναλυτική περιγραφή για το τρόπο σύνθεσης του καθώς και λεπτομέρειες για τα δεδομένα που περιέχει μπορούν να βρεθούν στην ιστοσελίδα του University of New Brunswick (UNB) [109].

Για την παραγωγή της δικτυακής κίνησης, η υποδομή προσομοίωσης που χρησιμοποιήθηκε περιείχε συνολικά 50 υπολογιστές που δρούσαν ως επιτιθέμενοι (με λειτουργικό σύστημα Kali Linux), καθώς και 30 servers και 420 υπολογιστές (με λειτουργικό σύστημα Ubuntu και Windows) ως τα θύματα, οργανωμένοι σε 5 τμήματα. Το τελικό dataset αποτελείται από τα αρχεία καταγραφής κίνησης (pcap) και τις καταγραφές συστήματος των συσκευών του δικτύου (log files), μαζί με συνολικά 79 features που εξήχθησαν από αυτά με την χρήση του λογισμικού CICFlowMeter-V3 [110]. Η εξαγωγή χαρακτηριστικών με την χρήση του CICFlowMeter-V3 βασίζεται στην έννοια του *flow* που ορίστηκε στην ενότητα 2.1. Το λογισμικό αυτό έχει την δυνατότητα να παράγει αμφίδρομα flows, ορίζοντας τις κατευθύνσεις forward (από την πηγή προς τον προορισμό) και backward (από τον προορισμό προς

την πηγή), με βάση το πρώτο πακέτο του flow. Η κατεύθυνση αυτή δηλώνεται σε αρκετά από τα τελικά χαρακτηριστικά του CSE-CIC-IDS2018.

Όλα τα αρχεία βρίσκονται αποθηκευμένα σε δημόσιο S3 bucket του AWS. Εκεί περιέχονται τα αρχεία καταγραφής δικτυακής κίνησης και συστήματος (pcap & log files) καθώς και αρχεία CSV που έχουν εξαχθεί από τα πρώτα, για βολικότερη χρήση σε εφαρμογές μηχανικής μάθησης και τεχνητής νοημοσύνης, και είναι οργανωμένα ανά μέρα.

5.1.2 Περιεχόμενα του dataset

Η παραγωγή δικτυακής κίνησης για τον σχηματισμό του συγκεκριμένου dataset χρησιμοποιεί την έννοια των *προφίλ (profile)* που περιγράφει με αφηρημένο τρόπο την συμπεριφορά χρηστών (ατόμων ή εφαρμογών) στο δίκτυο. Πιο συγκεκριμένα στην [109] αναφέρονται δύο τέτοια προφίλ: Τα B-profile, που συνοψίζουν την φυσιολογική κίνηση μοντελοποιούν την συμπεριφορά των χρηστών κάνοντας χρήση στατιστικών εργαλείων για την δημιουργία χαρακτηριστικών, καθώς και τα M-profiles που περιγράφουν σενάρια επιθέσεων.

Παρουσιάζονται συνοπτικά παρακάτω τα σενάρια επιθέσεων που υλοποιήθηκαν στο συγκεκριμένο dataset σύμφωνα με την ιστοσελίδα που το περιγράφει [109]:

- *Επιθέσεις άρνησης υπηρεσιών (Denial of Service attacks, DoS).* Σκοπός αυτών των επιθέσεων είναι η μείωση ή εξάλειψη της λειτουργικότητας που μπορεί να προσφέρει ένα σύστημα (π.χ. ένας διακομιστής). Κάτι τέτοιο μπορεί να συμβεί όταν κάποιος κακόβουλος χρήστης ή λογισμικό εξαντλεί τους διαθέσιμους πόρους ενός συστήματος, με αποτέλεσμα οι φυσιολογικοί χρήστες να μην μπορέσουν να εξυπηρετηθούν. Συχνά η δικτυακή κίνηση που προκαλεί την επίθεση προέρχεται από πολλαπλές κατανεμημένες πηγές, με τις επιθέσεις αυτές να ονομάζονται κατανεμημένες επιθέσεις άρνησης υπηρεσιών (Distributed Denial of Service attacks, DDoS). Για τις επιθέσεις αυτού του είδους, οι δημιουργοί του dataset χρησιμοποίησαν τα εξής εργαλεία λογισμικού: Slowloris, Hulk, GoldenEye, Slowhttptest, καθώς και τα Low/High Orbit Ion Canon (LOIC, HOIC) για επιθέσεις DDoS.

- Επιθέσεις που βασίζονται σε ευάλωτα σημεία (vulnerabilities) συστημάτων. Τέτοιου είδους ευάλωτα σημεία συχνά δεν γίνονται άμεσα αντιληπτά και παρέρχεται χρόνος μέχρι να διορθωθούν. Συνεπώς, στο ενδιάμεσο διάστημα, μπορούν να γίνουν επιθέσεις στα συστήματα που διαθέτουν αυτά τα ευάλωτα σημεία και να τα εκμεταλλευθούν. Ένα από τα πιο γνωστά παραδείγματα είναι το HeartBleed bug (CVE-2014-0160) [111], [112], με το οποίο η ευρέως χρησιμοποιούμενη κρυπτογραφική βιβλιοθήκη OpenSSL επιτρέπει σε επιτιθέμενους να έχουν πρόσβαση σε ευαίσθητα δεδομένα από την μνήμη των συστημάτων που επηρεάζονται. Για την εκμετάλλευση του Heartbleed στο dataset έγινε χρήση του εργαλείου heartleech.

- Παρείσφρηση (infiltration) στο δίκτυο. Για την προσομοίωση τέτοιας επίθεσης, ένα αρχείο που περιέχει κακόβουλο λογισμικό αποστέλλεται σε χρήστη του δικτύου μέσω e-mail. Όταν εγκατασταθεί το λογισμικό αυτό, μπορεί πλέον να λειτουργήσει μέσα στο δίκτυο εκμεταλλευόμενο αδυναμίες του. Κατά την παραγωγή του συγκεκριμένου dataset, μετά την είσοδο του κακόβουλου λογισμικού στο δίκτυο, εκτελείται διερεύνηση

θυρών (port scan) και απαρίθμηση διαθέσιμων υπηρεσιών με χρήση του λογισμικού Nmap.

- Επιθέσεις ωμής βίας (brute force attacks). Αυτές οι επιθέσεις έχουν ως σκοπό την επίτευξη κάποιοι στόχου (συχνά την εύρεση κωδικών πρόσβασης για κάποιο σύστημα ή υπηρεσία) δοκιμάζοντας εξαντλητικά όλες τις πιθανές λύσεις. Σημαντικό μειονέκτημά τους είναι ο μεγάλος χρόνος που χρειάζεται για να ολοκληρωθούν όταν το σύνολο των πιθανών λύσεων είναι μεγάλο. Παρόλα αυτά χρησιμοποιούνται συχνά και υπάρχει πληθώρα εργαλείων για αυτές. Στο συγκεκριμένο dataset, γίνεται χρήση του λογισμικού patator για brute force επίθεση σε SSH και FTP, με την χρήση λεξικού που περιέχει ενενήντα εκατομμύρια εγγραφές.

- Botnet. Πρόκειται για ένα δίκτυο από υπολογιστές, μολυσμένους από κάποιο κακόβουλο πρόγραμμα, οι οποίοι συχνά συντονίζονται και συνεργάζονται για κάποιο κοινό σκοπό. Οι δημιουργοί του CSE-CIC-IDS2018 χρησιμοποίησαν δύο τέτοια λογισμικά. Το πρώτο είναι το Zeus, το οποίο είναι ένα trojan που μπορεί να τρέχει σε κάποιες εκδόσεις των Windows. Χρησιμοποιείται συχνά για επιθέσεις man-in-the-middle στις οποίες μπορεί να υποκλέπτει τραπεζικές πληροφορίες (π.χ. καταγράφοντας την πληκτρολόγηση του χρήστη). Το δεύτερο λογισμικό που χρησιμοποιήθηκε ονομάζεται Ares και είναι ένα trojan ανοιχτού κώδικα, το οποίο μπορεί μεταξύ άλλων λειτουργιών, να καταγράφει την πληκτρολόγηση και την οθόνη, καθώς και να επιτρέπει απομακρυσμένη σύνδεση με τερματικό.

- Επιθέσεις εφαρμογών ιστού (Web attacks). Γίνεται χρήση της Damn Vulnerable Web App (DVWA), η οποία είναι μια PHP/MySQL εφαρμογή που περιέχει εσκεμμένα κενά ασφαλείας, με σκοπό, μεταξύ άλλων, την χρήση της για εξάσκηση επαγγελματιών στον χώρο της τεχνολογίας. Για την πραγματοποίηση και αυτοματοποίηση επιθέσεων cross-site scripting (XSS) και brute force επιθέσεων, οι δημιουργοί του dataset υλοποίησαν κώδικα αυτοματοποίησης με την χρήση του Selenium.

5.2 Προεπεξεργασία δεδομένων

Η αποθήκευση των αρχείων στο περιβάλλον όπου εκπονείται η εργασία γίνεται με την χρήση του AWS CLI και γίνεται χρήση όλων των αρχείων CSV του dataset. Τα βασικά πακέτα Python που χρησιμοποιούνται για την προεπεξεργασία των δεδομένων είναι τα πολύ διαδεδομένα numpy [113], pandas [114] και scikit-learn [115]. Καθώς το συνολικό μέγεθος των CSV αρχείων είναι σχεδόν 7GB, και δεν είναι αποδοτικό να φορτωθούν όλα τα δεδομένα στην μνήμη, χρησιμοποιείται και το πακέτο Dask [116], το οποίο επιτρέπει την παράλληλη (ή και κατανεμημένη) επεξεργασιά δεδομένων, χωρίς αυτά να βρίσκονται απαραίτητα στην μνήμη, παρέχοντας ένα υψηλού επιπέδου API για την διαχείρισή τους, παρόμοιο με αυτό του pandas.

5.2.1 Καθαρισμός δεδομένων

Τα αρχεία είναι οργανωμένα ανά μέρα, κάτι που μας επιτρέπει να ξεκινήσουμε την επεξεργασία τους ξεχωριστά. Το καθένα από αυτά περιέχει 79 features τα οποία έχουν παραχθεί με την χρήση του CICFlowMeter-V3.

Επειδή τα ονόματα των features που περιέχονται στην πρώτη γραμμή των αρχείων περιέχουν κενά στην αρχή ή τέλος του, τα αφαιρούμε για ευκολότερη διαχείρισή τους στην συνέχεια. Επιπλέον βεβαιωνόμαστε ότι έχουν αφαιρεθεί όλα τα features που σχετίζονται με το δίκτυο προσομοίωσης και τον ορισμό των flows. Ως πρώτο βήμα στην προεπεξεργασία των δεδομένων, αφαιρούμε τις γραμμές οι οποίες περιέχουν κενές, NA (απούσες) ή άπειρες τιμές. Για τις συγκεκριμένες εγγραφές μια επιλογή θα ήταν να γίνουν impute οι μη έγκυρες τιμές (δηλαδή να αντικατασταθούν από αντίστοιχες έγκυρες τιμές), καθώς όμως ο όγκος των δεδομένων είναι μεγάλος, μπορούμε να τις αφαιρέσουμε χωρίς να επηρεαστεί η μάθηση. Επιπλέον παρατηρείται ότι, πέραν της πρώτης, κάποιες γραμμές στο κάθε αρχείο επαναλαμβάνουν τα ονόματα των features, συνεπώς αφαιρούνται και αυτές. Έπειτα από αυτόν τον αρχικό καθαρισμό κάθε αρχείου, τα συνενώνουμε ώστε να συνεχιστεί η επεξεργασία τους, στο σύνολό τους, πιο αποδοτικά με την χρήση του Dask. Σε αυτό το σημείο, το dataset περιέχει 16.137.183 εγγραφές που περιγράφουν πακέτα, ενώ το αρχικό πλήθος τους ήταν 16.232.943, δηλαδή αφαιρέθηκε περίπου το 0.59% αυτών.

Στην συνέχεια, εξερευνώντας τα δεδομένα, παρατηρούμε ότι 8 από τα features έχουν μηδενική τυπική απόκλιση, δηλαδή ισοδύναμα, έχουν σταθερές τιμές για όλες τις εγγραφές του dataset (συγκεκριμένα τα εξής: *Bwd PSH Flags, Bwd URG Flags, Fwd Byts/b Avg, Fwd Pkts/b Avg, Fwd Blk Rate Avg, Bwd Byts/b Avg, Bwd Pkts/b Avg, Bwd Blk Rate Avg*). Αυτά τα χαρακτηριστικά δεν προσφέρουν κάτι στην εκπαίδευση, συνεπώς μπορούμε να τα αφαιρέσουμε. 14 εγγραφές έχουν λανθασμένο timestamp (π.χ. 1970-01-10 03:04:26) οι οποίες και αυτές διαγράφονται. Σε δύο εγγραφές που έχουν αρνητικές τιμές στα χαρακτηριστικά *Flow IAT Mean, Fwd IAT Min*, τα οποία έχουν κανονικά τιμές μεγαλύτερες ή ίσες του μηδενός, αντικαθιστούμε τις τιμές αυτών των χαρακτηριστικών με 0. Ακόμα, κάποιες εγγραφές στα χαρακτηριστικά *Init Fwd Win Byts, Init Bwd Win Byts* (και αυτά αυστηρά μη αρνητικά) έχουν την τιμή -1. Παρατηρώντας ότι η μεγαλύτερη τιμή για αυτά τα features στο dataset είναι 65535, δηλαδή 1111111111111111 στο δυαδικό, υποθέτουμε ότι το -1 έχει προκύψει από overflow μεταβλητών που είχαν δηλωθεί ως μη προσημασμένοι ακέραιοι 16 bit (uint16), οι οποίοι έχουν μέγιστη τιμή 65535. Συνεπώς, επιλέγουμε να αντικαταστήσουμε αυτές τις τιμές με 65535 + 1. Τέλος, αφαιρούμε το χαρακτηριστικό *timestamp* το οποίο δεν μας είναι χρήσιμο για την ανάλυση. Στο τέλος της παραπάνω διαδικασίας, το dataset αποτελείται από 16.137.169 εγγραφές και 69 features.

5.2.2 Προεπεξεργασία για την εκπαίδευση των μοντέλων

Για την εκπαίδευση των μοντέλων θα χωρίσουμε τα δεδομένα σε train, validation και test sets με την αρκετά συνήθη αναλογία 60%-20%-20%. Καθώς σε διαφορετικές μέρες της προσομοίωσης υπάρχουν διαφορετικά είδη κακόβουλων πακέτων, ο διαχωρισμός θα γίνει ανά ημέρα, ώστε να υπάρχει πιο δίκαιη εκπροσώπηση των κλάσεων.

Έτσι, για κάθε μέρα, αφού ανακατέψουμε τα δεδομένα εκείνης, κάνουμε train-test

stratified split ώστε να διατηρηθεί η αναλογία εμφάνισης των κλάσεων. Στο τέλος, συγχωνεύονται τα train data κάθε μέρας, και αντίστοιχα τα test. Αυτή η διαδικασία παραμένει ίδια τόσο για την περίπτωση του binary, όσο και του multiclass classfication. Από τα train δεδομένα, αφαιρούνται στην συνέχεια τα αντίστοιχα για το validation. Πριν την εκπαίδευση, κάνουμε min-max scaling στα δεδομένα, ώστε οι τιμές του κάθε χαρακτηριστικού τους να είναι μεταξύ 0 και 1. Σημειώνεται ότι κάνουμε fit τον scaler μόνο στα train δεδομένα, και στην συνέχεια μετατρέπουμε τόσο αυτά όσο και τα test δεδομένα.

Στους πίνακες 5.1 και 5.2 παρουσιάζεται αντίστοιχα η κατανομή των δεδομένων ανά μέρα καθώς και η συνολική κατανομή τους.

Ημέρα	Κλάση	# εγγραφών
2018-02-14	Benign	663.803
	FTP-BruteForce	193.354
	SSH-Bruteforce	187.589
2018-02-15	Benign	988.050
	DoS attacks-GoldenEye	41.508
	DoS attacks-Slowloris	10.990
2018-02-16	Benign	446.772
	DoS attacks-Hulk	461.912
	DoS attacks-SlowHTTPTest	139.890
2018-02-20	Benign	7.313.104
	DDoS attacks-LOIC-HTTP	576.191
2018-02-21	Benign	360.833
	DDOS attack-HOIC	686.012
	DDOS attack-LOIC-UDP	1.730
2018-02-22	Benign	1.042.594
	Brute Force - Web	249
	Brute Force - XSS	79
	SQL Injection	34
2018-02-23	Benign	1.042.301
	Brute Force - Web	362
	Brute Force - XSS	151
	SQL Injection	53
2018-02-28	Benign	538.666
	Infiltration	68.236
2018-03-01	Benign	235.778
	Infiltration	92.403
2018-03-02	Benign	758.334
	Bot	286.191

Πίνακας 5.1: *Κατανομή κλάσεων ανά ημέρα*

Κλάση	Κωδικοποίηση	# εγγραφών
Benign	0	13.390.235
DDOS attack-HOIC	1	686.012
DDoS attacks-LOIC-HTTP	2	576.191
DoS attacks-Hulk	3	461.912
Bot	4	286.191
FTP-BruteForce	5	193.354
SSH-Bruteforce	6	187.589
Infiltration	7	160.639
DoS attacks-SlowHTTPTest	8	139.890
DoS attacks-GoldenEye	9	41.508
DoS attacks-Slowloris	10	10.990
DDOS attack-LOIC-UDP	11	1.730
Brute Force - Web	12	611
Brute Force - XSS	13	630
SQL Injection	14	87

Πίνακας 5.2: *Συνολική κατανομή κλάσεων*

Συνολικά υπάρχουν 15 κλάσεις, οι οποίες κωδικοποιούνται από 0 έως 14. Παρατηρούμε ότι υπάρχει ανισορροπία και οι εγγραφές που αντιστοιχούν στα φυσιολογικά πακέτα είναι αρκετά περισσότερες από αυτές των κακόβουλων, ενώ το πλήθος των δειγμάτων για τις τρεις τελευταίες κλάσεις είναι αμελητέο, συνεπώς δεν θα συμπεριληφθούν στην ανάλυση.

Κεφάλαιο 6

Υλοποίηση

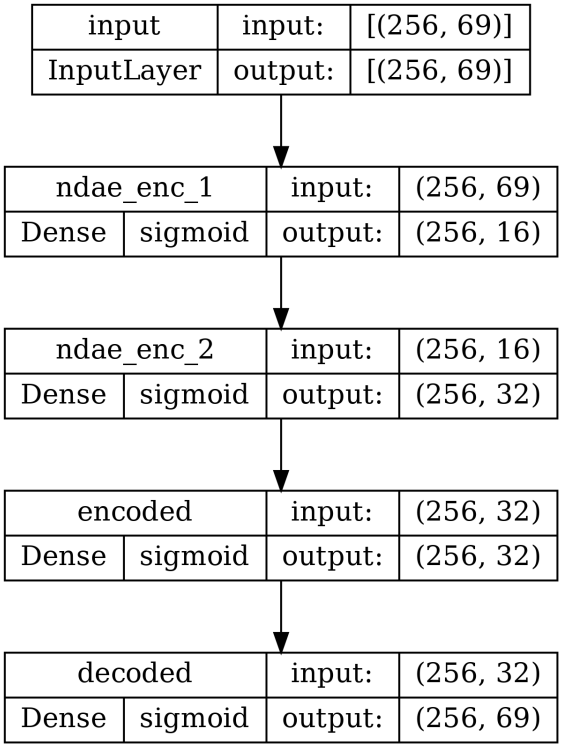
Σε αυτό το κεφάλαιο περιγράφεται η υλοποίηση του IDS και της διάταξης federated learning, οι μετρικές που χρησιμοποιήθηκαν με σκοπό την αξιολόγησή της επίδοσης του μοντέλου, και τα πειράματα που εκτελέστηκαν. Η υλοποίηση πραγματοποιήθηκε με χρήση των TensorFlow [117] και Keras [118] σε Python 3. Για την εκτέλεση πειραμάτων federated learning χρησιμοποιήθηκε η βιβλιοθήκη TensorfLow Federated [119] που προσφέρει την κατάλληλη υποδομή για προσομοίωση κατανεμημένου περιβάλλοντος καθώς και υλοποιήσεις αλγορίθμων όπως ο federated averaging. Για την διαχείριση των δεδομένων (π.χ. batching, shuffling, one-hot encoding, caching, prefetching) αξιοποιήθηκε το API των data pipelines που προσφέρει το TensorFlow. Το σύνολο της υλοποίησης πραγματοποιήθηκε σε υπολογιστή με λειτουργικό σύστημα Ubuntu 20.04.4 LTS 64-bit, επεξεργαστή AMD Ryzen 7 4800H, μνήμη χωρητικότητας 16GB και κάρτα γραφικών NVIDIA GeForce RTX 2060.

6.1 Υλοποίηση υβριδικού μοντέλου

Όπως αναφέρθηκε στο κεφάλαιο 4, το σύστημα ανίχνευσης εισβολής και τα πειράματα που εκτελέστηκαν βασίζονται σε μια αρχιτεκτονική βαθιάς μάθησης που συνδυάζει δύο επιμέρους μοντέλα: έναν non symmetrical deep autonencoder (NDAE) μη επιβλεπόμενης μάθησης και ένα βαθύ feed forward νευρωνικό δίκτυο επιβλεπόμενης μάθησης (θα αναφέρεται απλά ως DNN στο εξής). Αυτό το υβριδικό μοντέλο που αρχικά εκπαιδεύεται χωρίς federated learning θα χρησιμοποιηθεί και ως μέτρο αναφοράς (θα αναφέρεται ως baseline μοντέλο στο εξής), ώστε να μπορούμε στην συνέχεια να συγκρίνουμε την συμπεριφορά του με το μοντέλο (ίδιας αρχιτεκτονικής) το οποίο εκπαιδεύεται με federated learning. Σε όλα τα πειράματα το batch size επιλέγεται να είναι 256, έπειτα από δοκιμές εναλλακτικών τιμών.

6.1.1 Υλοποίηση NDAE

Το NDAE αποτελεί το τμήμα του μοντέλου που εκπαιδεύεται μη επιβλεπόμενα, κάνοντας χρήση μόνο unlabeled (μη επισημασμένων) δεδομένων. Αποτελείται από τέσσερα dense επίπεδα νευρώνων με την έξοδο του προτελευταίου επιπέδου να αποτελεί την κωδικοποιημένη αναπαράσταση των δεδομένων, και η έξοδος του τελευταίου να αποτελεί το reconstruction το οποίο επιθυμούμε να προσομοιάζει τα δεδομένα εισόδου. Το πλήθος των νευρώνων σε



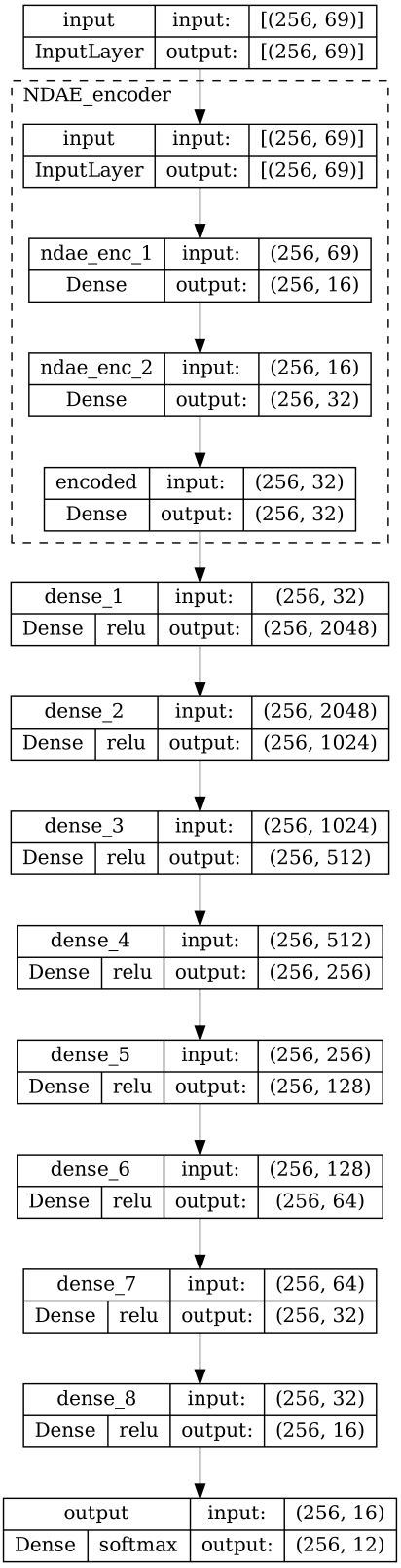
Σχήμα 6.1: *Οι αρχιτεκτονική του NDAE*

κάθε επίπεδο είναι αντίστοιχα $(16, 32, 32, 69)$, παρόμοια με την αρχιτεκτονική στο [79], με την διάσταση εξόδου να ταυτίζεται με την διάσταση εισόδου. Σε όλα τα επίπεδα χρησιμοποιείται σιγμοειδής συνάρτηση ενεργοποίησης, η χρήση της οποίας απαιτείται εξαρχής για το τελευταίο επίπεδο, στα πλαίσια της ανακατασκευής των δεδομένων εισόδου καθώς κατά τη προεπεξεργασία έγινε scaling αυτών στο διάστημα $[0, 1]$. Η συνάρτηση απώλειας που χρησιμοποιείται κατά την εκπαίδευση είναι η binary cross entropy (BCE). Στο σχήμα 6.1 παρουσιάζονται συνοπτικά οι επιλογές για την αρχιτεκτονική του NDAE. Όπως φαίνεται η κωδικοποιημένη μορφή των δεδομένων έχει διαστατικότητα 32, αρκετά μικρότερη από την αρχική (μείωση κατά περίπου 53,6%).

6.1.2 Υλοποίηση feed forward DNN

Το feed forward DNN είναι το τμήμα του μοντέλου που εκπαιδεύεται επιβλεπόμενα και το οποίο πραγματοποιεί ταξινόμηση labeled δεδομένων. Προϋπόθεσή αποτελεί να έχει προηγηθεί η εκπαίδευση του NDAE, καθώς χρησιμοποιούμε το encoding μέρος του για την κωδικοποίηση των δεδομένων εισόδου του DNN. Αποτελείται αρχικά από 8 επίπεδα με αντίστοιχα πλήθη νευρώνων $(2048, 1024, 512, 256, 128, 64, 32, 16)$, αρχιτεκτονική που προέκυψε έπειτα από δοκιμές εναλλακτικών. Έπειτα από αυτά, υπάρχει ένα τελευταίο επίπεδο εξόδου για την τελική ταξινόμηση, όπου το μέγεθος του είναι 2 για την περίπτωση binary classification, ενώ για την περίπτωση του multiclass classification ισούται με το πλήθος των κλάσεων, δηλαδή 12. Για τις ετικέτες των δεδομένων σε κάθε περίπτωση χρησιμοποιείται one-hot encoding. Σε όλα τα επίπεδα χρησιμοποιείται η relu συνάρτηση ενεργοποίησης, με εξαίρεση το επίπεδο εξόδου του δικτύου όπου χρησιμοποιείται softmax. Ως συνάρτηση απώλειας για

την εκπαίδευση του DNN χρησιμοποιείται η categorical cross entropy.



Σχήμα 6.2: *Η αρχιτεκτονική του συνολικού υβριδικού μοντέλου (encoder και DNN).*

Στο σχήμα 6.2 παρουσιάζεται συνοπτικά η αρχιτεκτονική ολόκληρου του υβριδικού μοντέλου για mutliclass classfication (για binary classification η μοναδική διαφορά είναι το

πλήθος νευρώνων στο επίπεδο εξόδου). Σημειώνεται, ότι η ανανέωση των βαρών στο NDAE γίνεται μόνο κατά την εκπαίδευση του (πριν την εκπαίδευση του DNN). Όταν απομονώνουμε από το NDAE τον encoder, τον χρησιμοποιούμε μόνο για την κωδικοποίηση των δεδομένων, χωρίς δηλαδή τα βάρη του να εκπαιδεύονται ξανά.

Στα πλαίσια της υβριδικής μάθησης, χωρίζουμε τα training data με το 75% αυτών να χρησιμοποιηθεί για την μη επιβλεπόμενη εκπαίδευση του NDAE (αγνοώντας δηλαδή πλήρως τις ετικέτες τους), και το υπόλοιπο 25% των δεδομένων χρησιμοποιείται για την επιβλεπόμενη μάθηση του DNN, αφού πρώτα κωδικοποιηθούν από τον εκπαιδευμένο πλέον encoder του NDAE. Η συγκεκριμένη δυσαναλογία στον διαχωρισμό τον δεδομένων έχει σκοπό να προσομοιώσει την ανισορροπία μεταξύ unlabel labeled που παρατηρείται στην πράξη, ωστόσο διαφορετικές αναλογίες μπορούν να δοκιμαστούν για βελτιστοποίηση της επίδοσης του μοντέλου.

6.2 Υλοποίηση federated learning

Πριν από την έναρξη της συνεργατικής εκπαίδευσης, μοιράζουμε τα training δεδομένα στους clients, βάσει κάποιας κατανομής για το εκάστοτε πείραμα (περισσότερα στην ενότητα 6.4). Αναφερόμαστε στα δεδομένα εκπαίδευσης του κάθε client ως *τοπικά δεδομένα*, τα οποία καθ' όλη την διάρκεια του federated learning δεν ανταλλάσσονται ούτε μεταξύ των clients, ούτε με τον server. Τόσο η εκπαίδευση του NDAE, όσο και του DNN γίνεται μόνο με federated learning, με την μοναδική προϋπόθεσή να είναι ότι πρέπει να έχει ολοκληρωθεί πρώτα η εκπαίδευση του NDAE, ώστε να αξιοποιηθεί ο encoder του για την κωδικοποίηση των δεδομένων που χρησιμοποιούνται από το DNN.

Σε αντίθεση με την εκπαίδευση των NDAE και DNN στο baseline μοντέλο, που απαιτεί μόνο τον ορισμό των συνήθων υπερπαραμέτρων, για την υλοποίηση του federated learning χρειάζεται επιπλέον ο ορισμός συγκεκριμένων επιλογών και μεγεθών που καθορίζουν την πορεία της εκπαίδευσης.

Το πρώτο από αυτά είναι το πλήθος των clients που συνεργάζονται για την εκπαίδευση του κοινού μοντέλου. Για τα πειράματα τις παρούσας εργασίας, στα πλαίσια προσομοίωσης cross-silo federated learning, επιλέγεται να συμμετέχουν σε κάθε γύρο εκπαίδευσης 4 σταθεροί clients οι οποίοι είναι αξιόπιστα διαθέσιμοι. Ακόμα, σε κάθε γύρο, οι clients εκπαιδεύουν τα τοπικά τους μοντέλα για 3 epochs με τον εκάστοτε client να χρησιμοποιεί μόνο τα τοπικά του δεδομένα, τα οποία ανακατεύονται πριν από κάθε epoch. Ο αριθμός των clients και των εποχών τοπικής εκπαίδευσης επιλέχθηκαν έπειτα από περιορισμένες δοκιμές, με την εύρεση της βέλτιστης τιμής τους να παραμένει αντικείμενο μελλοντικής εργασίας. Με την ολοκλήρωση της εκπαίδευσής των τοπικών μοντέλων στους clients για τον τρέχοντα γύρο, γίνεται αποστολή των ανανεώσεων των βαρών στον server που θα πραγματοποιήσει το aggregation και θα ανανεώσει το συνολικό μοντέλο. Η παραπάνω διαδικασία αντικατοπτρίζει τον τυπικό γύρο του federated averaging.

Με το πέρας κάθε γύρου, και την ανανέωση του κοινού συνολικού μοντέλου στον server, χρησιμοποιούμε τα validation δεδομένα για την αξιολόγηση του συνολικού μοντέλου μέχρι εκείνη την στιγμή, και την καταγραφή της πορείας της ομοσπονδιακής μάθησης. Όταν ολοκληρωθεί η διαδικασία του federated learning, το τελικό κοινό μοντέλο αξιολογείται βάσει

των test data. Η αξιολόγηση μπορεί να γίνει είτε κεντρικά στον server, είτε κατανεμημένα αφού πρώτα τα test data φτάσουν στους clients και στην συνέχεια συμψηφίσουμε τα επιμέρους αποτελέσματα της αξιολόγησης του κάθε client. Σε κάθε περίπτωση το τελικό κοινό μοντέλο διατίθεται σε όλους τους clients για μελλοντική τοπική τους χρήση.

6.3 Περιγραφή πειραμάτων

Σε αυτήν την ενότητα περιγράφονται αναλυτικά τα πειράματα που εκτελέστηκαν για την αξιολόγηση του συστήματος ανίχνευσης εισβολής όταν εκπαιδεύεται με federated learning, καθώς και την συμπεριφορά του όταν τα δεδομένα δεν ικανοποιούν την i.i.d. υπόθεση.

6.3.1 Baseline μοντέλο

Αρχικά, ως μέτρο σύγκρισης, εκπαιδεύεται το μοντέλο χωρίς federated learning (baseline), ώστε να έχουμε ξεκάθαρη εικόνα της επίδοσής του, συγκριτικά με το αντίστοιχο μοντέλο federated learning. Για τον σκοπό αυτό, εκπαιδεύεται πρώτα ο NDAE χρησιμοποιώντας τα unlabeled δεδομένα. Ως optimizer επιλέγεται ο Adam [120] με learning rate 0.01. Για την αποφυγή overfitting χρησιμοποιούμε early stopping με κριτήριο το validation loss, όπως αυτό προκύπτει στο τέλος του κάθε epoch, ορίζοντας ως παράμετρο ελάχιστης επιθυμητής μεταβολής $\delta = 0.00001$ και patience 5 epochs (δηλαδή τον αριθμό των epochs στον οποίο δεν παρατηρείται θετική μεταβολή μεγαλύτερη του δ πριν διακοπεί πρόωρα η εκπαίδευση). Μετά την εφαρμογή του early stopping, επαναφέρουμε τα βάρη με τα οποία επιτεύχθηκε η καλύτερη επίδοση βάσει του validation loss.

Έχοντας εκπαιδεύσει τον NDAE, ¨παγώνουμε¨ τα βάρη του, και εξάγουμε από αυτόν τον encoder, δηλαδή τα επίπεδα που είναι υπεύθυνα για το encoding. Ο encoder αυτός χρησιμοποιείται για την κωδικοποίηση των δεδομένων πριν την είσοδο τους στο DNN. Για την εκπαίδευση του DNN με την χρήση των κωδικοποιημένων δεδομένων εισόδου, χρησιμοποιείται ως optimizer ο Adam με learning rate 0.0001. Όπως και στην εκπαίδευση του NDAE, γίνεται χρήση early stopping βάσει του validation loss, patience 5 εποχών και αυτή την φορά $\delta = 0.0001$. Η επιλογή του δ έγινε αφού εκτελέστηκαν δοκιμές ώστε να φανεί σε ποια τιμή συγκλίνει το validation loss. Όπως και στην περίπτωση εκπαίδευσης του NDAE στο τελικό μοντέλο, μετά το early stopping επαναφέρουμε τα βάρη με τα οποία επιτεύχθηκε η καλύτερη επίδοση βάσει του validation loss. Η παραπάνω διαδικασία εκπαίδευσης του baseline υβριδικού μοντέλου εκτελέστηκε για την περίπτωση binary classification και για αυτή του multiclass classification ξεχωριστά.

6.3.2 Federated learning μοντέλο

Στα πλαίσια της σύγκρισης της συμπεριφοράς και επίδοσης του μοντέλου σε περιβάλλον cross-silo federated learning, θα χρησιμοποιήσουμε την ίδια ακριβώς αρχιτεκτονική για το NDAE και το DNN, εκπαιδεύοντας τα αυτή την φορά με την χρήση του αλγορίθμου federated averaging, όπως περιγράφηκε στην ενότητα 6.2. Συνοπτικά, υπενθυμίζουμε ότι χρησιμοποιούνται 4 clients με έναν server για το aggregation, ενώ τόσο το μοντέλο NDAE όσο και το DNN εκπαιδεύονται με χρήση federated learning. Ο αριθμός των εποχών τοπικής

εκπαίδευσης σε κάθε γύρο είναι σταθερός και ίσος με 3. Σχετικά με την τελική αξιολόγηση του συνολικού μοντέλου που προκύπτει από το federated learning, παρατηρήθηκε ότι το αποτέλεσμα παραμένει όμοιο, είτε αυτή γίνει κεντρικά, είτε ως μέσος όρος των επιδόσεων των επιμέρους clients. Σε κάθε περίπτωση η τελική αξιολόγηση γίνεται με την χρήση των test data τα οποία δεν συμμετείχαν στην εκπαίδευση.

Για να γίνει η σύγκριση του baseline και του federated μοντέλου επί ίσοις όροις, επιλέγουμε οι υπερπαράμετροι (π.χ. learning rate, optimizer) που επιλέχθηκαν για την εκπαίδευση των NDAE και DNN στο baseline μοντέλο, να παραμείνουν ίδιοι και σταθεροί στην εκπαίδευση των τοπικών μοντέλων στους clients σε κάθε γύρο του federated learning. Η διαφορά στο federated averaging είναι ότι στο τέλος του κάθε γύρου, κατά το aggregation που πραγματοποιείται στον server, χρησιμοποιούμε SGD για την ανανέωση των βαρών του κοινού μοντέλο, με learning rate 1.0 για την εκπαίδευση του NDAE και 0.5 για την εκπαίδευση του DNN.

Μια ακόμη διαφορά μεταξύ της εκπαίδευσης baseline μοντέλου, και του μοντέλου που εκπαιδεύεται με federated learning, είναι η υλοποίηση του early stopping. Καθως στην περίπτωση του federated learning, σε κάθε γύρο, το κοινό συνολικό μοντέλο προκύπτει μετά το aggregation και την ανανέωση του στον server, χρησιμοποιούμε τότε τα validation data για τον υπολογισμό των μετρικών και του validation loss. Συνεπώς, η εφαρμογή early stopping γίνεται στο επίπεδο των γύρων, και όχι των εποχών όπως στην περίπτωση του baseline μοντέλου. Οι παράμετροι συμπεριφοράς του early stopping παραμένουν ίδιοι με τους αντίστοιχους της εκπαίδευσης του baseline μοντέλου χωρίς federated learning: η μετρική αξιολόγισης είναι το validation loss, το δ παραμένει ίδιο για την εκπαίδευση του NDAE και του DNN (0.00001 και 0.0001 αντίστοιχα), ενώ το patience (που παραμένει 5) αναφέρεται, στην περίπτωση του federated learning, στον αριθμό των γύρων που δεν παρατηρείται βελτίωση μεγαλύτερη του δ στο validation loss, όπως αυτό προκύπτει από το ανανεωμένο κοινό μοντέλο.

Πέρα από την σύγκριση του baseline υβριδικού μοντέλου (δηλαδή χωρίς την χρήση federated learning), και του ίδιου μοντέλου όταν εκπαιδεύεται με federated learning σε binary και multiclass classification, επιθυμούμε να εξετάσουμε την συμπεριφορά και την επίδοση του μοντέλου federated learning, όταν δεν ισχύει η υπόθεση i.i.d. για τα δεδομένα. Έως τώρα, στα πειράματα federated learning που εκτελέστηκαν, τα δεδομένα κατανέμονταν στους clients τυχαία και ομοιόμορφα. Θα εξετάσουμε δύο σενάρια στα πλαίσια του multiclass classification με federated learning στα οποία επιβάλουμε συγκεκριμένη κατανομή των δεδομένων στους clients με το κάθε σενάριο να εξετάζει διαφορετικά χαρακτηριστικά.

Για το πρώτο σενάριο (θα αναφέρεται ως σενάριο 1 στο εξής), κατανέμουμε τα δεδομένα εκπαίδευσης της κάθε κλάσης κακόβουλης δικτυακής κίνησης στους clients με αναλογίες *(0.7, 0.1, 0.1, 0.1)* εκ περιτροπής (round robin). Δηλαδή για μια συγκεκριμένη κλάση i ο πρώτος client θα έχει το 70% δειγμάτων που την εκπροσωπούν, ενώ οι υπόλοιποι τρεις θα έχουν από 10%. Για την επόμενη κλάση $i + 1$, ο δεύτερος client θα έχει το 70% δειγμάτων που την εκπροσωπούν ενώ οι υπόλοιποι τρεις θα έχουν από 10% και ούτω καθεξής για όλες τις κλάσεις επιθέσεων. Για την κλάση των πακέτων φυσιολογικής δικτυακής κίνησης, τα δείγματα μοιράζονται τυχαία και ομοιόμορφα στους clients, με το σκεπτικό ότι και στην πράξη ο κάθε client που συμμετέχει θα διαθέτει πληθώρα δεδομένων φυσιολογικής λειτουρ-

γίας. Έτσι, κάθε client διαθέτει δείγματα από κάθε κλάση κακόβουλων δεδομένων αν και υπάρχει σημαντική ανισορροπία στο πλήθος τους για καθεμία από αυτές, μεταξύ των clients.

Για το δεύτερο σενάριο (θα αναφέρεται ως σενάριο 2 στο εξής) ακολουθείται διαφορετική προσέγγιση, όπου εξετάζουμε την συμπεριφορά του federated learning μοντέλου και της εκπαίδευσής του όταν σε κάποιον client απουσιάζουν πλήρως δεδομένα κάποιων κακόβουλων κλάσεων. Πιο συγκεκριμένα, για μια κλάση i, δεν διατίθενται καθόλου δεδομένα που την εκπροσωπούν στον πρώτο client, ενώ στους υπόλοιπους τρεις μοιράζονται τυχαία με αναλογίες *(0.6, 0.2, 0.2)*. Η διαδικασία αυτή πραγματοποιείται και πάλι με round robin τρόπο, δηλαδή για την επόμενη κακόβουλη κλάση $i + 1$ θα απουσιάζουν πλήρως δεδομένα που την εκπροσωπούν από τον δεύτερο client και θα μοιραστούν στους εναπομείναντες τρεις clients, και ούτω καθεξής. Όπως και στο σενάριο 1, τα δεδομένα που προέρχονται από φυσιολογική δικτυακή κίνηση μοιράζονται τυχαία και ομοιόμορφα στους clients.

Για τα σενάρια 1 και 2 πέρα από τις διαφορές στην επίδοση του κοινού τελικού μοντέλου που προκύπτει από το federated learning, μας ενδιαφέρει και η συμπεριφορά εκπαίδευσης που θα παρουσίαζε ο κάθε client στην περίπτωση που εκπαίδευε μόνος του το baseline μοντέλο χρησιμοποιώντας μόνο τα τοπικά δεδομένα του με την προβληματική κατανομή. Έτσι, για καθένα από τα δύο σενάρια, εκπαιδεύουμε τους clients μεμονωμένα, και στην συνέχεια θα συγκρίνουμε την επίδοση των μοντέλων που προέκυψαν, σε σχέση με τον κοινό συνεργατικό federated learning μοντέλο.

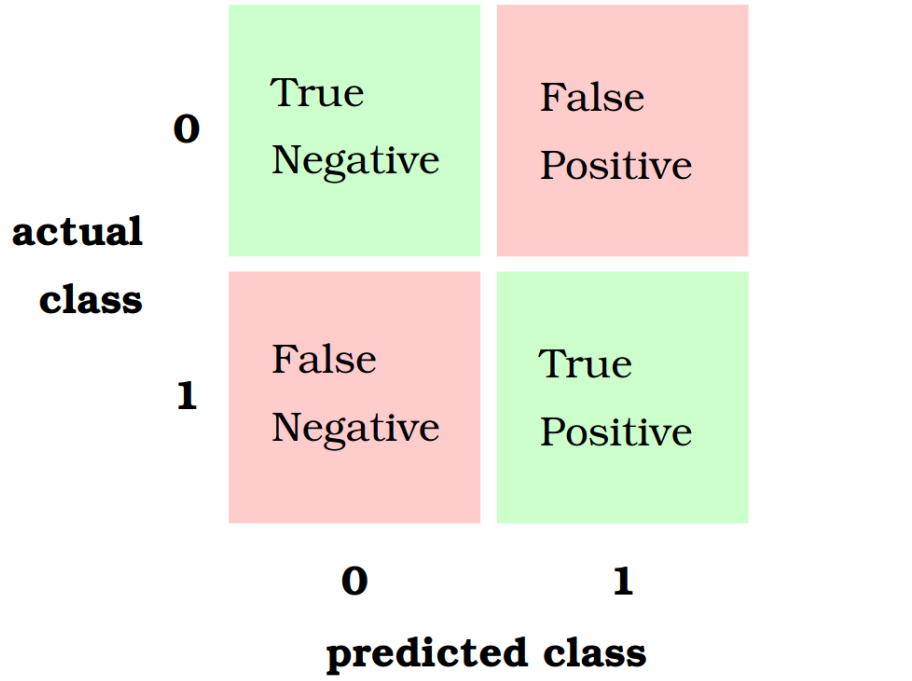
Παρακάτω συνοψίζονται τα πειράματα που εκτελέστηκαν με το υβριδικό μοντέλου που περιγράφηκε και στα οποία βασίζονται τα συμπεράσματα που θα προκύψουν:

- Binary & multiclass classification χωρίς federated learning (baseline).
- Binary & multiclass classification σε cross-silo federated learning περιβάλλον.
- Multiclass classification federated learning, με non-i.i.d. και imbalanced δεδομένα εκπαίδευσης, επιβαλλόμενης κατανομής στους clients:
  - Σενάριο 1: Αναλογίες κλάσεων κακόβουλης δικτυακής κίνησης $(0.7, 0.1, 0.1, 0.1)$ (με round robin διαμοιρασμό).
  - Σενάριο 2: Αναλογίες κλάσεων κακόβουλης δικτυακής κίνησης $(0.0, 0.6, 0.2, 0.2)$, με απουσία δεδομένων που εκπροσωπούν την εκάστοτε κακόβουλη κλάση σε έναν client (με round robin διαμοιρασμό).

6.4 Μετρικές αξιολόγησης

Σε αυτή την ενότητα αναφέρονται οι ποσοτικές μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου σε όλα τα πειράματα, πέραν δηλαδή από παρατηρήσεις και συμπεράσματα που προκύπτουν από την εξέταση της διαδικασίας εκπαίδευσης. Τόσο για την περίπτωση του binary, όσο και για την περίπτωση του multiclass classification χρησιμοποιούνται οι ίδιες μετρικές, και συγκεκριμένα οι διαδεδομένες για προβλήματα ταξινόμησης: Accuracy, Precision, Recall, F_1 score.

Χρήσιμη για τον ακριβή ορισμό της κάθε μετρικής αλλά και για την οπτικοποίηση της επίδοσης του μοντέλου είναι η έννοια του *confusion matrix (πίνακας σύγχυσης)*. Στην περίπτωση αξιολόγησης ενός ταξινομητή, στο confusion matrix φαίνεται παραστατικά το πλήθος (ή ισοδύναμα το ποσοστό) των δεδομένων που ταξινομήθηκαν σωστά για την κάθε κλάση, το πλήθος αυτών που ταξινομήθηκαν λανθασμένα, καθώς και την κλάση που προέβλεψε ο ταξινομητής για τα δεδομένα που ταξινομήθηκαν λανθασμένα. Στο σχήμα 6.3 φαίνεται ένα απλό confusion matrix για binary classification με κλάσεις 0 και 1 (όμοια επεκτείνεται η μορφή του και στην περίπτωση multiclass classification):



Σχήμα 6.3: *Δομή confusion matrix για 2 κλάσεις*

Κατά σύμβαση, μια από τις δύο κλάσεις θεωρείται θετική και η άλλη αρνητική. Με παρόμοιο τρόπο στην περίπτωση πολλαπλών κλάσεων, κάθε μια από αυτές μπορεί να θεωρηθεί θετική, με τις υπόλοιπες αρνητικές. Ανάλογα με ποια κλάση θεωρηθεί θετική, ορίζονται τα μεγέθη True Positive (TP), True Negative (TN), False Positive (FP) και False Negative (FN), με τα πρώτα δύο να είναι ο αριθμός των σωστά ταξινομημένων δεδομένων ανά κλάση (πράσινο χρώμα στο παράδειγμα), και αντίστοιχα τα τελευταία δύο να είναι ο αριθμός των λανθασμένα ταξινομημένων δεδομένων (κόκκινο χρώμα στο παράδειγμα). Με βάση αυτά τα μεγέθη ορίζονται οι εξής τέσσερις μετρικές που θα χρησιμοποιήσουμε:

Accuracy

Το accuracy (ορθότητα) του μοντέλου, ορίζεται ως ο λόγος του πλήθους των ορθά ταξινομημένων δεδομένων ανά το συνολικό πλήθος των δεδομένων, δηλαδή:

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

Precision

Το precision (ακρίβεια) του μοντέλου ορίζεται από τον παρακάτω λόγο:

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

και εκφράζει το ποσοστό των δεδομένων τα οποία το μοντέλο ταξινόμησε ορθώς ως θετικά, σε σχέση με το πλήθος όλων τον δεδομένων που ταξινομήθηκαν ως θετικά.

Recall

Το recall (ευαισθησία) του μοντέλου ορίζεται από τον παρακάτω λόγο:

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

και εκφράζει το ποσοστό των δεδομένων τα οποία το μοντέλο ταξινόμησε ορθώς ως θετικά, σε σχέση με το πλήθος όλων τον δεδομένων που στην πραγματικότητα ανήκουν στην θετική κλάση.

F_1 score

Το F_1 score αποτελεί τον αρμονικό μέσο των precision και recall συνδυάζοντας τις δύο αυτές μετρικές, δίνοντάς τους την ίδια βαρύτητα, και ορίζεται από τον τύπο:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Averaging μετρικών

Για τις τρεις μετρικές που χαρακτηρίζουν την συμπεριφορά του μοντέλου ως προς μια από τις κλάσεις, δηλαδή precision, recall και F_1 score, μπορούμε να υπολογίσουμε τους μέσους όρους τους ώστε να συνοψίσουμε την επίδοση του μοντέλου. Οι δύο βασικοί τρόποι που θα χρησιμοποιήσουμε είναι το **macro** και το **weighted** averaging. Το macro averaging υπολογίζει τον μέσο όρο των τιμών της μετρικής για την κάθε κλάση, ενώ το weighted averaging υπολογίζει τον *σταθμισμένο* μέσο όρο των τιμών της μετρικής για την κάθε κλάση, χρησιμοποιώντας ως συντελεστές βαρύτητας την συχνότητα εμφάνισης της στα δεδομένα αξιολόγησης. Συνεπώς το weighted averaging επηρεάζεται στην περίπτωση που υπάρχει ανισορροπία στα δεδομένα, σε αντίθεση με το macro averaging.

Κεφάλαιο 7

Παρουσίαση αποτελεσμάτων

Σε αυτό το κεφάλαιο παρουσιάζουμε γραφικά τα αποτελέσματα των πειραμάτων, με την σειρά παρουσίασης τους να ακολουθεί την σειρά με την οποία παρουσιάστηκαν τα πειράματα στην ενότητα 6.3. Η αξιολόγηση του μοντέλου, η σύγκριση της επίδοσής του και η συζήτηση σχετικά με την συμπεριφορά που επιδεικνύει στα διαφορετικά πειράματα θα πραγματοποιηθεί στο επόμενο κεφάλαιο.

7.1 Baseline μοντέλο

Το baseline μοντέλο, δηλαδή αυτό που εκπαιδεύεται χωρίς federated learning, θα αξιοποιηθεί τόσο με στόχο την αναγνώριση εισβολής, όσο και με στόχο την κατηγοριοποίηση της, δηλαδή για binary και για multiclass classification αντίστοιχα.

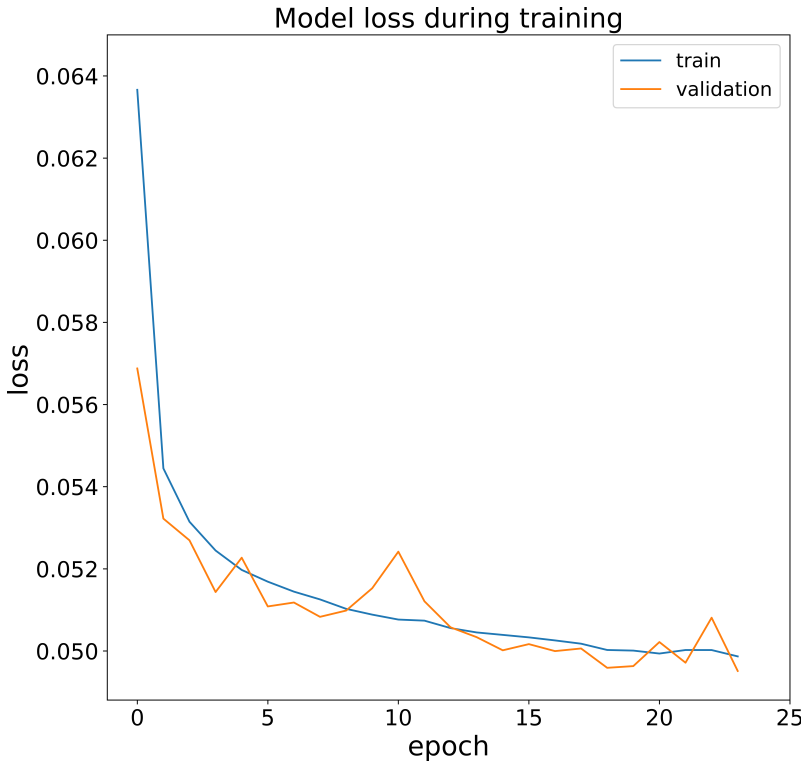
7.1.1 Binary classification

Η πορεία εκπαίδευσης του συνολικού baseline μοντέλου, συνοψίζεται από την γραφική παράσταση των τιμών του training και validation loss ανά εποχή, όπως φαίνεται στο σχήμα 7.1. Παρατηρούμε ότι το συγκεκριμένο μοντέλο εκπαιδεύεται ομαλά, με το validation loss να ακολουθεί φθίνουσα πορεία μέχρι τους τελευταίους γύρους και να λαμβάνει παραπλήσιες τιμές με το training loss.

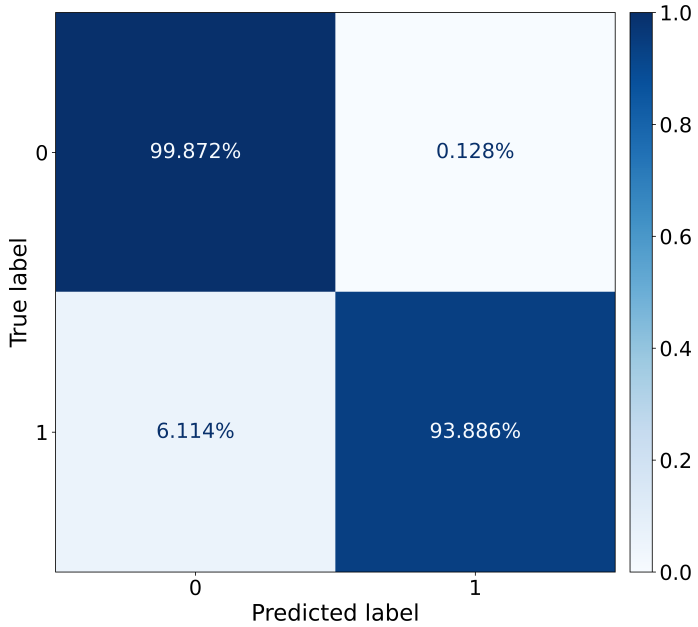
Μετά το πέρας της εκπαίδευσης, η αξιολόγηση του γίνεται βάσει των test data. Οι τιμές των μετρικών αξιολόγησης παρουσιάζονται αναλυτικά στον πίνακα 7.1, ενώ στο σχήμα 7.2 φαίνεται το αντίστοιχο confusion matrix.

Μετρική	Τιμή
Accuracy	98.853%
Macro F_1	0.9792
Macro Precision	0.9905
Macro Recall	0.9688
Weighted F_1	0.9884
Weighted Precision	0.9886
Weighted Recall	0.9885

Πίνακας 7.1: *Τιμές μετρικών αξιολόγησης για το baseline binary μοντέλο*



Σχήμα 7.1: *Training και validation loss κατά την εκπαίδευση του baseline binary μοντέλου*

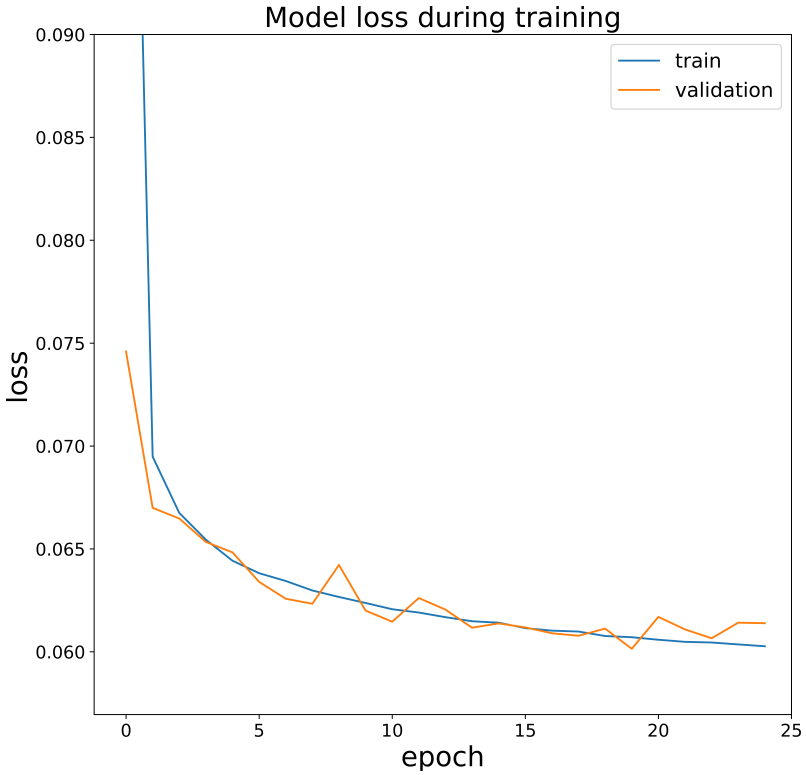


Σχήμα 7.2: *Confusion matrix του baseline binary μοντέλου βάσει των test data*

7.1.2 Multiclass classification

Για την baseline multiclass περίπτωση η αρχιτεκτονική του baseline μοντέλου παραμένει ίδια, με εξαίρεση το output layer του DNN στο οποίο επιλέγουμε 12 νευρώνες, ίσο αριθμό με το πλήθος των κλάσεων. Στον πίνακα 7.2 παραθέτονται οι τιμές των μετρικών αξιολόγησης στα test data, ενώ όπως και στην περίπτωση του binary baseline μοντέλο, στο σχήμα 7.3

φαίνεται η μεταβολή των training και validation loss κατά την εκπαίδευση του μοντέλου.

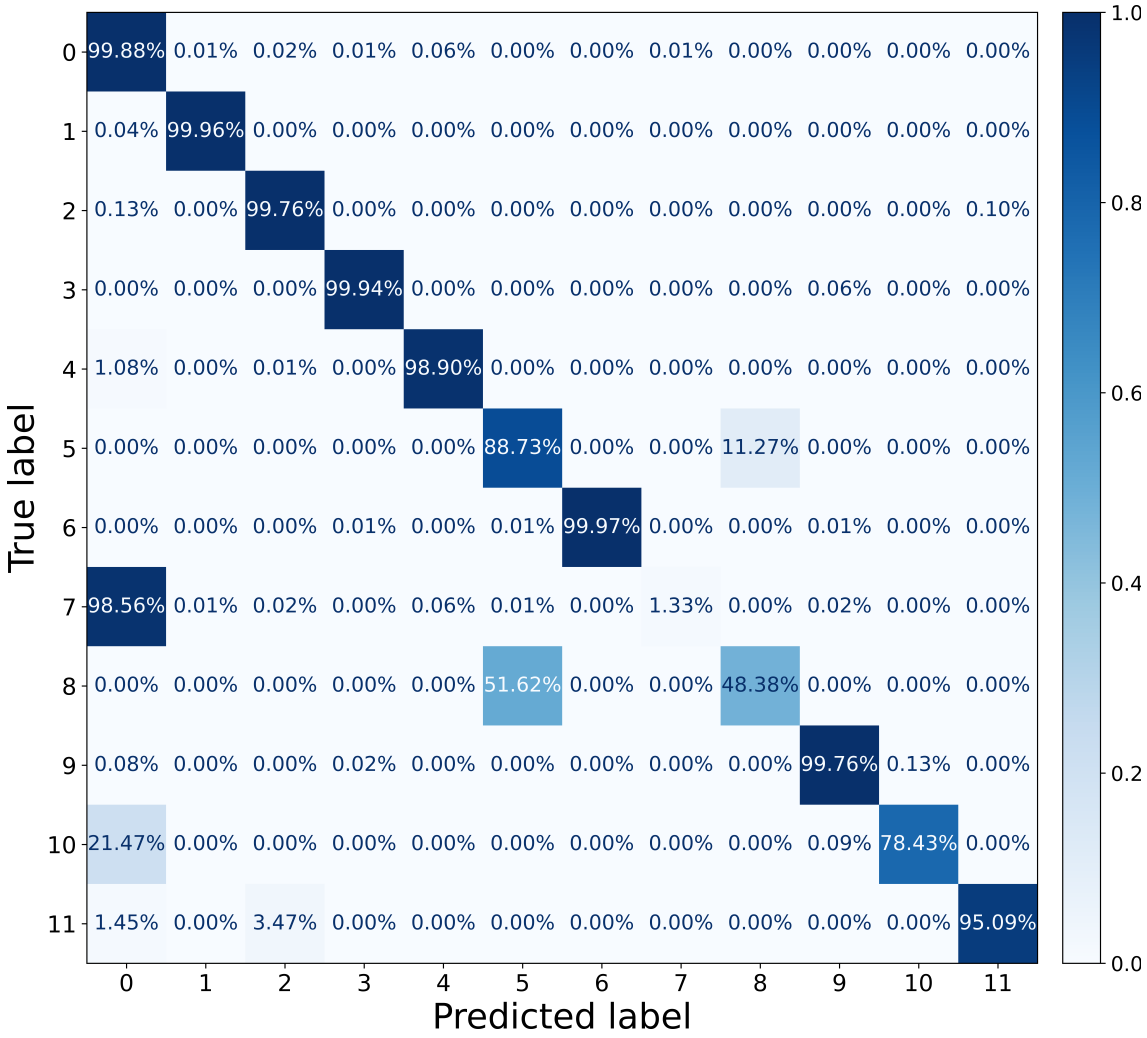


Σχήμα 7.3: *Training και validation loss κατά την εκπαίδευση του baseline multiclass μοντέλου*

Μετρική	Τιμή
Accuracy	98.286%
Macro F_1	0.8383
Macro Precision	0.8860
Macro Recall	0.8418
Weighted F_1	0.9779
Weighted Precision	0.9785
Weighted Recall	0.9829

Πίνακας 7.2: *Τιμές μετρικών αξιολόγησης για το baseline multiclass μοντέλο*

Για να αντιληφθούμε καλύτερα την επίδοση του μοντέλου ως προς την καθεμία κλάση, σχεδιάζουμε το confusion matrix βάσει των test data στο σχήμα 7.4. Φαίνεται ότι για την πλειοψηφία των κλάσεων, το baseline μοντέλο έχει καλή επίδοση, με δύο εξαιρέσεις στις οποίες η συμπεριφορά του διαφέρει. Για την κλάση 8, παρατηρούμε πως περίπου τα μισά δεδομένα της ταξινομούνται εσφαλμένα ως επιθέσεις τις κλάσης 5, ενώ το μοντέλο αποτυγχάνει να εντοπίσει την κακόβουλη κίνηση της κλάσης 7 (infiltration), ταξινομώντας την ως φυσιολογική. Θετικό είναι το γεγονός ότι το μοντέλο πετυχαίνει και στην binary και στην multiclass περίπτωση πολύ χαμηλό ποσοστό ψευδώς θετικών αποτελεσμάτων, δηλαδή πακέτων που ταξινόμησε ως επιθέσεις ενώ ήταν φυσιολογικά.



Σχήμα 7.4: *Confusion matrix του baseline multiclass μοντέλου βάσει των test data*

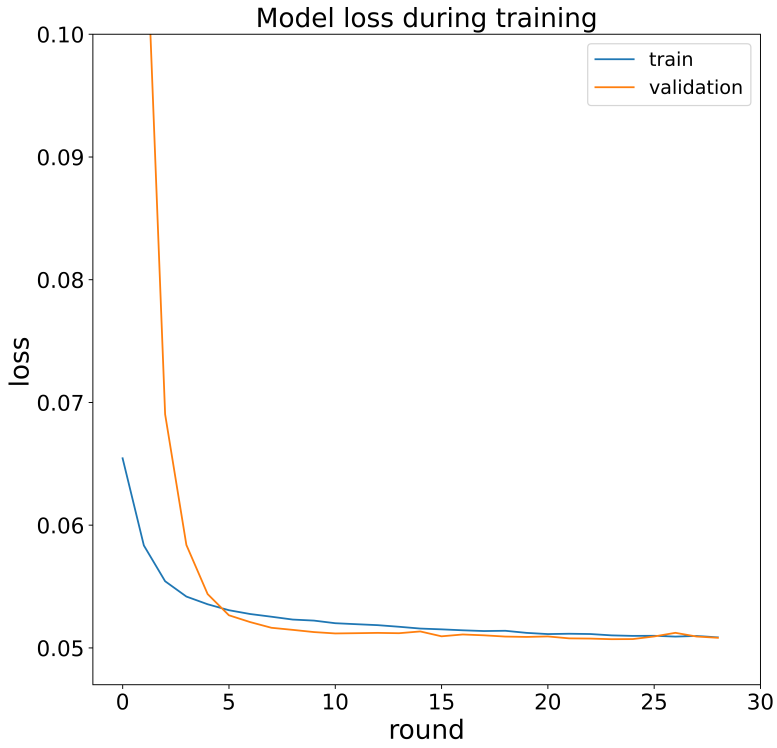
7.2 Federated learning μοντέλο

Συνεχίζοντας, παρουσιάζονται τα αποτελέσματα των πειραμάτων που αφορούν το binary και το multiclass classification, στην περίπτωση που το υβριδικό μοντέλο (δηλαδή τόσο το NDAE, όσο και το DNN), εκπαιδευτεί με την χρήση federated averaging, σύμφωνα με την αρχιτεκτονική και την περιγραφή των πειραμάτων που παρουσιάστηκε στις ενότητες 6.2 και 6.3 αντίστοιχα. Υπενθυμίζουμε ότι στο federated learning το validation loss υπολογίζεται στο τέλος κάθε γύρου, χρησιμοποιώντας το ανανεωμένο κοινό μοντέλο όπως προέκυψε μετά το aggregation, ενώ το training loss ως μέσος όρος των επιμέρους τιμών του κατά την εκπαίδευση των clients.

7.2.1 Binary classification

Στο σχήμα 7.5 παρουσιάζεται η πορεία της εκπαίδευσης ως συνάρτηση των training και validation loss ανά γύρο του federated averaging. Όπως φαίνεται, το validation loss συγκλίνει πολύ ομαλά κοντά στην βέλτιστη τιμή του, ακολουθώντας την αντίστοιχη τιμή του training loss, ενώ σχετική σταθεροποίηση του validation loss παρατηρείται μόλις από τους

πρώτους πέντε γύρους της συνεργατικής μάθησης.



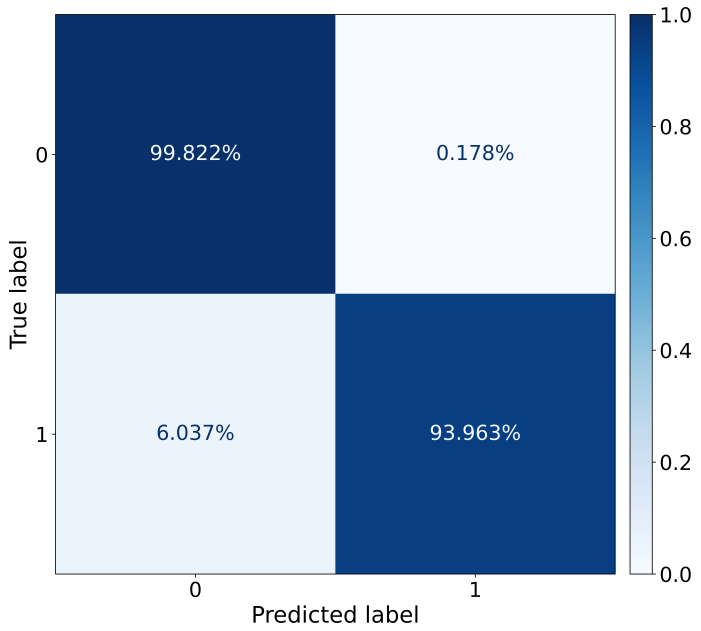
Σχήμα 7.5: *Training και validation loss κατά την εκπαίδευση του feaderated learning binary μοντέλου*

Έχοντας εκπαιδεύσει συνεργατικά το μοντέλο, το αξιολογούμε βάσει των μετρικών, οι τιμές των οποίων φαίνονται στον πίνακα 7.3.

Μετρική	Τιμή
Accuracy	98.824%
Macro F_1	0.9788
Macro Precision	0.9893
Macro Recall	0.9689
Weighted F_1	0.9881
Weighted Precision	0.9883
Weighted Recall	0.9882

Πίνακας 7.3: *Τιμές μετρικών αξιολόγησης για το federated learning binary μοντέλο*

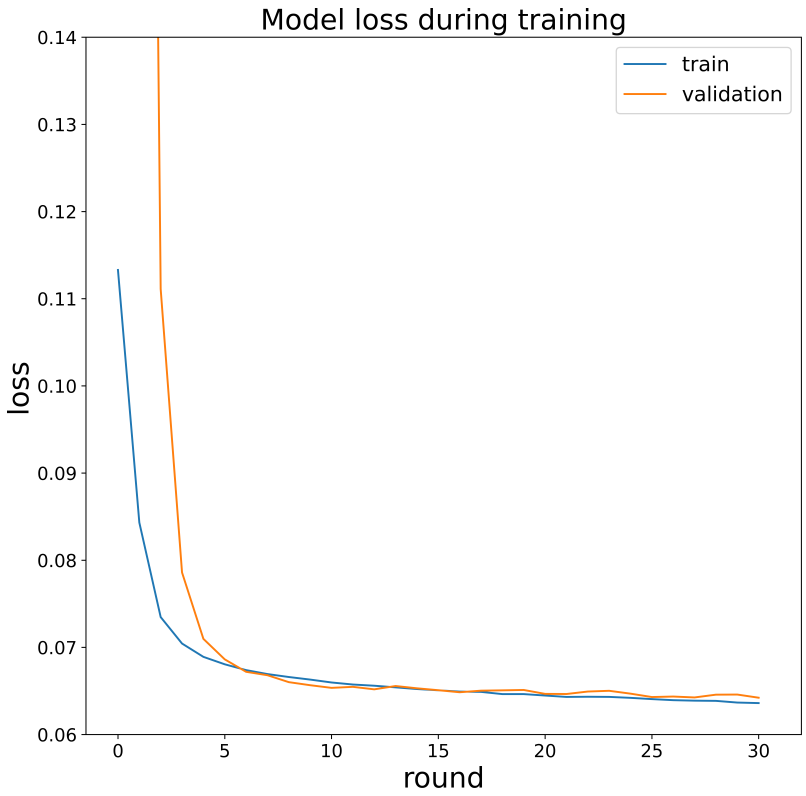
Όμοια με την περίπτωση του baseline binary μοντέλου, παραθέτουμε και το confusion matrix στο σχήμα 7.6 όπως υπολογίστηκε στα test data μετά το πέρας της federated εκπαίδευσης. Παρατηρούμε από το confusion matrix ότι το federated learning μοντέλο συμπεριφέρεται σχεδόν ίδια με το baseline μοντέλο στην binary ταξινόμηση.



Σχήμα 7.6: *Confusion matrix του federated learning binary μοντέλου βάσει των test data*

7.2.2 Multiclass classification

Αλλάζοντας μόνο το πλήθος των νευρώνων του output layer στο DNN (ώστε να ισούται με το πλήθος των κλάσεων), χρησιμοποιούμε federated learning για την εκπαίδευση του multiclass μοντέλου, η πορεία του οποίου φαίνεται στο σχήμα 7.7.

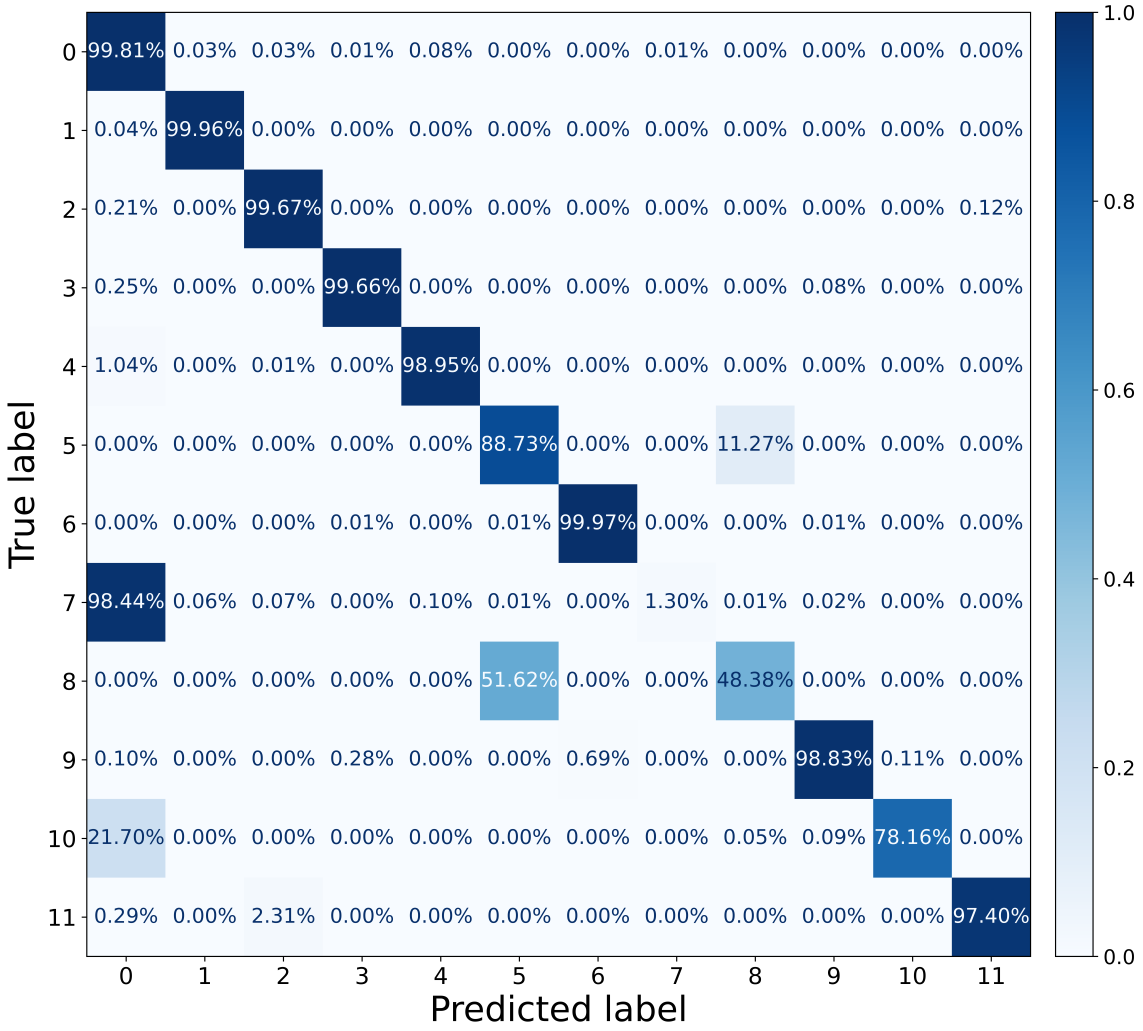


Σχήμα 7.7: *Training και validation loss κατά την εκπαίδευση του federated learning multiclass μοντέλου*

Οι τιμές των μετρικών αξιολόγησης με βάση τα test data παρουσιάζονται στον πίνακα 7.4, ενώ στην συνέχεια στο σχήμα 7.8 φαίνεται το αντίστοιχο confusion matrix.

Μετρική	Τιμή
Accuracy	98.221%
Macro F_1	0.8358
Macro Precision	0.8810
Macro Recall	0.8423
Weighted F_1	0.9772
Weighted Precision	0.9778
Weighted Recall	0.9822

Πίνακας 7.4: *Τιμές μετρικών αξιολόγησης για το federated learning multiclass μοντέλο*



Σχήμα 7.8: *Confusion matrix του federated learning multiclass μοντέλου βάσει των test data*

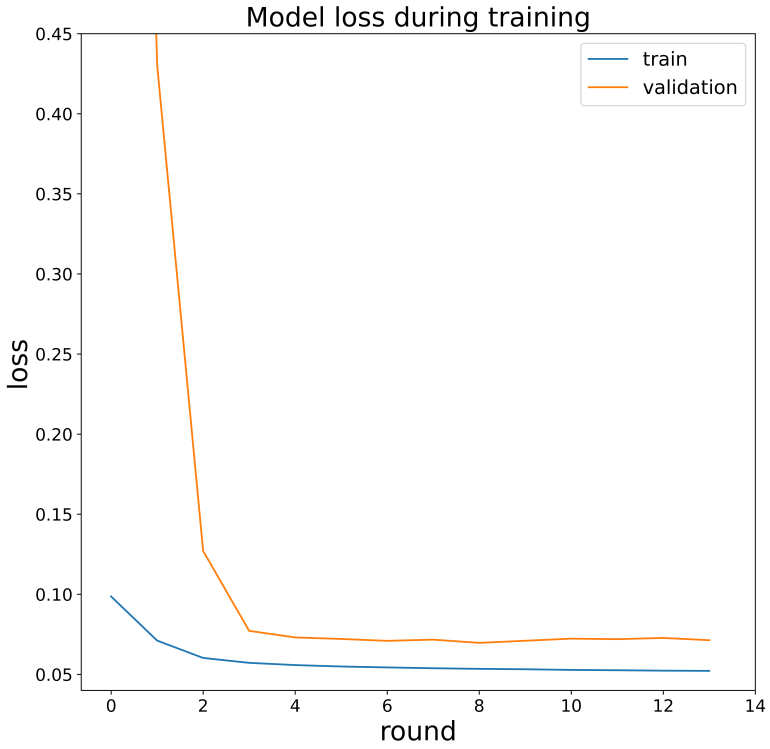
Και πάλι φαίνεται πως το συνεργατικό μοντέλο συγκλίνει ομαλά και γρήγορα, ενώ από το confusion matrix βλέπουμε ότι έχουμε όμοια συμπεριφορά με το multiclass baseline μοντέλο, χωρίς υποβάθμιση της επίδοσής του.

7.3 Federated learning με non-i.i.d. multiclass δεδομένα

Σκοπός των συγκεκριμένων πειραμάτων, είναι να εξετάσουμε κατά πόσο το multiclass federated learning μοντέλο είναι ανθεκτικό στην περίπτωση όπου τα δεδομένα εκπαίδευσης πάψουν να υπόκεινται στην i.i.d. υπόθεση. Στην ενότητα αυτή παρουσιάζουμε τα αποτελέσματα εκπαίδευσης και τις τιμές των μετρικών αξιολόγησης, και στο επόμενο κεφάλαια θα το συγκρίνουμε αναλυτικότερα, τόσο με την επίδοση του federated learning μοντέλου με i.i.d. δεδομένα, όσο και με την επιμέρους επίδοση των client όταν εκπαιδεύονται μεμονωμένα στα non-i.i.d. δεδομένα.

7.3.1 Σενάριο 1

Στο σενάριο 1, τα δεδομένα κάθε κλάσης που αντιπροσωπεύει κακόβουλη κίνηση κατανέμονται εκ περιτροπής στους clients με αναλογίες $(0.7, 0.1, 0.1, 0.1)$, όπως αναλύθηκε στην υποενότητα 6.3.2. Η πορεία εκπαίδευσης του κοινού μοντέλου που εκπαιδεύεται με federated learning για το σενάριο φαίνεται στο σχήμα 7.9



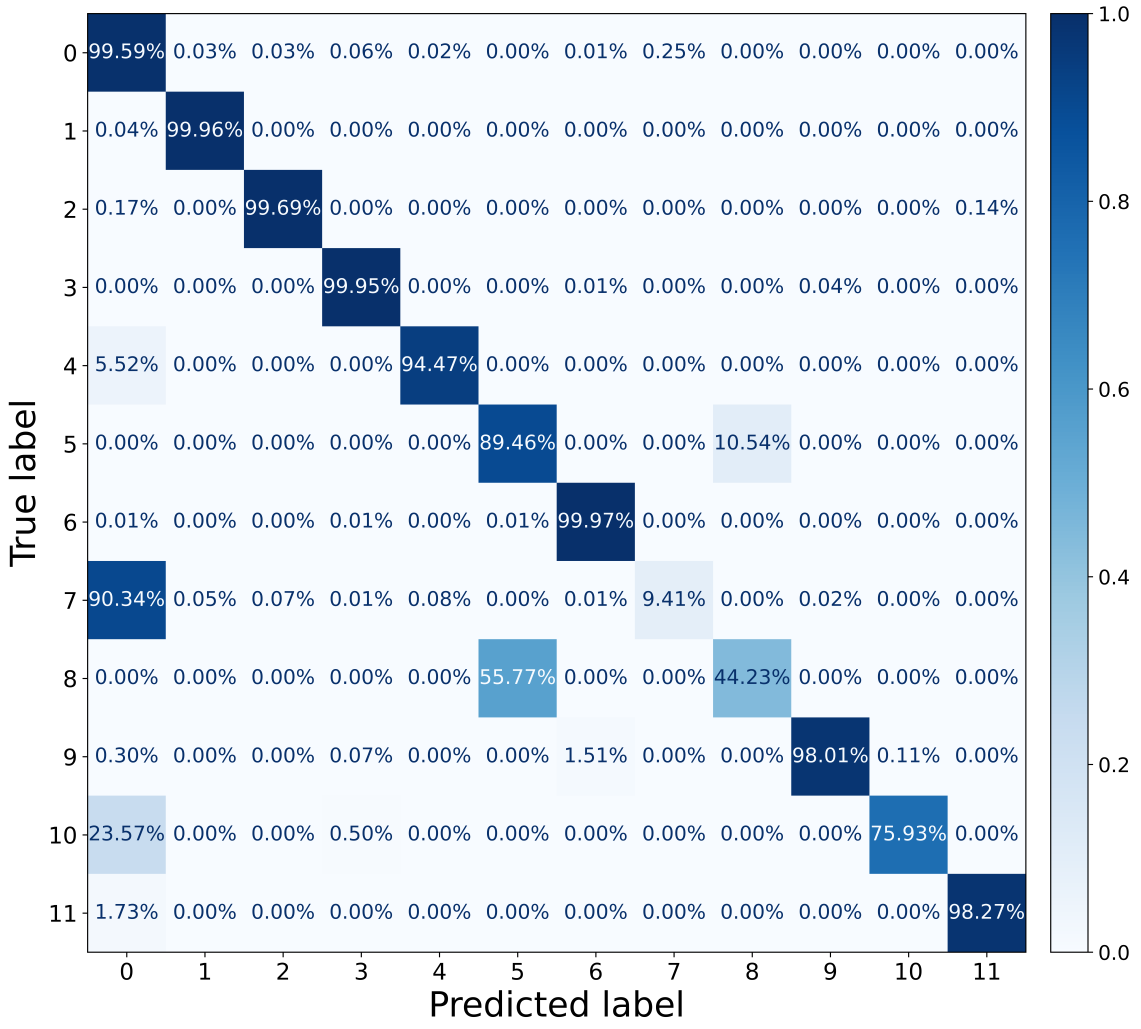
Σχήμα 7.9: *Training και validation loss κατά την εκπαίδευση του federated learning multiclass μοντέλου στο σενάριο 1 non-i.i.d. δεδομένων*

Όπως φαίνεται, παρά τον μη ομοιόμορφο τρόπο με τον οποία έχουν χωριστεί τα δεδομένα εκπαίδευσης στους clients, το κοινό μοντέλο εξακολουθεί να εκπαιδεύεται ομαλά και μόλις σε τρεις γύρους έχει πλησιάσει την βέλτιστη τιμή validation loss για το συγκεκριμένο πείραμα. Μετά την ολοκλήρωση της εκπαίδευσης, αξιολογούμε το κοινό μοντέλο με την χρήση των test data. Οι τιμές των μετρικών αξιολόγησης παρουσιάζονται στον πίνακα 7.5, ενώ στο σχήμα 7.10 σχεδιάζεται το confusion matrix του μοντέλου για τα test data.

Μετρική	Τιμή
Accuracy	98.015%
Macro F_1	0.8386
Macro Precision	0.8618
Macro Recall	0.8408
Weighted F_1	0.9768
Weighted Precision	0.9757
Weighted Recall	0.9801

Πίνακας 7.5: *Τιμές μετρικών αξιολόγησης για το federated learning multiclass μοντέλο στο σενάριο 1 non-i.i.d. δεδομένων*

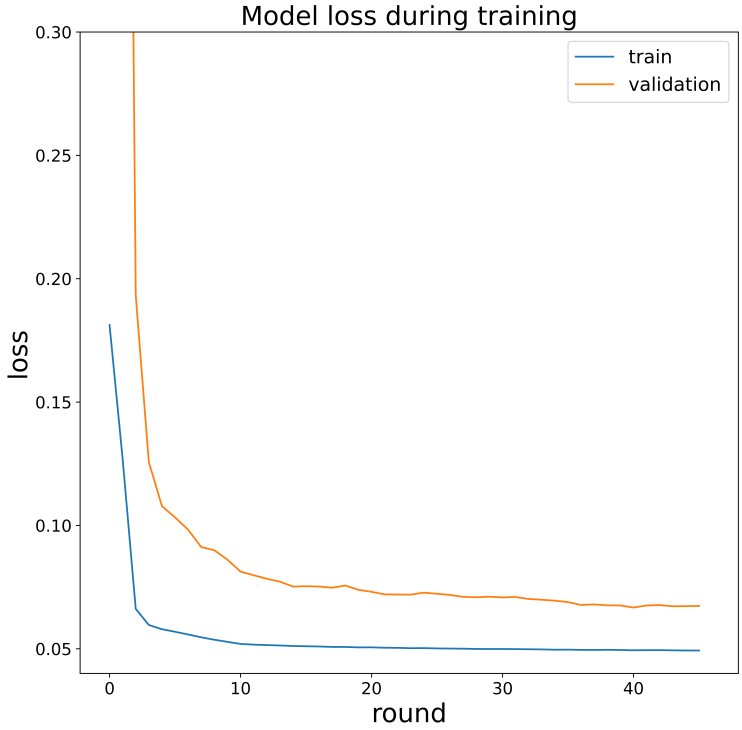
Βλέπουμε ότι τα αποτελέσματα είναι παρεμφερή με την περίπτωση του federated learning multiclass μοντέλο, όταν δεν είχε επιβληθεί συγκεκριμένη ανομοιόμορφη κατανομή στα δεδομένα εκπαίδευσης.



Σχήμα 7.10: *Confusion matrix του federated learning multiclass μοντέλου βάσει των test data για το σενάριο 1 non-i.i.d. δεδομένων*

7.3.2 Σενάριο 2

Στο σενάριο 2, τα δεδομένα κάθε κλάσης που αντιπροσωπεύει κακόβουλη κίνηση κατανέμονται στους clients με αναλογίες $(0.0, 0.6, 0.2, 0.2)$, δηλαδή για την εκάστοτε κλάση κάποιος client δεν θα περιέχει καθόλου δεδομένα της, όπως αναλύθηκε στην υποενότητα 6.3.2. Η πορεία εκπαίδευσης του κοινού μοντέλου που εκπαιδεύεται με federated learning για το σενάριο 2 φαίνεται στο σχήμα 7.11



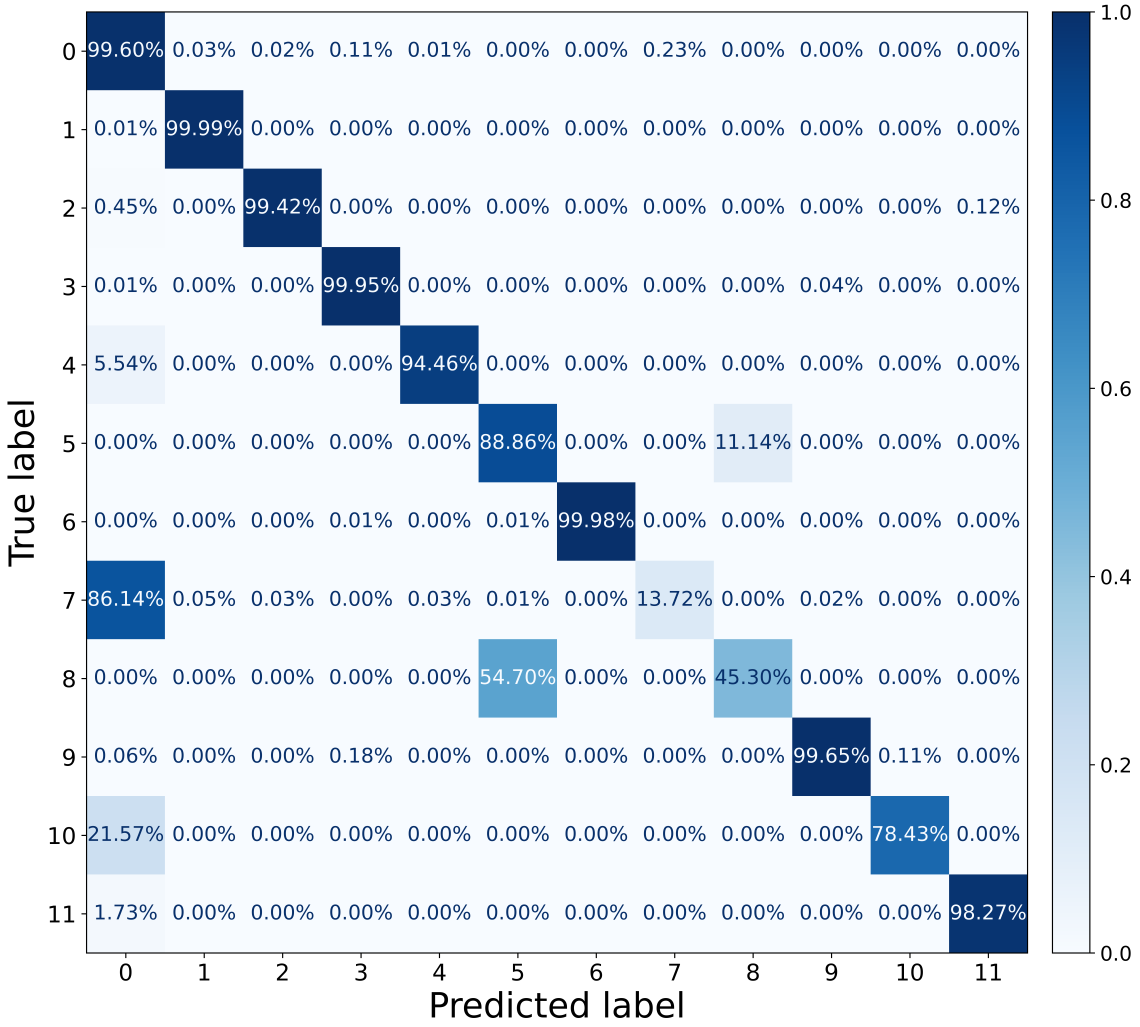
Σχήμα 7.11: *Training και validation loss κατά την εκπαίδευση του feaderated learning multiclass μοντέλου στο σενάριο 2 non-i.i.d. δεδομένων*

Παρατηρούμε ότι και τώρα, παρά το ότι τα δεδομένα εκπαίδευσης κατανέμονται ανομοιόμορφα στους clients, και απουσιάζουν ορισμένες κλάσεις από τον καθένα, το μοντέλο συγκλίνει ομαλά, σε περισσότερους γύρους federated averaging. Αφού ολοκληρωθεί η εκπαίδευση, αξιολογούμε το μοντέλο, με τις τιμές των μετρικών να φαίνονται στον πίνακα 7.6.

Μετρική	Τιμή
Accuracy	98.068%
Macro F_1	0.8478
Macro Precision	0.8736
Macro Recall	0.8480
Weighted F_1	0.9777
Weighted Precision	0.9771
Weighted Recall	0.9807

Πίνακας 7.6: *Τιμές μετρικών αξιολόγησης για το federated learning multiclass μοντέλο στο σενάριο 2 non-i.i.d. δεδομένων*

Βλέπουμε ότι το μοντέλο δείχνει ανοχή στις διαφορετικές κατανομές των δεδομένων και συνεχίζει να επιδεικνύει accuracy άνω του 98%. Όπως και για τα υπόλοιπα πειράματα, στο σχήμα 7.12 παραθέτουμε το confusion matrix, ώστε να φανεί αναλυτικά η επίδοση του μοντέλου στο σενάριο 2 για όλες τις κλάσεις.



Σχήμα 7.12: *Confusion matrix του federated learning multiclass μοντέλου βάσει των test data για το σενάριο 2 non-i.i.d. δεδομένων*

Κεφάλαιο 8

Σύγκριση και συζήτηση αποτελεσμάτων

Έχοντας παρουσιάσει για το κάθε πείραμα που εκτελέστηκε την πορεία εκπαίδευσής του και τις τιμές των μετρικών αξιολόγησης, όπως υπολογίστηκαν βάσει των test data, θα αφιερώσουμε το παρόν κεφάλαιο στην σύγκριση των αποτελεσμάτων και στα συμπεράσματα που προκύπτουν από αυτήν.

8.1 Το federated learning μοντέλο σε σχέση με το baseline

Το πρώτο θέμα που εξετάζεται είναι η μεταφορά του υβριδικού μοντέλου από το παραδοσιακό περιβάλλον εκπαίδευσης σε αυτό του federated learning. Έχοντας επιβεβαιώσει ότι το baseline μοντέλο εκπαιδεύεται ομαλά και πετυχαίνει καλό accuracy βάσει των test data, τόσο στο binary classification, όσο και στο multiclass classification, επαναλάβαμε την διαδικασία εκπαίδευσης, αυτή την φορά με cross-silo federated learning. Η αρχιτεκτονική και οι υπερπαράμετροι του υβριδικού μοντέλου παραμένουν σταθερά. Το κοινό συνεργατικό μοντέλο που προκύπτει αξιολογήθηκε χρησιμοποιώντας τα ίδια test data. Η σύγκριση της επίδοσης του baseline μοντέλου με το αντίστοιχο federated learning παρουσιάζεται συγκεντρωτικά στον πίνακα 8.1.

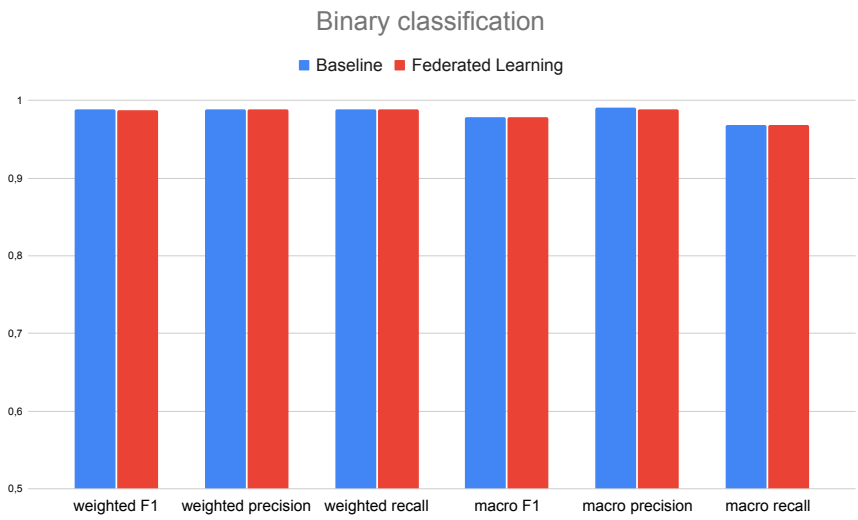
Μετρική	Baseline Binary	FL Binary	Baseline Multi	FL Multi
Accuracy	98.853%	98.824%	98.286%	98.221%
Weighted F_1	0.9884	0.9881	0.9779	0.9772
Weighted Precision	0.9886	0.9883	0.9785	0.9778
Weighted Recall	0.9885	0.9882	0.9829	0.9822
Macro F_1	0.9792	0.9788	0.8383	0.8358
Macro Precision	0.9905	0.9893	0.8860	0.8810
Macro Recall	0.9688	0.9689	0.8418	0.8423

Πίνακας 8.1: *Σύγκριση μετρικών αξιολόγησης binary και multiclass classification μεταξύ baseline και federated learning μοντέλου*

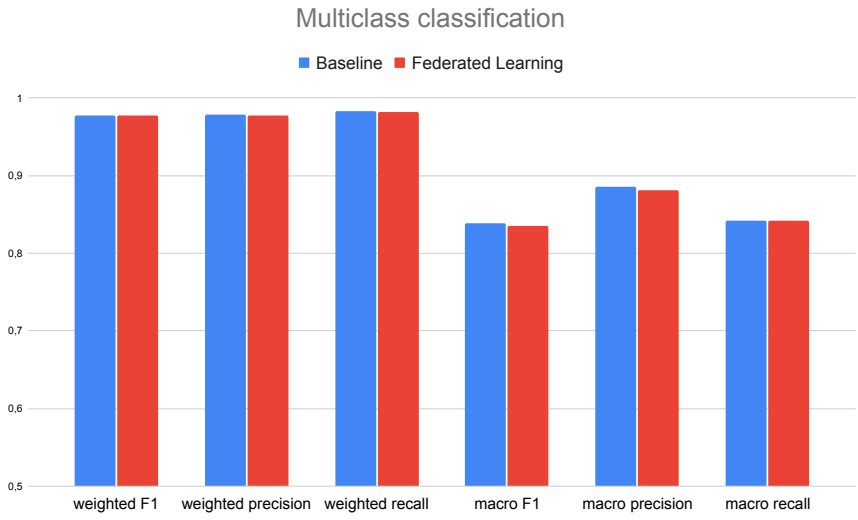
Παρατηρούμε ότι τόσο στην περίπτωση του binary όσο και στην περίπτωση του multiclass classification το μοντέλο federated learning πετυχαίνει σχεδόν το ίδιο accuracy με το αντίστοιχο baseline. Παρόμοια συμπεριφορά εμφανίζεται και στις υπόλοιπες μετρικές, με το federated learning να έχει κιόλας καλύτερη επίδοση σε κάποιες από αυτές. Επιπλέον, βλέπουμε ότι το multiclass accuracy παρουσιάζει χαμηλή πτώση (~ 0.6%) σε σχέση με τις αντίστοιχες binary περιπτώσεις, παρά το γεγονός ότι το μοντέλο καλείται να διαχωρίσει 12 αντί για 2 κλάσεις.

Μια παρατήρηση που αξίζει να σημειωθεί, είναι η διαφορά στις τιμές των μετρικών αξιολόγησης κατά το multiclass classification. Βλέπουμε ότι η μέση τιμή των μετρικών όταν χρησιμοποιείται απλός μέσος όρος (macro) είναι χαμηλότερος από τις αντίστοιχες τιμές του σταθμισμένου μέσου όρου (weighted). Αυτό μπορεί να εξηγηθεί από το γεγονός ότι το μοντέλο που έχει επιλεχθεί υστερεί στην ταξινόμηση συγκεκριμένων κακόβουλων κλάσεων (στην 7, και σε μικρότερο βαθμό στην 8), όπως φαίνεται και στα confusion matrices του κεφαλαίου 7. Καθώς το dataset είναι imbalanced, και οι συγκεκριμένες κλάσεις δεν αποτελούν πλειοψηφία, η χρήση σταθμισμένου μέσου δεν αναδεικνύει αυτή την ολιγωρία.

Για καλύτερη οπτικοποίηση της σύγκρισης του baseline με το federated learning μοντέλο, παρουσιάζονται γραφικά οι μετρικές που αναλύθηκαν στα σχήματα 8.1 και 8.2



Σχήμα 8.1: *Γραφική σύγκριση των μετρικών αξιολόγησης μεταξύ baseline και federated learning μοντέλου στο binary classsification*



Σχήμα 8.2: *Γραφική σύγκριση των μετρικών αξιολόγησης μεταξύ baseline και federated learning μοντέλου στο multiclass classsification*

8.2 Federated learning με non-i.i.d. δεδομένα εκπαίδευσης

Σε αυτή την ενότητα επιβάλλουμε συγκεκριμένες ανομοιόμορφες και μη ισορροπημένες κατανομές στα δεδομένα εκπαίδευσης και εξετάζουμε την επίδραση που αυτό θα έχει στην επίδοση του federated learning μοντέλου κατά το multiclass classification. Διακρίνουμε δύο σενάρια, όπως αυτά περιγράφηκαν αναλυτικά στα προηγούμενα κεφάλαια.

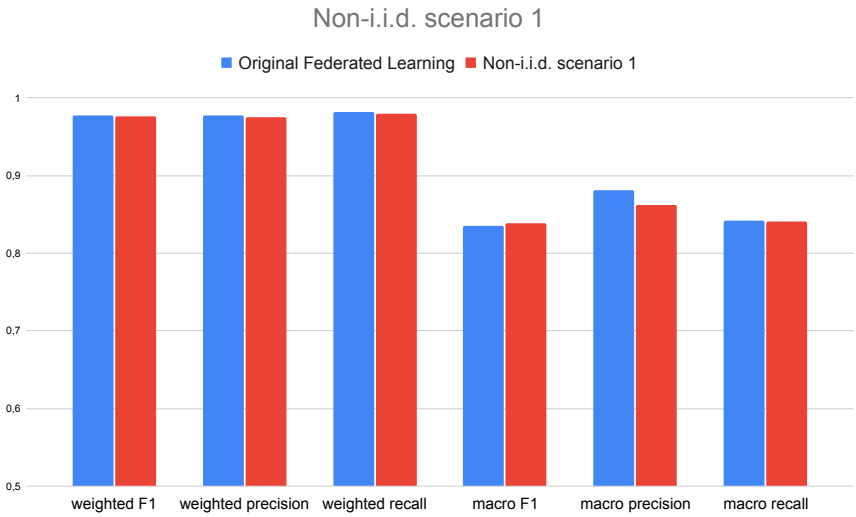
8.2.1 Σύγκριση με την περίπτωση ομοιόμορφα τυχαία κατανεμημένων δεδομένων εκπαίδευσης

Αρχικά θα εξετάσουμε την μεταβολή στην επίδοση του federated learning μοντέλου όταν προσομοιώθηκε το σενάριο 1 (κατανομή $(0.7, 0.1, 0.1, 0.1)$ ανά κακόβουλη κλάση) και το σενάριο 2 (κατανομή $(0.0, 0.6, 0.2, 0.2)$ ανά κακόβουλη κλάση). Σημειώνεται ότι όπως και σε όλα τα προηγούμενα πειράματα τα test data παραμέμουν ίδια. Ακόμα, θέλοντας να εξετάσουμε την ανθεκτικότητα του federated learning μοντέλου όταν τα δεδομένα εκπαίδευσης διαφέρουν, διατηρούμε τις υπερπαραμέτρους εκπαίδευσης ίδιες, χωρίς δηλαδή να γίνει fine tuning για τα δύο αυτά σενάρια. Οι τιμές των μετρικών αξιολόγησης παρουσιάζονται στον πίνακα 8.2.

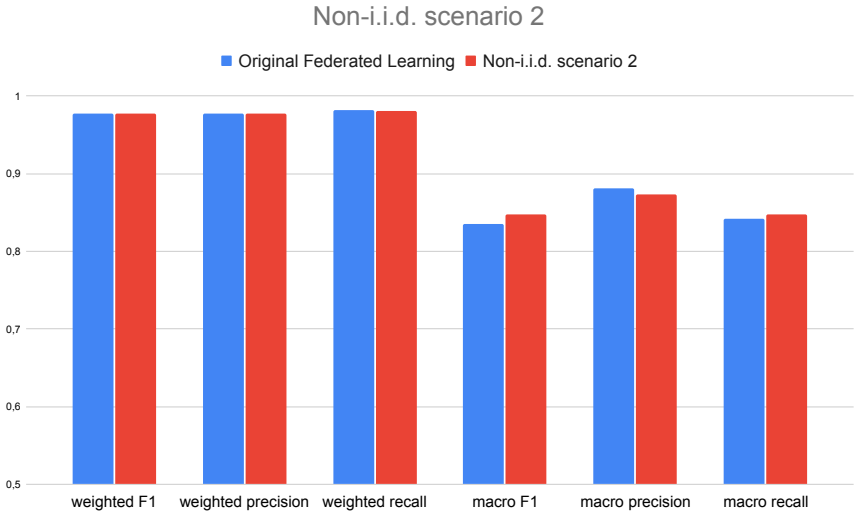
Μετρική	FL Original	Scenario 1	Scenario 2
Accuracy	98.221%	98.015%	98.068%
Weighted F_1	0.9772	0.9768	0.9777
Weighted Precision	0.9778	0.9757	0.9771
Weighted Recall	0.9822	0.9801	0.9807
Macro F_1	0.8358	0.8386	0.8478
Macro Precision	0.8810	0.8618	0.8736
Macro Recall	0.8423	0.8408	0.8480

Πίνακας 8.2: *Σύγκριση επίδοσης federated learning μοντέλου σε multiclass classification με πειράματα non-i.i.d. δεδομένων εκπαίδευσης*

Παρατηρούμε πως παρά το γεγονός ότι και στα δύο σενάρια επιβάλλεται ανομοιόμορφη κατανομή που αναιρεί την i.i.d. υπόθεση για τα δεδομένα, το federated learning μοντέλο καταφέρνει να αντεπεξέλθει, διατηρώντας accuracy μεγαλύτερο του 98%. Το weighted F_1 score παραμένει σχεδόν σταθερό, ενώ και στα δύο σενάριο το macro F_1 score βελτιώνεται συγκριτικά με το αντίστοιχο στην περίπτωση της τυχαίας και ομοιόμορφης κατανομής των δεδομένων εκπαίδευσης. Τα αποτελέσματα αυτά υποδεικνύουν πως το υβριδικό μοντέλο όταν εκπαιδεύεται με federated learning σε cross-silo περιβάλλον, μπορεί να παρουσιάσει ανθεκτικότητα στην εκπαίδευση σε διαφορετικές κατανομές των δεδομένων στους clients που συνεργάζονται. Όπως και στην προηγούμενη ενότητα, παρουσιάζουμε γραφικά την σύγκριση της επίδοσης των σεναρίων 1 και 2 σε σχέση με το αρχικό federated learning μοντέλο στα σχήματα 8.3 και 8.4.



Σχήμα 8.3: *Γραφική σύγκριση των μετρικών αξιολόγησης του federated learning μοντέλου στην αρχική κατανομή δεδομένων εκπαίδευσης και στο σενάριο 1 non-i.i.d. δεδομένων*



Σχήμα 8.4: *Γραφική σύγκριση των μετρικών αξιολόγησης του federated learning μοντέλου στην αρχική κατανομή δεδομένων εκπαίδευσης και στο σενάριο 2 non-i.i.d. δεδομένων*

8.2.2 Σύγκριση με την επίδοση των μεμονωμένων clients

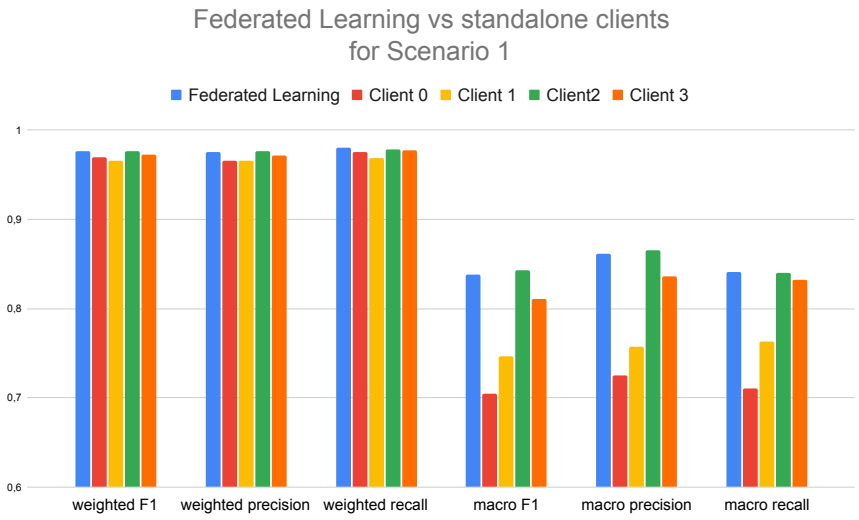
Στην περίπτωση των non-i.i.d. δεδομένων εκπαίδευσης στους clients, ιδιαίτερο ενδιαφέρον παρουσιάζει η σύγκριση του κοινού συνεργατικού μοντέλου που προέκυψε από το federated learning σε σχέση με την ικανότητα που παρουσιάζουν οι επιμέρους clients στο να εκπαιδεύσουν ένα ατομικό μοντέλο, κάνοντας χρήση μόνο των τοπικών τους δεδομένων. Στην συνέχεια παρουσιάζονται τόσο οι μετρικές αξιολόγησης για τον κάθε client όταν αυτός εκπαιδευτεί ατομικά, όσο και η πορεία της εκπαίδευσης τους. Τα test data παραμένουν ίδια με τα υπόλοιπα πειράματα, ενώ κοινή παραμένει επίσης η αρχιτεκτονική και οι υπερ-

παράμετροι που χρησιμοποιούνται για την εκπαίδευση των μεμονωμένων μοντέλων σε κάθε client.

Ξεκινώντας με το σενάριο 1, έχοντας επιβάλει σε κάθε client μη i.i.d. δεδομένα με κατανομή $(0.7, 0.1, 0.1, 0.1)$ ανά κακόβουλη κλάση, αξιολογούμε τα επιμέρους μοντέλα των clients. Στον πίνακα 8.3 φαίνονται αναλυτικά οι τιμές των μετρικών αξιολόγησης, οι οποίες παρουσιάζονται και γραφικά στο σχήμα 8.5.

Μετρική	Federated Learning	Client 0	Client 1	Client 2	Client 3
Accuracy	98.015%	97.579%	96.852%	97.815%	97.795%
Weighted F_1	0.9768	0.9693	0.9655	0.9764	0.9723
Weighted Precision	0.9757	0.9655	0.9658	0.9762	0.9719
Weighted Recall	0.9801	0.9758	0.9685	0.9781	0.9780
Macro F_1	0.8386	0.7049	0.7465	0.8427	0.8112
Macro Precision	0.8618	0.7247	0.7569	0.8657	0.8364
Macro Recall	0.8408	0.7110	0.7635	0.8402	0.8319

Πίνακας 8.3: *Σύγκριση τιμών μετρικών αξιολόγησης του federated learning μοντέλου με τα ατομικά μοντέλα που εκπαιδεύει ο κάθε client για το σενάριο 1 non-i.i.d. δεδομένων*



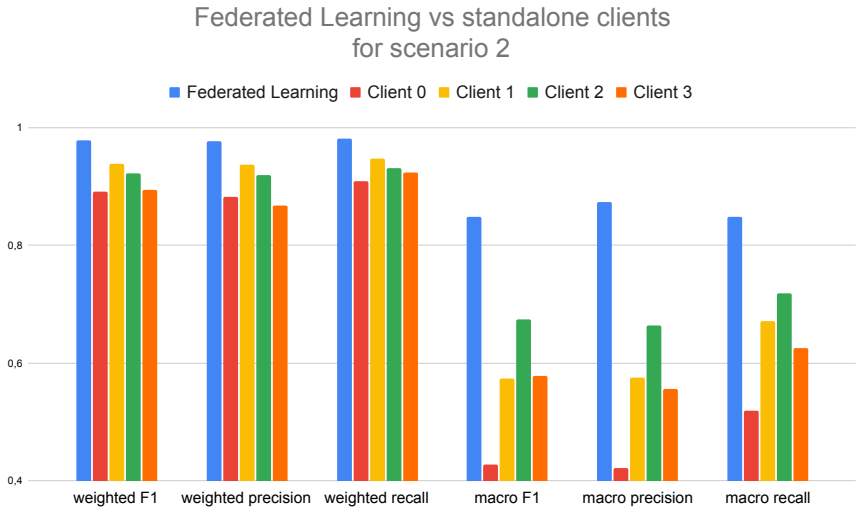
Σχήμα 8.5: *Γραφική σύγκριση των μετρικών αξιολόγησης του federated learning μοντέλου με τα ατομικά μοντέλα που εκπαιδεύει ο κάθε client για το σενάριο 1 non-i.i.d. δεδομένων*

Παρατηρούμε ότι το accuracy και σχεδόν όλες οι τιμές των weighted μετρικών για τα ατομικά μοντέλα είναι χαμηλότερες από αυτές του συνεργατικού κοινού μοντέλου. Το γεγονός ότι η διαφορά είναι μικρή, μπορεί να αιτιολογηθεί καθώς και το baseline μοντέλο είχε δείξει ότι μπορεί να αντεπεξέλθει σε unbalanced δεδομένα εκπαίδευσης στα οποία όμως υπάρχουν όλες οι κλάσεις με τα φυσιολογικά πακέτα να υπερτερούν, όπως αυτά που υπάρχουν στον κάθε client. Ακόμα, η χρήση των weighted μετρικών αποκρύπτει τις αδυναμίες του εκάστοτε μοντέλου ως προς τις κλάσεις που υποεκπροσωπούνται. Η ισχύς του μοντέλου federated learning σε σχέση με τα ατομικά μοντέλα φαίνεται καθαρά στις macro μετρικές, που υπολογίζουν τον μέσο όρο αποδίδοντας χωρίς συντελεστές βαρύτητας σε κάθε κλάση.

Ίδια μεθοδολογία ακολουθείται και για την κατανομή (0.0, 0.6, 0.2, 0.2) ανά κακόβουλη κλάση του σεναρίου 2. Στο συγκεκριμένο σενάριο, λόγω της κατανομής που έχει επιβληθεί, από τον κάθε client θα απουσιάζουν ορισμένες κακόβουλες κλάσεις από τα δεδομένα εκπαίδευσής του. Όμοια με το σενάριο 1, παραθέτουμε στον πίνακα 8.4 τις τιμές των μετρικών αξιολόγησης για το σενάριο 2, οι οποίες παρουσιάζονται και γραφικά στο σχήμα 8.6.

Μετρική	Federated Learning	Client 0	Client 1	Client 2	Client 3
Accuracy	98.068%	90.842%	94.721%	93.116%	92.393%
Weighted F_1	0.9777	0.8914	0.9385	0.9219	0.8942
Weighted Precision	0.9771	0.8819	0.9368	0.9190	0.8682
Weighted Recall	0.9807	0.9084	0.9472	0.9312	0.9239
Macro F_1	0.8478	0.4278	0.5731	0.6741	0.5778
Macro Precision	0.8736	0.4213	0.5753	0.6638	0.5559
Macro Recall	0.8480	0.5196	0.6714	0.7190	0.6262

Πίνακας 8.4: *Σύγκριση τιμών μετρικών αξιολόγησης του federated learning μοντέλου με τα ατομικά μοντέλα που εκπαιδεύει ο κάθε client για το σενάριο 2 non-i.i.d. δεδομένων*



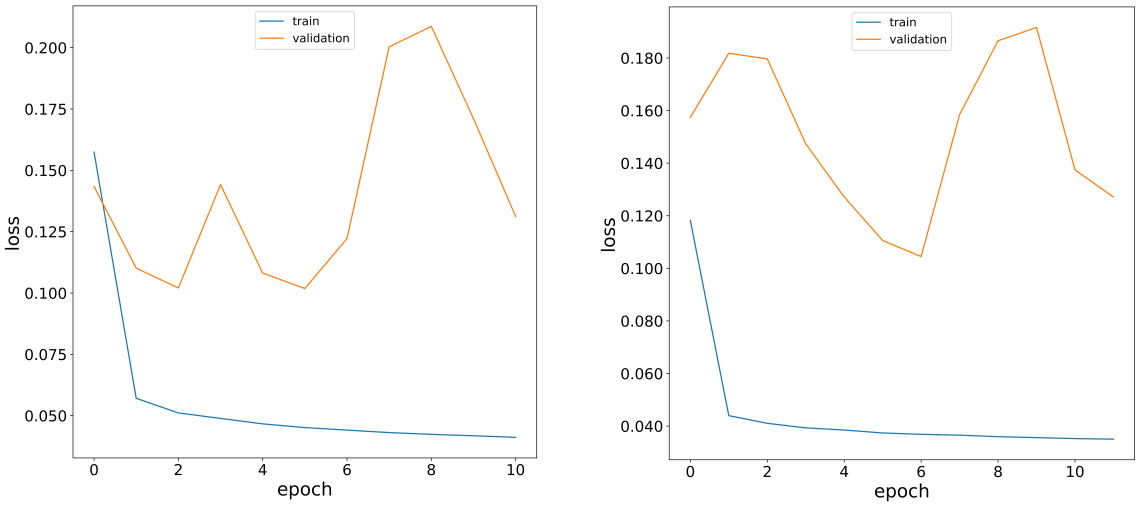
Σχήμα 8.6: *Γραφική σύγκριση των μετρικών αξιολόγησης του federated learning μοντέλου με τα ατομικά μοντέλα που εκπαιδεύει ο κάθε client για το σενάριο 2 non-i.i.d. δεδομένων*

Από τα παραπάνω βλέπουμε για το σενάριο 2 περαιτέρω μείωση στο accuracy και τις weighted μετρικές, σε σχέση με το συνεργατικό μοντέλο που εκπαιδεύτηκε με federated learning. Αυτό πιθανώς οφείλεται στο γεγονός ότι ο κάθε client καλείται να ταξινομήσει, στα test data, δεδομένα κλάσεων που δεν έχει συναντήσει κατά την εκπαίδευση. Αυτή η μείωση αναδεικνύει ένα ακόμα πλεονέκτημα του fedearated learning, καθώς το κοινό μοντέλο που προκύπτει έχει εκπαιδευτεί έμμεσα, σε κάποιο βαθμό, για κάθε κλάση που υπάρχει στα train data, ασχέτως αν αυτή δεν εμφανίζεται σε κάποιον client. Η υπεροχή του federated learning είναι ακόμα πιο εμφανής, βάσει των macro averaged μετρικών, οι οποίες δεν επηρεάζονται από την ανισορροπία εμφάνισης κλάσεων στα δεδομένα, όπου τα ατομικά μοντέλα φαίνεται να αποτυγχάνουν.

8.2.3 Συμπεριφορά εκπαίδευσης σε σύγκριση με μεμονωμένους clients

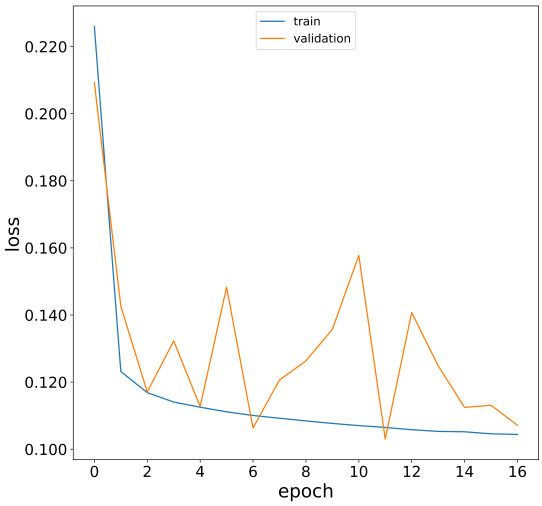
Για την διερεύνηση περισσότερων διαφορών που προκύπτουν μεταξύ της χρήσης federated learning και της εκπαίδευσης ατομικών μοντέλων από τους client, με χρήση μόνο των τοπικών δεδομένων τους, όταν αίρουμε την i.i.d. υπόθεση, θα εξετάσουμε την πορεία της εκπαίδευσής τους βάσει του training και validation loss. Το validation dataset που χρησιμοποιείται σε κάθε client είναι κοινό και ίδιο με το αντίστοιχο validation dataset που χρησιμοποιήθηκε κατά την εκπαίδευση του federated learning μοντέλου στο εκάστοτε σενάριο. Αυτά τα validation data έχουν ίδιο πλήθος με τα αντίστοιχα test data και περιέχουν δεδομένα όλων των κλάσεων.

Στο σχήμα 8.7 που ακολουθεί φαίνεται η πορεία εκπαίδευσης client κατά την ατομική εκπαίδευση τους για τα τοπικά δεδομένα που διαθέτουν στο σενάριο 1 non-i.i.d. δεδομένων. Για διευκόλυνση της σύγκρισης, παραθέτουμε ξανά αμέσως μετά και την αντίστοιχη πορεία εκπαίδευσης του federated learning μοντέλο στο σχήμα 8.8.

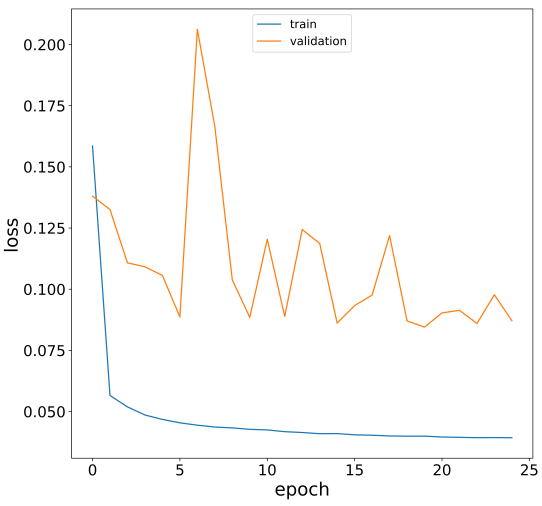


Client 0

Client 1

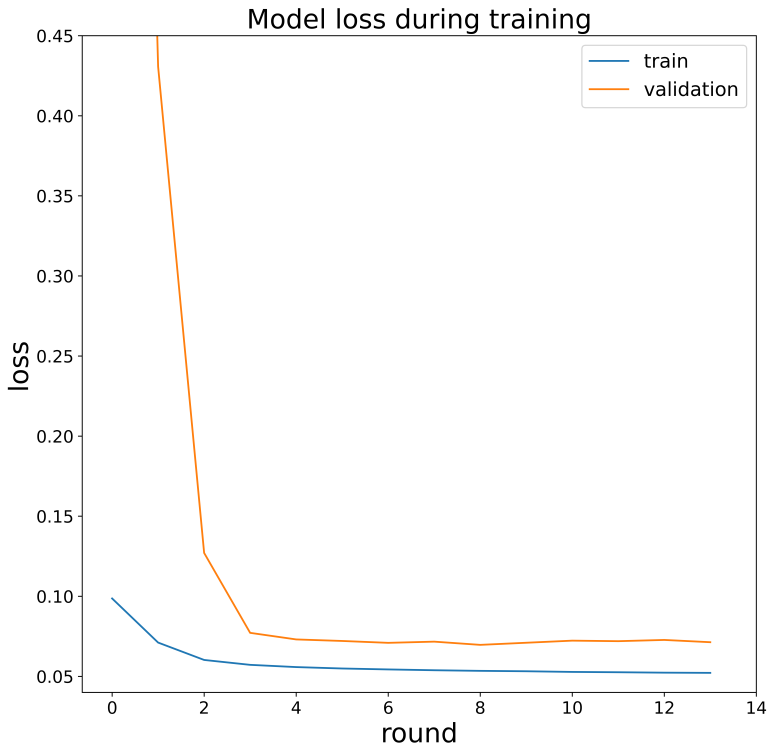


Client 2



Client 3

Σχήμα 8.7: *Πορεία εκπαίδευσης του κάθε client μεμονωμένα στο non-i.d.d. σενάριο 1*



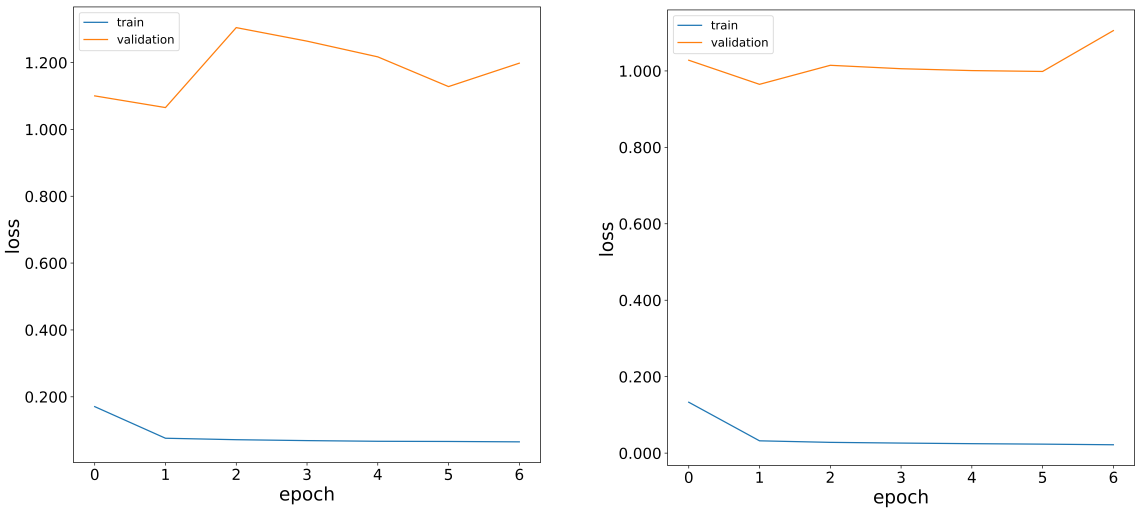
Σχήμα 8.8: *Train και validation loss κατά την εκπαίδευση του federated learning multiclass μοντέλου στο σενάριο 1 non-i.i.d. δεδομένων (επανάληψη για διευκόλυνση σύγκρισης)*

Βάσει των σχημάτων 8.7 και 8.8 είναι φανερή η υπεροχή του federated learning όσον αφορά την πορεία εκπαίδευσης και τα χαρακτηριστικά της. Βλέπουμε πως οι μεμονωμένοι clients παρουσιάζουν ετερογενή συμπεριφορά κατά την εκπαίδευση, κάτι το οποίο μπορεί να αιτιολογηθεί από την διαφορετική κατανομή των δεδομένων εκπαίδευσης στον καθένα από αυτούς. Έτσι, στο συγκεκριμένο παράδειγμα, ενώ οι clients 2 και 3 φαίνεται να προσεγγίζουν μια τιμή σύγκλισης, οι clients 0 και 1 παρουσιάζουν πιο αποκλίνουσα συμπεριφορά. Ασφαλώς, κάτι τέτοιο θα μπορούσε να βελτιωθεί με την χρήση τεχνικών που συνεισφέρουν στην ομαλή σύγκλιση ή επιτρέποντας στα τοπικά μοντέλα να εκπαιδευτούν για περισσότερες εποχές. Στη βάση όμως της σύγκρισης με το κοινό μοντέλο federated learning, διατηρώντας ίδιες υπερπαραμέτρους εκπαίδευσης, τα τοπικά μοντέλα έχουν υποδεέστερη συμπεριφορά.

Περαιτέρω διαφορές που παρατηρούνται κατά την εκπαίδευση των μεμονωμένων client στο σενάριο 1 με χρήση μόνο τοπικών δεδομένων, πέρα από την σημαντική αδυναμία σταθερής μείωσης του validation loss και σύγκλισης, είναι η μεταβλητή διάρκεια της εκπαίδευσης σε κάθε client (σε εποχές εκπαίδευσης), καθώς και το γεγονός ότι τα τελικά validation losses που προέκυψαν είναι υψηλότερα από το αντίστοιχο του federated learning κοινού μοντέλου. Για το κοινό μοντέλο, βλέπουμε στο σχήμα 8.8, ότι η σύγκλιση κατά την εκτέλεση του federated averaging επιτυγχάνεται ομαλά και σύντομα.

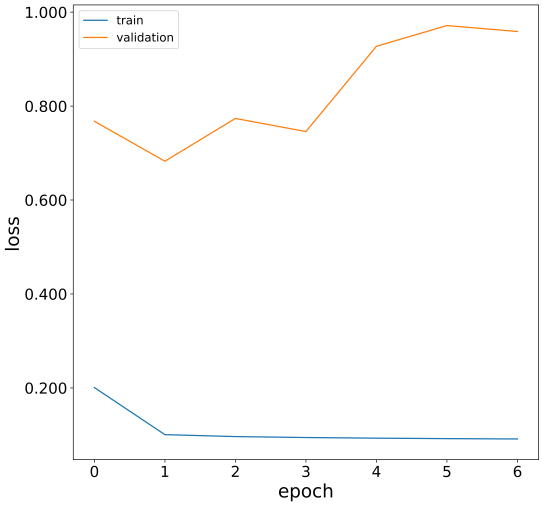
Όμοια με το σενάριο 1, θα εξετάσουμε την συμπεριφορά κατά την εκπαίδευση των μεμονωμένων clients και στο σενάριο 2, όταν εκπαιδεύονται χρησιμοποιώντας μόνο τα τοπικά τους δεδομένα, χωρίς δηλαδή την χρήση federated learning. Υπενθυμίζουμε ότι η κατανομή δεδομένων εκπαίδευσης στους clients που έχει επιβληθεί στο σενάριο 2 ανά κλάση κακόβουλης κίνησης είναι $(0.0, 0.6, 0.2, 0.2)$, δηλαδή η πορεία της εκπαίδευσης παρουσιάζει

ενδιαφέρον καθώς σε κάθε client απουσιάζουν ορισμένες διαφορετικές κλάσεις. Στο σχήμα 8.9 παρουσιάζεται συγκεντρωτικά η μεταβολή του train και validation loss, ανά εποχή.

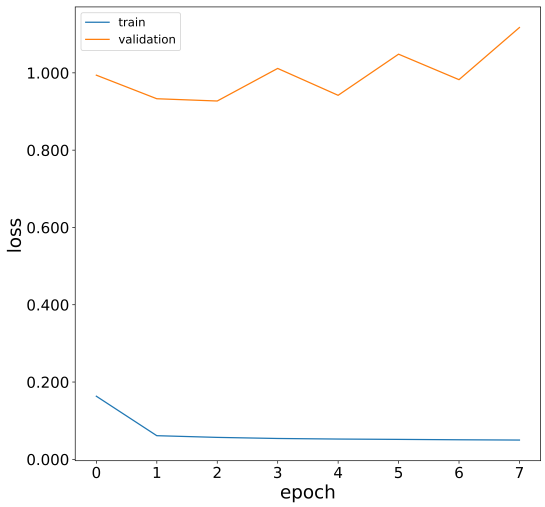


Client 0

Client 1



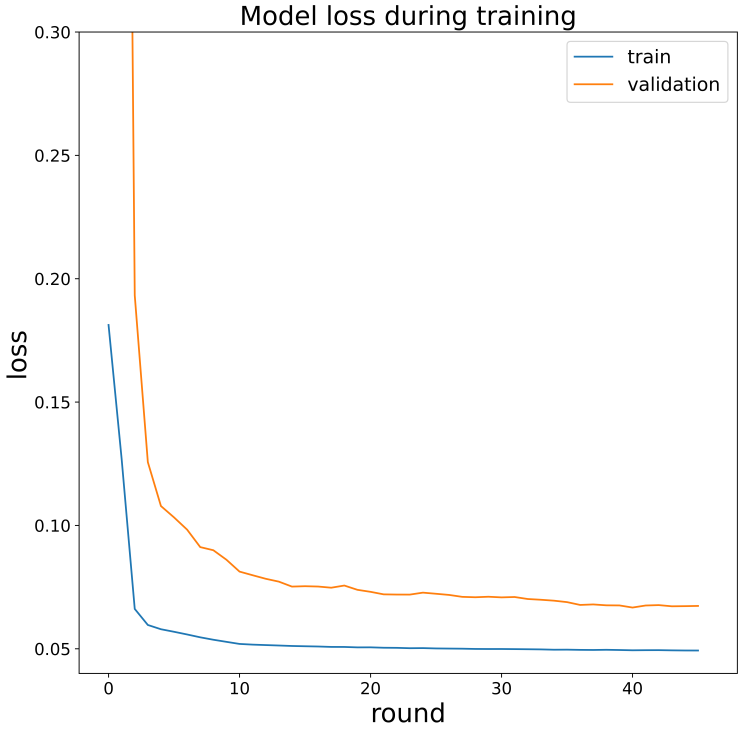
Client 2



Client 3

Σχήμα 8.9: *Πορεία εκπαίδευσης του κάθε client μεμονωμένα στο non-i.i.d. σενάριο 2*

Όπως και προηγουμένως, για διευκόλυνση της σύγκρισης στο σενάριο 2, παραθέτουμε ξανά στην επόμενη σελίδα και την αντίστοιχη πορεία εκπαίδευσης του κοινού federated learning μοντέλου στο σχήμα 8.10.



Σχήμα 8.10: *Train και validation loss κατά την εκπαίδευση του federated learning multiclass μοντέλου στο σενάριο 2 non-i.i.d. δεδομένων (επανάληψη για διευκόλυνση σύγκρισης)*

Βάσει των παραπάνω σχημάτων για το σενάριο 2, βλέπουμε πως και σε αυτή την περίπτωση, το μοντέλο που εκπαιδεύεται με federated learning υπερτερεί. Φαίνεται ότι οι επιμέρους clients, όταν εκπαιδεύονται ατομικά, δεν συγκλίνουν και η τιμή του validation loss παραμένει υψηλή. Αυτό δικαιολογεί τις χαμηλότερες τιμές μετρικών αξιολόγησης που πέτυχαν στα test data (8.6), τα οποία περιέχουν δεδομένα από όλες τις κλάσεις. Αντίθετα η πορεία εκπαίδευσης του federated learning μοντέλου για αυτό το σενάριο δείχνει ότι είναι σε θέση να συγκλίνει ομαλά και γρήγορα, με το validation loss να μειώνεται συνεχώς κατά την εκτέλεση του federated averaging. Έτσι, φαίνεται πως το κοινό μοντέλο που προκύπτει διαθέτει διακριτική ικανότητα για όλες της κλάσεις, ακόμα και όταν αυτές δεν εκπροσωπούνται σε κάθε client.

Μέρος III

Επίλογος

Κεφάλαιο 9

Συμπεράσματα

Στην παρούσα εργασία υλοποιήθηκε ένα συνεργατικό σύστημα ανίχνευσης εισβολής, βασισμένο σε βαθιά μηχανική μάθηση, εκπαιδευμένο με χρήση federated learning. Το υβριδικό μοντέλο που χρησιμοποιήθηκε συνδυάζει έναν μη συμμετρικό βαθύ autoencoder, ο οποίος εκπαιδεύεται με χρήση αποκλειστικά μη επισημασμένων δεδομένων, και ένα βαθύ νευρωνικό δίκτυο, το οποίο χρησιμοποιεί κωδικοποιημένα επισημασμένα δεδομένα για να εκπαιδευτεί στο να ταξινομεί δικτυακές απειλές. Σε αυτό το κεφάλαιο συνοψίζουμε τα συμπεράσματα που προέκυψαν από την υλοποίηση, τα πειράματα, και τις συγκρίσεις των αποτελεσμάτων τους.

9.1 Baseline υβριδικό μοντέλο

Αν και η υλοποίηση ενός μοντέλου που να επικρατεί, ως προς την επίδοσή του, των σύγχρονων IDS δεν αποτελεί το αντικείμενο της παρούσας εργασίας, το baseline μοντέλο που χρησιμοποιείται πρέπει να πληροί ορισμένες προϋποθέσεις. Για πιο ρεαλιστική προσομοίωση είναι σημαντικό να μπορεί να εκμεταλλεύεται την ύπαρξη μεγαλύτερου αριθμού μη επισημασμένων δεδομένων, καθώς αυτά απαντώνται κατά πλειοψηφία στην πράξη. Επιπλέον, η επίδοσή του βάσει συγκεκριμένων μετρικών αξιολόγησης είναι αναγκαίο να παραμένει υψηλή, τόσο στο binary όσο και στο multiclass classification παρά την ανισορροπία κλάσεων στα δεδομένα εκπαίδευσης.

Το υβριδικό μοντέλο που επιλέχτηκε χρησιμοποιεί και πλήρως unlabeled δεδομένα εκπαίδευσης, τα οποία λαμβάνονται από το dataset ώστε το πλήθος τους να είναι τριπλάσιο σε σχέση με τα labeled δεδομένα. Ακόμα, το CSE-CIC-IDS2018 που χρησιμοποιείται περιέχει πολύ περισσότερα πακέτα φυσιολογικής κίνησης συγκριτικά με κακόβουλα, είναι δηλαδή μη ισορροπημένο. Κατόπιν εκπαίδευσής του, βλέπουμε πως το μοντέλο που επιλέχτηκε αποδίδει ικανοποιητικά, βάσει των μετρικών, τόσο σε binary όσο και σε multiclass classification, διατηρώντας επιπλέον πολύ χαμηλό ποσοστό ψευδώς θετικών αποτελεσμάτων.

9.2 Χρήση Federated learning

Η αξιολόγηση του υβριδικού μοντέλου που εκπαιδεύεται με federated learning, πέρα από την συμπεριφορά του κατά την εκπαίδευση, γίνεται και με την σύγκριση του με το αντίστοιχο baseline μοντέλο όταν αυτό εκπαιδεύεται χρησιμοποιώντας όλα τα δεδομένα με

μη συνεργατικό τρόπο. Βάσει των αποτελεσμάτων και της σύγκρισης που παρουσιάζονται στα κεφάλαια 7 και 8, βλέπουμε ότι η χρήση federated learning για την εκπαίδευση ενός υβριδικού μοντέλου, δεν μειώνει αξιοσημείωτα την επίδοσή του ούτε στο binary ούτε στο multiclass classification. Αυτό είναι ενθαρρυντικό, καθώς φαίνεται πως δεν θυσιάζεται η αποτελεσματικότητα του μοντέλου στο σύνηθες trade-off έναντι της διασφάλισης της ιδιωτικότητας των δεδομένων. Πέραν της αύξησης της ιδιωτικότητας, η χρήση federated learning επιφέρει και άλλα πλεονεκτήματα, όπως την αξιοποίηση δεδομένων κατά την συνεργατική εκπαίδευση στα οποία οι επιμέρους clients δεν έχουν πρόσβαση υπό άλλες συνθήκες.

Μια σημαντική πρόκληση που καλούνται να αντιμετωπίσουν τα συστήματα ανίχνευσης επιθέσεων είναι η ετερογένεια που παρουσιάζουν τα δεδομένα, καθώς και το γεγονός ότι σε διαφορετικά δίκτυα υπάρχει σημαντικά διαφορετική κατανομή στο είδος της κίνησης που χρειάζεται να διαχειριστούν. Ένα χαρακτηριστικό του federated learning είναι ότι δεν προϋποθέτει i.i.d. δεδομένα για την εκπαίδευση του μοντέλου. Έτσι, για την πληρέστερη αξιολόγηση του συνεργατικού IDS, ορίζονται δύο σενάρια στα οποία επιβάλλεται συγκεκριμένη κατανομή των δεδομένων εκπαίδευσης, ώστε να μοντελοποιηθεί η ανισορροπία κακόβουλων κλάσεων μεταξύ των clients, καθώς και η απουσία ορισμένων κλάσεων σε καθέναν από αυτούς.

Τα αποτελέσματα των πειραμάτων που βασίζονται σε αυτά τα δύο σενάρια δείχνουν ότι το συνεργατικό μοντέλο δεν επηρεάζεται σημαντικά από τις προβληματικές κατανομές που επιβάλλονται, διατηρώντας την ικανότητα να διαχωρίζει τόσο την κακόβουλη από την φυσιολογική κίνηση όσο και την ταξινομεί. Ακόμα, μελετώντας την εκπαίδευση του κάθε μεμονωμένου client σε μη i.i.d. δεδομένα (στην περίπτωση δηλαδή όπου δεν γίνεται χρήση federated learning), φαίνεται πως δεν θα μπορούσαν να συγκλίνουν ομαλά κάνοντας χρήση μόνο του τοπικού τους training dataset, σε αντίθεση με το συνεργατικό federated learning μοντέλο στο οποίο δεν παρουσιάζεται αντίστοιχο πρόβλημα.

Βάσει των παραπάνω, μπορούμε να συμπεράνουμε ότι η χρήση federated learning για την ανάπτυξη συστημάτων ανίχνευσης εισβολής είναι πολλά υποσχόμενη και δύναται να διευρύνει τις δυνατότητες εποικοδομητικής συνεργασίας μεταξύ οργανισμών, υποδομών και δικτύων που παρουσιάζουν ετερογένεια τόσο στην υπολογιστική ισχύ, όσο και στα διαθέσιμα δεδομένα, με κοινό στόχο την ασφάλεια τους και την ιδιωτικότητα των δεδομένων τους. Δικαιολογείται έτσι το ερευνητικό ενδιαφέρον για την εξερεύνηση των ιδιοτήτων συστημάτων που βασίζονται στο federated learning, καθώς και η υλοποίηση μεθόδων για την βελτιστοποίηση της λειτουργίας τους.

Κεφάλαιο 10

Μελλοντικό έργο

Η σχετικά πρόσφατη εμφάνιση του federated learning προσφέρει την δυνατότητα για έρευνα σε ένα ευρύ ερευνητικό πεδίο, τόσο για την βελτιστοποίηση του, όσο και για προσπάθειες εφαρμογής του σε τομείς συστημάτων ασφαλείας που στηρίζονται στην βαθιά μηχανική μάθηση. Σε αυτό το κεφάλαιο, θα αναφερθούν μερικά δυνατά μελλοντικά έργα που μπορούν να πραγματοποιηθούν στον χώρο της ανάπτυξης συστημάτων ανίχνευσης εισβολής που βασίζονται σε μοντέλα τα οποία εκπαιδεύονται με federated learning, και ειδικότερα σε πιθανές μελλοντικές επεκτάσεις του υβριδικού συστήματος που παρουσιάζεται στην παρούσα εργασία.

Καθώς υπάρχει διαθέσιμη μεγάλη ποικιλία αρχιτεκτονικών που αφορούν μοντέλα επιβλεπόμενης και μη επιβλεπόμενης μάθησης, είναι σημαντικό να βρεθεί ο κατάλληλος συνδυασμός τους, που να αποδίδει βέλτιστα στην ανίχνευση και ταξινόμηση κακόβουλης δικτυακής κίνησης. Ιδανικά, ένα συνεργατικό IDS που βασίζεται σε federated learning για την εκπαίδευσή του, θα πρέπει να είναι σε θέση να αντιμετωπίζει όσο το δυνατό μεγαλύτερο εύρος επιθέσεων, ακόμα και zero-day attacks οι οποίες δεν έχουν αναγνωριστεί στο παρελθόν και το IDS δεν έχει εκτεθεί σε αυτές κατά την εκπαίδευσή του. Έτσι η εύρεση κατάλληλης αρχιτεκτονικής, με βέλτιστη επιλογή υπερπαραμέτρων, που να ισορροπεί ανάμεσα σε υψηλή επίδοση και ευκολία εκπαίδευσης, ανάγεται σε μείζον ζήτημα.

Η βελτιστοποίηση federated συστημάτων ανίχνευσης εισβολής, αποτελεί σημαντική ερευνητική πρόκληση ώστε να επεκταθεί η διάδοσή τους. Κύριες κατευθύνσεις είναι η επιτυχής προσαρμογή τους στην ετερογένεια συστημάτων και δεδομένων, το κόστος επικοινωνίας, καθώς και αύξηση της ασφάλειας και ιδιωτικότητας που προσφέρουν στους συμμετέχοντες. Πιο συγκεκριμένα, σε ότι αφορά την δικτυακή επιβάρυνση για την αναγκαία επικοινωνία κατά το federated learning, ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη της μεταβολής στην απόδοση τους όταν εφαρμόζονται τεχνικές που σκοπεύουν στην μείωση του όγκου δεδομένων που ανταλλάσσονται αλλά και γενικότερα του κόστους επικοινωνίας που απαιτείται. Ένα παράδειγμα για έρευνα προς αυτή την κατεύθυνση είναι η εφαρμογή μεθόδων συμπίεσης των δεδομένων επικοινωνίας κατά το federated learning, όπως η χρήση subsampling ή probabilistic quantization, που αναφέρθηκαν στην υποενότητα 2.3.4. Ένας ακόμη χώρος με ερευνητικό και πρακτικό ενδιαφέρον είναι η περαιτέρω αποκεντροποίηση της διαδικασίας με την αφαίρεση του aggregating server και την χρήση blockhain και άλλων consensus μηχανισμών.

Πέρα της αυξημένης ιδιωτικότητας που προσφέρει εγγενώς το federated learning καθώς

τα τοπικά δεδομένα εκπαίδευσης των clients δεν εκτίθενται στο δίκτυο, πολλές φορές υπάρχει η ανάγκη αυξημένης προστασίας με πιο σύνθετες τεχνικές, καθώς υπάρχει η δυνατότητα εξαγωγής συμπερασμάτων σχετικά με τα δεδομένα και τους χρήστες που τα παρήγαγαν ακόμα και από την ανταλλαγή των βαρών. Χαρακτηριστικό παράδειγμα μεθόδου που εγγυάται φορμαλιστικά συγκεκριμένο επίπεδο ιδιωτικότητας είναι το differential privacy. Ακόμα, τεχνικές κρυπτογράφησης για την επικοινωνία μεταξύ του server και των clients αλλά και μέθοδοι homomorphic encryption για την εκτέλεση υπολογισμών σε κρυπτογραφημένα δεδομένα, θα προσέφεραν υψηλότερη ασφάλεια και προστασία της ιδιωτικότητας στους συμμετέχοντες. Σημαντική πρόκληση στην ευρύτερη υιοθέτηση τέτοιων τεχνικών, την οποία καλούνται να εξετάσουν οι ερευνητές, είναι η διατήρηση της επίδοσης και αποδοτικότητας όταν λαμβάνονται αναγκαία μέτρα για την ενίσχυση της ασφάλειας και της ιδιωτικότητας.

Βιβλιογραφία

[1] Hung Jen Liao, Chun Hung Richard Lin, Ying Chih Lin και Kuang Yuan Tung. *Intrusion detection system: A comprehensive review*. *Journal of Network and Computer Applications*, 36(1):16–24, 2013. https://doi.org/10.1016/j.jnca.2012.09.004.

[2] Dilara Gümüşbaş, Tulay Yıldırım, Angelo Genovese και Fabio Scotti. *A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems*. *IEEE Systems Journal*, 15(2):1717–1731, 2021. https://doi.org/10.1109/JSYST.2020.2992966.

[3] D.E. Denning. *An Intrusion-Detection Model*. *IEEE Transactions on Software Engineering*, SE-13(2):222–232, 1987. https://doi.org/10.1109/TSE.1987.232894.

[4] Karen Scarfone και Peter Mell. *Guide to intrusion detection and prevention systems (IDPS)*. *NIST special publication*, 800-94, 2007. https://www.doi.org/10.6028/NIST.SP.800-94.

[5] Hervé Debar, Marc Dacier και Andreas Wespi. *Towards a taxonomy of intrusion-detection systems*. *Computer Networks*, 31(8):805–822, 1999. https://www.doi.org/10.1016/S1389-1286(98)00017-6.

[6] Zeeshan Ahmad, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah και Farhan Ahmad. *Network intrusion detection system: A systematic study of machine learning and deep learning approaches*. *Transactions on Emerging Telecommunications Technologies*, 32(1):4150, 2021. https://doi.org/10.1002/ett.4150.

[7] Hongyu Liu και Bo Lang. *Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey*. *Applied Sciences*, 9(20):4396, 2019. https://doi.org/10.3390/app9204396.

[8] Shaashwat Agrawal, Sagnik Sarkar, Ons Aouedi, Gokul Yenduri, Kandaraj Piamrat, Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta και Thippa Reddy Gadekallu. *Federated Learning for Intrusion Detection System: Concepts, Challenges and Future Directions*. *arXiv preprint arXiv:2106.09527v1*, 2021. https://doi.org/10.48550/arXiv.2106.09527.

[9] J. Rajahalme, A. Conta, B. Carpenter και S. Deering. *IPv6 Flow Label Specification*. RFC 3697, RFC Editor, Μάρτιος 2004. https://www.doi.org/10.17487/RFC3697.

[10] N. Brownlee, C. Mills C. και G. Ruth. *Traffic Flow Measurement: Architecture*. RFC 2722, RFC Editor, Οκτώβριος 1999. https://www.doi.org/10.17487/RFC2722.

[11] Iman Sharafaldin, Arash Habibi Lashkari και Ali A Ghorbani. *Toward generating a new intrusion detection dataset and intrusion traffic characterization*. *ICISSp*, 1:108–116, 2018.

[12] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes και Andreas Hotho. *A survey of network-based intrusion detection data sets*. *Computers & Security*, 86:147–167, 2019. https://doi.org/10.1016/j.cose.2019.06.005.

[13] B. Claise. *Cisco Systems NetFlow Services Export Version 9*. RFC 3954, RFC Editor, Οκτώβριος 2004. https://www.doi.org/10.17487/RFC3954.

[14] B. Claise, B. Trammell και P. Aitken. *Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information*. RFC 7011, RFC Editor, Σεπτέμβριος 2013. https://www.doi.org/10.17487/RFC7011.

[15] P. Phaal και M. Lavine. *sFlow Version 5*. Τεχνική Αναφορά, sFlow.org, Ιούλιος 2004. https://sflow.org/sflow_version_5.txt.

[16] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker και Jonathan Turner. *OpenFlow: enabling innovation in campus networks*. *ACM SIGCOMM computer communication review*, 38(2):69–74, 2008. https://doi.org/10.1145/1355734.1355746.

[17] *nfdump: Netflow processing tools*. https://github.com/phaag/nfdump. Ημερομηνία πρόσβασης: 27-05-2022.

[18] *Yet Another Flowmeter (YAF)*. https://tools.netsa.cert.org/yaf/. Ημερομηνία πρόσβασης: 27-05-2022.

[19] *KDD Cup 1999 Data*. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. Ημερομηνία πρόσβασης: 27-05-2022.

[20] Ismail Butun, Salvatore D. Morgera και Ravi Sankar. *A survey of intrusion detection systems in wireless sensor networks*. *IEEE communications surveys & tutorials*, 16(1):266–282, 2014. https://www.doi.org/10.1109/SURV.2013.050113.00191.

[21] Ansam Khraisat, Iqbal Gondal, Peter Vamplew και Joarder Kamruzzaman. *Survey of intrusion detection systems: techniques, datasets and challenges*. *Cybersecurity*, 2(1):1–22, 2019. https://www.doi.org/10.1186/s42400-019-0038-7.

[22] Stefan Axelsson. *Intrusion detection systems: A survey and taxonomy*. Τεχνική Αναφορά 99, Department of Computer Engineering, Chalmers University, Απρίλιος 2000.

[23] Hanan Hindy, David Brosset, Ethan Bayne, Amar Kumar Seeam, Christos Tachtatzis, Robert Atkinson και Xavier Bellekens. *A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems*. *IEEE Access*, 8:104650–104675, 2020. https://doi.org/10.1109/access.2020.3000179.

[24] Ansam Khraisat, Iqbal Gondal και Peter Vamplew. *An anomaly intrusion detection system using C5 decision tree classifier*. *Trends and Applications in Knowledge Discovery and Data Mining*, σελίδες 149–155. Springer, 2018. https://doi.org/10.1007/978-3-030-04503-6_14.

[25] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu και Sen Zhao. *Advances and Open Problems in Federated Learning*. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. https://doi.org/10.1561/2200000083.

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson και Blaise Aguera y Arcas. *Communication-efficient learning of deep networks from decentralized data*. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, τόμος 54, σελίδες 1273–1282. PMLR, 2017.

[27] Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin και Karl H. Johansson. *A survey of distributed optimization*. *Annual Reviews in Control*, 47:278–305, 2019. https://doi.org/10.1016/j.arcontrol.2019.05.006.

[28] Jakub Konečnỳ, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh και Dave Bacon. *Federated learning: Strategies for improving communication efficiency*. *arXiv preprint arXiv:1610.05492v2*, 2016. https://doi.org/10.48550/arXiv.1610.05492.

[29] Qiang Yang, Yang Liu, Tianjian Chen και Yongxin Tong. *Federated machine learning: Concept and applications*. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. https://doi.org/10.1145/3298981.

[30] Li Li, Yuxi Fan, Mike Tse και Kuo-Yi Lin. *A review of applications in federated learning*. *Computers & Industrial Engineering*, 149:106854, 2020. https://doi.org/10.1016/j.cie.2020.106854.

[31] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen και Qiang Yang. *A secure federated transfer learning framework*. *IEEE Intelligent Systems*, 35(4):70–82, 2020.

[32] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang και Zheng Chen. *Cross-Domain Sentiment Classification via Spectral Feature Alignment*. *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, σελίδα 751–760, New York, NY, USA, 2010. Association for Computing Machinery. https://doi.org/10.1145/1772690.1772767.

[33] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha και Gautam Srivastava. *A survey on security and privacy of federated learning*. *Future Generation Computer Systems*, 115:619–640, 2021. https://doi.org/10.1016/j.future.2020.10.007.

[34] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu και Bingsheng He. *A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection*. *IEEE Transactions on Knowledge and Data Engineering*, 2021. https://doi.org/10.1109/tkde.2021.3124599.

[35] Latanya Sweeney. *Simple demographics often identify people uniquely*. *Health (San Francisco)*, 671(2000):1–34, 2000.

[36] Tian Li, Anit Kumar Sahu, Ameet Talwalkar και Virginia Smith. *Federated Learning: Challenges, Methods, and Future Directions*. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. https://doi.org/10.1109/MSP.2020.2975749.

[37] Takayuki Nishio και Ryo Yonetani. *Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge*. *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*. IEEE, Μάιος 2019. https://doi.org/10.1109/icc.2019.8761315.

[38] Jie Xu και Heqiang Wang. *Client Selection and Bandwidth Allocation in Wireless Federated Learning Networks: A Long-Term Perspective*. *IEEE Transactions on Wireless Communications*, 20(2):1188–1200, 2021. https://doi.org/10.1109/TWC.2020.3031503.

[39] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin και Vikas Chandra. *Federated Learning with Non-IID Data*. *arXiv preprint arXiv:1806.00582v1*, 2018. https://doi.org/10.48550/arXiv.1806.00582.

[40] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller και Wojciech Samek. *Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data*. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2020. https://doi.org/10.1109/TNNLS.2019.2944481.

[41] Hao Wang, Zakhary Kaplan, Di Niu και Baochun Li. *Optimizing Federated Learning on Non-IID Data with Reinforcement Learning*. *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, σελίδες 1698–1707, 2020. https://doi.org/10.1109/INFOCOM41043.2020.9155494.

[42] Hyesung Kim, Jihong Park, Mehdi Bennis και Seong-Lyun Kim. *Blockchained On-Device Federated Learning*. *IEEE Communications Letters*, 24(6):1279–1283, 2020. https://doi.org/10.1109/LCOMM.2019.2921755.

[43] Stefano Savazzi, Monica Nicoli και Vittorio Rampa. *Federated Learning With Cooperating Devices: A Consensus Approach for Massive IoT Networks*. *IEEE Internet of Things Journal*, 7(5):4641–4654, 2020. https://doi.org/10.1109/JIOT.2020.2964162.

[44] Chuan Ma, Jun Li, Ming Ding, Howard H. Yang, Feng Shu, Tony Q.S. Quek και H. Vincent Poor. *On safeguarding privacy and security in the framework of federated learning*. *IEEE network*, 34(4):242–248, 2020. https://doi.org/10.1109/MNET.001.1900506.

[45] Lingjuan Lyu, Han Yu και Qiang Yang. *Threats to federated learning: A survey*. *arXiv preprint arXiv:2003.02133v1*, 2020. https://doi.org/10.48550/arXiv.2003.02133.

[46] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal και Seraphin Calo. *Analyzing Federated Learning through an Adversarial Lens*. *Proceedings of the 36th International Conference on Machine Learning*, επιμελητές Kamalika Chaudhuri και Ruslan Salakhutdinov, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 634–643. PMLR, Ιούνιος 2019.

[47] Ligeng Zhu, Zhijian Liu και Song Han. *Deep Leakage from Gradients*. *Advances in Neural Information Processing Systems*, επιμελητές H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox και R. Garnett, τόμος 32. Curran Associates, Inc., 2019.

[48] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge και Michael Moeller. *Inverting Gradients - How easy is it to break privacy in federated learning?* *Advances in Neural Information Processing Systems*, επιμελητές H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, τόμος 33, σελίδες 16937–16947. Curran Associates, Inc., 2020.

[49] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek και H. Vincent Poor. *Federated Learning With Differential Privacy: Algorithms and Performance Analysis*. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020. https://doi.org/10.1109/TIFS.2020.2988575.

[50] Meng Hao, Hongwei Li, Guowen Xu, Sen Liu και Haomiao Yang. *Towards Efficient and Privacy-Preserving Federated Deep Learning*. *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, σελίδες 1–6, 2019. https://doi.org/10.1109/ICC.2019.8761267.

[51] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang και Yi Zhou. *A Hybrid Approach to Privacy-Preserving Federated Learning*. *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, AISec'19, σελίδες 1–11, New York, NY, USA, 2019. Association for Computing Machinery. https://doi.org/10.1145/3338501.3357370.

[52] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar και Li Zhang. *Deep Learning with Differential Privacy*. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, σελίδες 308–318, New York, NY, USA, 2016. Association for Computing Machinery. https://doi.org/10.1145/2976749.2978318.

[53] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith και Brian Thorne. *Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption*. *arXiv preprint arXiv:1711.10677v1*, 2017. https://doi.org/10.48550/arXiv.1711.10677.

[54] Chih Fong Tsai, Yu Feng Hsu, Chia Ying Lin και Wei Yang Lin. *Intrusion detection by machine learning: A review*. *Expert Systems with Applications*, 36(10):11994–12000, 2009. https://doi.org/10.1016/j.eswa.2009.05.029.

[55] Wun-Hwa Chen, Sheng-Hsun Hsu και Hwang-Pin Shen. *Application of SVM and ANN for intrusion detection*. *Computers & Operations Research*, 32(10):2617–2634, 2005. https://doi.org/10.1016/j.cor.2004.03.019. Applications of Neural Networks.

[56] Srinivas Mukkamala, Guadalupe Janoski και Andrew Sung. *Intrusion detection using neural networks and support vector machines*. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, τόμος 2, σελίδες 1702–1707. IEEE, 2002. https://doi.org/10.1109/IJCNN.2002.1007774.

[57] Kinan Ghanem, Francisco J. Aparicio-Navarro, Konstantinos G. Kyriakopoulos, Sangarapillai Lambotharan και Jonathon A. Chambers. *Support Vector Machine for Network Intrusion and Cyber-Attack Detection*. *2017 Sensor Signal Processing for Defence Conference (SSPD)*, σελίδες 1–5, 2017. https://doi.org/10.1109/SSPD.2017.8233268.

[58] Yihua Liao και V. Rao Vemuri. *Use of K-Nearest Neighbor classifier for intrusion detection*. *Computers & Security*, 21(5):439–448, 2002. https://doi.org/10.1016/S0167-4048(02)00514-X.

[59] Zhenghui Ma και Ata Kaban. *K-Nearest-Neighbours with a novel similarity measure for intrusion detection*. *2013 13th UK Workshop on Computational Intelligence (UKCI)*, σελίδες 266–271, 2013. https://doi.org/10.1109/UKCI.2013.6651315.

[60] Ozgur Depren, Murat Topallar, Emin Anarim και M. Kemal Ciliz. *An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks*. *Expert Systems with Applications*, 29(4):713–722, 2005. https://doi.org/10.1016/j.eswa.2005.05.002.

[61] Nabila Farnaaz και M.A. Jabbar. *Random Forest Modeling for Network Intrusion Detection System*. *Procedia Computer Science*, 89:213–217, 2016. https://doi.org/10.1016/j.procs.2016.06.047. Twelfth International Conference on Communication Networks, ICCN 2016, August 19– 21, 2016, Bangalore, India Twelfth International

Conference on Data Mining and Warehousing, ICDMW 2016, August 19-21, 2016, Bangalore, India Twelfth International Conference on Image and Signal Processing, ICISP 2016, August 19-21, 2016, Bangalore, India.

[62] R. Kumari, Sheetanshu, M. K. Singh, R. Jha και N.K. Singh. *Anomaly detection in network traffic using K-mean clustering*. *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, σελίδες 387–393, 2016. https://doi.org/10.1109/RAIT.2016.7507933.

[63] Alhamza Alalousi, Rozmie Razif, Mosleh Abualhaj, Mohammed Anbar και Shahrul Nizam. *A preliminary performance evaluation of K-means, KNN and EM unsupervised machine learning methods for network flow classification*. *International Journal of Electrical and Computer Engineering*, 6(2):778–784, 2016. https://doi.org/10.11591/ijece.v6i2.pp778-784.

[64] Javed Asharf, Nour Moustafa, Hasnat Khurshid, Essam Debie, Waqas Haider και Abdul Wahab. *A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions*. *Electronics*, 9(7), 2020. https://doi.org/10.3390/electronics9071177.

[65] Arwa Aldweesh, Abdelouahid Derhab και Ahmed Z. Emam. *Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues*. *Knowledge-Based Systems*, 189:105124, 2020. https://doi.org/10.1016/j.knosys.2019.105124.

[66] R. Vinayakumar, Mamoun Alazab, K. P. Soman, Prabaharan Poornachandran, Ameer Al-Nemrat και Sitalakshmi Venkatraman. *Deep Learning Approach for Intelligent Intrusion Detection System*. *IEEE Access*, 7:41525–41550, 2019. https://doi.org/10.1109/ACCESS.2019.2895334.

[67] Zheng Wang. *Deep Learning-Based Intrusion Detection With Adversaries*. *IEEE Access*, 6:38367–38384, 2018. https://doi.org/10.1109/ACCESS.2018.2854599.

[68] *NSL-KDD Dataset*. https://www.unb.ca/cic/datasets/nsl.html. Ημερομηνία πρόσβασης: 06-06-2022.

[69] Chuanlong Yin, Yuefei Zhu, Jinlong Fei και Xinzheng He. *A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks*. *IEEE Access*, 5:21954–21961, 2017. https://doi.org/10.1109/ACCESS.2017.2762418.

[70] Jihyun Kim, Jaehyun Kim, Huong Le Thi Thu και Howon Kim. *Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection*. *2016 International Conference on Platform Technology and Service (PlatCon)*, σελίδες 1–5, 2016. https://doi.org/10.1109/PlatCon.2016.7456805.

[71] Ralf C. Staudemeyer. *Applying long short-term memory recurrent neural networks to intrusion detection*. *South African Computer Journal*, 56(1):136–154, 2015.

[72] Baraa Ismael Farhan και Ammar D. Jasim. *Performance analysis of intrusion detection for deep learning model based on CSE-CIC-IDS2018 dataset*. *Indonesian Journal of Electrical Engineering and Computer Science*, 26:1165–1172, Μάιος 2022. https://doi.org/10.11591/ijeecs.v26.i2.pp1165-1172.

[73] Congyuan Xu, Jizhong Shen, Xin Du και Fan Zhang. *An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units*. *IEEE Access*, 6:48697–48707, 2018. https://doi.org/10.1109/ACCESS.2018.2867564.

[74] Wen-Hui Lin, Hsiao-Chung Lin, Ping Wang, Bao-Hua Wu και Jeng-Ying Tsai. *Using convolutional neural networks to network intrusion detection for cyber threats*. *2018 IEEE International Conference on Applied System Invention (ICASI)*, σελίδες 1107–1110, 2018. https://doi.org/10.1109/ICASI.2018.8394474.

[75] Md Moin Uddin Chowdhury, Frederick Hammond, Glenn Konowicz, Chunsheng Xin, Hongyi Wu και Jiang Li. *A few-shot deep learning approach for improved intrusion detection*. *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, σελίδες 456–462. IEEE, 2017. https://doi.org/10.1109/UEMCON.2017.8249084.

[76] Jiyeon Kim, Jiwon Kim, Hyunjung Kim, Minsun Shim και Eunjung Choi. *CNN-based network intrusion detection against denial-of-service attacks*. *Electronics*, 9(6):916, 2020. https://doi.org/10.3390/electronics9060916.

[77] Fahimeh Farahnakian και Jukka Heikkonen. *A deep auto-encoder based approach for intrusion detection system*. *2018 20th International Conference on Advanced Communication Technology (ICACT)*, σελίδες 178–183, 2018. https://doi.org/10.23919/ICACT.2018.8323688.

[78] Sasanka Potluri και Christian Diedrich. *Accelerated deep neural networks for enhanced Intrusion Detection System*. *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*, σελίδες 1–8, 2016. https://doi.org/10.1109/ETFA.2016.7733515.

[79] Nathan Shone, Tran Nguyen Ngoc, Vu Dinh Phai και Qi Shi. *A Deep Learning Approach to Network Intrusion Detection*. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1):41–50, 2018. https://doi.org/10.1109/TETCI.2017.2772792.

[80] Sunanda Gamage και Jagath Samarabandu. *Deep learning methods in network intrusion detection: A survey and an objective comparison*. *Journal of Network and Computer Applications*, 169:102767, 2020. https://doi.org/10.1016/j.jnca.2020.102767.

[81] Muhamad Erza Aminanto, Rakyong Choi, Harry Chandra Tanuwidjaja, Paul D. Yoo και Kwangjo Kim. *Deep Abstraction and Weighted Feature Selection for Wi-Fi Impersonation Detection*. *IEEE Transactions on Information Forensics and Security*, 13(3):621–636, 2018. https://doi.org/10.1109/TIFS.2017.2762828.

[82] Baoan Zhang, Yanhua Yu και Jie Li. *Network Intrusion Detection Based on Stacked Sparse Autoencoder and Binary Tree Ensemble Method*. *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, σελίδες 1–6, 2018. https://doi.org/10.1109/ICCW.2018.8403759.

[83] Md. Zahangir Alom, VenkataRamesh Bontupalli και Tarek M. Taha. *Intrusion detection using deep belief networks*. *2015 National Aerospace and Electronics Conference (NAECON)*, σελίδες 339–344, 2015. https://doi.org/10.1109/NAECON.2015.7443094.

[84] Ying Zhang, Peisong Li και Xinheng Wang. *Intrusion Detection for IoT Based on Improved Genetic Algorithm and Deep Belief Network*. *IEEE Access*, 7:31711–31722, 2019. https://doi.org/10.1109/ACCESS.2019.2903723.

[85] Joffrey L. Leevy και Taghi M. Khoshgoftaar. *A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 big data*. *Journal of Big Data*, 7(1):1–19, 2020. https://doi.org/10.1186/s40537-020-00382-x.

[86] Baraa I. Farhan και Ammar D. Jasim. *A Survey of Intrusion Detection Using Deep Learning in Internet of Things*. *Iraqi Journal For Computer Science and Mathematics*, 3(1):83–93, Ιανουάριος 2022. https://doi.org/10.52866/ijcsm.2022.01.01.009.

[87] Peng Lin, Kejiang Ye και Cheng-Zhong Xu. *Dynamic Network Anomaly Detection System by Using Deep Learning Techniques*. *Cloud Computing - CLOUD 2019*, επιμελητές Dilma Da Silva, Qingyang Wang και Liang Jie Zhang, σελίδες 161–176, Cham, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-23502-4_12.

[88] Muhammad Ashfaq Khan και Juntae Kim. *Toward Developing Efficient Conv-AE-Based Intrusion Detection System Using Heterogeneous Dataset*. *Electronics*, 9(11), 2020. https://doi.org/10.3390/electronics9111771.

[89] Jiyeon Kim, Yulim Shin και Eunjung Choi. *An intrusion detection model based on a convolutional neural network*. *Journal of Multimedia Information System*, 6(4):165–172, 2019. https://doi.org/10.33851/jmis.2019.6.4.165.

[90] Marta Catillo, Massimiliano Rak και Umberto Villano. *2L-ZED-IDS: A Two-Level Anomaly Detector for Multiple Attack Classes*. *Web, Artificial Intelligence and Network Applications*, επιμελητές Leonard Barolli, Flora Amato, Francesco Moscato, Tomoya Enokido και Makoto Takizawa, σελίδες 687–696, Cham, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-44038-1_63.

[91] Feng Zhao, Hao Zhang, Jia Peng, Xiaohong Zhuang και Sang-Gyun Na. *A semi-self-taught network intrusion detection system*. *Neural Computing and Applications*, 32(23):17169–17179, 2020. https://doi.org/10.1007/s00521-020-04914-7.

[92] Enrique Mármol Campos, Pablo Fernández Saura, Aurora González-Vidal, José L. Hernández-Ramos, Jorge Bernal Bernabé, Gianmarco Baldini και Antonio Skarmeta. *Evaluating Federated Learning for intrusion detection in Internet of Things: Review and challenges*. *Computer Networks*, 203:108661, 2022. https://doi.org/10.1016/j.comnet.2021.108661.

[93] Aitor Belenguer, Javier Navaridas και Jose A. Pascual. *A review of Federated Learning in Intrusion Detection Systems for IoT*. *arXiv preprint arXiv:2204.12443v2*, 2022. https://doi.org/10.48550/arXiv.2204.12443.

[94] Francisco Assis Moreirado Nascimento και Fabiano Hessel. *A Decentralized Federated Learning Architecture for Intrusion Detection in IoT Systems*. *Advanced Information Networking and Applications*, επιμελητές Leonard Barolli, Farookh Hussain και Tomoya Enokido, σελίδες 256–268, Cham, 2022. Springer International Publishing. https://doi.org/10.1007/978-3-030-99587-4_22.

[95] Dapeng Man, Fanyi Zeng, Wu Yang, Miao Yu, Jiguang Lv και Yijing Wang. *Intelligent intrusion detection based on federated learning for edge-assisted internet of things*. *Security and Communication Networks*, 2021. https://doi.org/10.1155/2021/9361348.

[96] Othmane Friha, Mohamed Amine Ferrag, Lei Shu, Leandros Maglaras, Kim-Kwang Raymond Choo και Mehdi Nafaa. *FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things*. *Journal of Parallel and Distributed Computing*, 165:17–31, 2022. https://doi.org/10.1016/j.jpdc.2022.03.003.

[97] Ajesh Koyatan Chathoth, Abhyuday Jagannatha και Stephen Lee. *Federated Intrusion Detection for IoT with Heterogeneous Cohort Privacy*. *arXiv preprint arXiv:2101.09878v1*, 2021. https://doi.org/10.48550/arXiv.2101.09878.

[98] Jiachao Zhang, Peiran Yu, Le Qi, Song Liu, Haiyu Zhang και Jianzhong Zhang. *FLDDoS: DDoS Attack Detection Model based on Federated Learning*. *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, σελίδες 635–642, 2021. https://doi.org/10.1109/TrustCom53373.2021.00095.

[99] Jingyi Li, Zikai Zhang, Yidong Li, Xinyue Guo και Huifang Li. *FIDS: Detecting DDoS Through Federated Learning Based Method*. *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, σελίδες 856–862, 2021. https://doi.org/10.1109/TrustCom53373.2021.00121.

[100] Roberto Doriguzzi-Corin και Domenico Siracusa. *FLAD: Adaptive Federated Learning for DDoS Attack Detection*. *arXiv preprint arXiv:2205.06661v2*, 2022. https://doi.org/10.48550/arxiv.2205.06661.

[101] Parya Haji Mirzaee, Mohammad Shojafar, Zahra Pooranian, Pedram Asefy, Haitham Cruickshank και Rahim Tafazolli. *FIDS: A Federated Intrusion Detection System for*

5G Smart Metering Network. *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, σελίδες 215–222, 2021. https://doi.org/10.1109/MSN53354.2021.00044.

[102] Beibei Li, Yuhao Wu, Jiarui Song, Rongxing Lu, Tao Li και Liang Zhao. *DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber–Physical Systems*. *IEEE Transactions on Industrial Informatics*, 17(8):5615–5624, 2021. https://doi.org/10.1109/TII.2020.3023430.

[103] Tian Dong, Song Li, Han Qiu και Jialiang Lu. *An Interpretable Federated Learning-based Network Intrusion Detection Framework*. *arXiv preprint arXiv:2201.03134v1*, 2022. https://doi.org/10.48550/arXiv.2201.03134.

[104] Burak Cetin, Alina Lazar, Jinoh Kim, Alex Sim και Kesheng Wu. *Federated Wireless Network Intrusion Detection*. *2019 IEEE International Conference on Big Data (Big Data)*, σελίδες 6004–6006, 2019. https://doi.org/10.1109/BigData47090.2019.9005507.

[105] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou και Stefanos Gritzalis. *Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset*. *IEEE Communications Surveys & Tutorials*, 18(1):184–208, 2016. https://doi.org/10.1109/COMST.2015.2402161.

[106] Ana Cholakoska, Bjarne Pfitzner, Hristijan Gjoreski, Valentin Rakovic, Bert Arnrich και Marija Kalendar. *Differentially Private Federated Learning for Anomaly Detection in EHealth Networks*. *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, σελίδες 514–518, New York, NY, USA, 2021. Association for Computing Machinery. https://doi.org/10.1145/3460418.3479365.

[107] Pavel Laskov, Patrick Düssel, Christin Schäfer και Konrad Rieck. *Learning Intrusion Detection: Supervised or Unsupervised?* *Image Analysis and Processing - ICIAP 2005*, επιμελητές Fabio Roli και Sergio Vitulano, σελίδες 50–57, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. https://doi.org/10.1007/11553595_6.

[108] *A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)*. https://registry.opendata.aws/cse-cic-ids2018/. Ημερομηνία πρόσβασης: 23-05-2022.

[109] *CSE-CIC-IDS2018 on AWS*. https://www.unb.ca/cic/datasets/ids-2018.html. Ημερομηνία πρόσβασης: 23-05-2022.

[110] *CICFlowMeter (formerly ISCXFlowMeter)*. https://www.unb.ca/cic/research/applications.html#CICFlowMeter. Ημερομηνία πρόσβασης: 23-05-2022.

[111] *CVE-2014-0160 Detail*. https://www.cve.org/CVERecord?id=CVE-2014-0160. Ημερομηνία πρόσβασης: 23-05-2022.

[112] *The Heartbleed Bug*. https://heartbleed.com. Ημερομηνία πρόσβασης: 23-05-2022.

[113] *NumPy*. https://numpy.org. Ημερομηνία πρόσβασης: 24-05-2022.

[114] *pandas*. https://pandas.pydata.org. Ημερομηνία πρόσβασης: 24-05-2022.

[115] *scikit-learn*. https://scikit-learn.org. Ημερομηνία πρόσβασης: 24-05-2022.

[116] *Dask*. https://dask.org. Ημερομηνία πρόσβασης: 24-05-2022.

[117] *TensorFlow*. https://www.tensorflow.org. Ημερομηνία πρόσβασης: 20-06-2022.

[118] *Keras*. https://keras.io. Ημερομηνία πρόσβασης: 20-06-2022.

[119] *TensorFlow Federated: Machine Learning on Decentralized Data*. https://www.tensorflow.org/federated. Ημερομηνία πρόσβασης: 20-06-2022.

[120] Diederik P. Kingma και Jimmy Ba. *Adam: A Method for Stochastic Optimization*. *arXiv preprint arXiv:1412.6980v9*, 2014. https://doi.org/10.48550/arXiv.1412.6980.