Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών *Ασφάλεια Υπολογιστών και Δικτύων*

Μεταπτυχιακή Διατριβή

Security and Privacy Issues in Federated Learning Systems

Τέγγερης Ξάνθος

Επιβλέπων Καθηγητής

Μαυρίδης Ιωάννης

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών *Ασφάλεια Υπολογιστών και Δικτύων*

Μεταπτυχιακή Διατριβή

Security and Privacy Issues in Federated Learning Systems

Τέγγερης Ξάνθος

Επιβλέπων Καθηγητής

Μαυρίδης Ιωάννης

Περίληψη

Η έννοια του Federated Learning (FL) περιλαμβάνει την κατάρτιση στατιστικών μοντέλων σε κέντρα δεδομένων, όπως κινητά τηλέφωνα ή νοσοκομεία, διατηρώντας ταυτόχρονα τα δεδομένα τοπικά. Η εκπαίδευση σε ετερογενή και δυνητικά τεράστια δίκτυα εισάγει πολλές προκλήσεις που απορρέουν από την απόκλιση από το βασικό πρότυπο προσέγγισης για μεγάλης κλίμακας μηχανική μάθηση (machine learning), κατανεμημένης βελτιστοποίησης και ανάλυσης δεδομένων για τη διατήρηση της ιδιωτικής ζωής.

Το federated Learning (FL) περιλαμβάνει την αποκεντρωμένη μηχανική μάθηση σε ενιαίο στατιστικό μοντέλο από δεδομένα αποθηκευμένα σε δεκάδες έως δυνητικά εκατομμύρια απομακρυσμένες συσκευές. Η επικοινωνία είναι ένα κρίσιμο σημείο όσο αφορά την ασφάλεια του συστήματος για τα ενοποιημένα δίκτυα, τα οποία όταν συνδέονται με ζητήματα σχετικά με το απόρρητο ή με την ασφαλή αποστολή πρωτογενή δεδομένα απαιτούν την παραμονή τους στη συσκευή όπου δημιουργούνται.

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η μελέτη της σχετικής βιβλιογραφίας και η κωδικοποίηση των ζητημάτων ασφάλειας και ιδιωτικότητα των συστημάτων Federated Learning (FL). Επιπρόσθετα, γίνεται πρόταση και συζήτηση λύσεων για την αντιμετώπιση αυτών των θεμάτων. Στη συνέχεια, γίνεται ανάπτυξη πειραματικής εφαρμογής για τη μελέτη στη βάση σεναρίων για την αντιμετώπιση του data poisoning με βάση το model averaging. Τέλος, τα βασικά ερευνητικά ερωτήματα είναι η εύρεση των απειλών για την ασφάλεια, για την ιδιωτικότητα αλλά και τα κριτήρια που αξιολογούνται οι παραπάνω απειλές. Επιπρόσθετα, ποιά είναι τα χαρακτηριστικά των διαφόρων συστημάτων FL που επηρεάζουν τη διαμόρφωση του μοντέλου απειλών αλλά και οι τρόποι με τους οποίους μπορούν να αντιμετωπιστούν οι παραπάνω απειλές.

Abstract

The concept of Federated Learning (FL) involves compiling statistical models on data centers, such as cell phones or hospitals, while keeping the data local. Training in heterogeneous and potentially vast networks introduces many challenges arising from deviating from the basic standard approach to large-scale machine learning, distributed optimization, and data analysis for privacy.

FL includes decentralized machine learning in a single statistical model from data stored on tens to potentially millions of remote devices. Communication is a critical bottleneck for integrated networks which, when linked to privacy issues or the secure sending of primary data, requires them to remain on the device where they are created. Thus, the object of the present dissertation is the study of the relevant literature and codification of the security and privacy issues of the Federated Learning (FL) systems. In addition, solutions are proposed and discussed to address these issues. Then, an experimental application is developed for the study based on scenarios for the treatment of data poisoning based on the model averaging. Finally, the main research questions are to find the threats to security, privacy, and the criteria by which the above threats are evaluated. In addition, what are the characteristics of the various FL systems that affect the configuration of the threat model and the ways in which the above threats can be addressed.

Ευχαριστίες

Για τη διεκπεραίωση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Μαυρίδη Ιωάννη για τη συνεργασία του. Επίσης, όλοι οι καθηγητές του τμήματος με καθοδήγησαν και μου έδωσαν όλα τα εφόδια για την αρχή της σταδιοδρομίας μου. Τέλος, ευχαριστώ την οικογένεια μου για την ηθική της υποστήριξη.

Κατάλογος Εικόνων

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

Οι τρέχουσες εξελίξεις στο χώρο της τεχνητής νοημοσύνης ανέδειξαν τη σπουδαιότητα των συστημάτων Federated Machine Learning (FL) ως μια πολλά υποσχόμενη λύση στις προκλήσεις που αντιμετωπίζουν τα κλασσικά συγκεντρωτικά συστήματα Machine Learning (ML). Από τις κυριότερες αναφέρονται δυο: η πρώτη αφορά το γεγονός ότι, στις περισσότερες βιομηχανίες, τα δεδομένα διατηρούνται σε «απομονωμένα νησιά» (isolated islands), ενώ η δεύτερη αφορά την ενίσχυση του απορρήτου και της ασφάλειας των δεδομένων.

Τα συστήματα Federated Learning παρέχουν εκ πρώτης όψεως λύσεις στα θέματα της ιδιωτικότητα των πρωτογενή δεδομένων, ξεπερνώντας σημαντικούς περιορισμούς που θέτει η νομοθεσία και οι διεθνείς κανονισμοί. Ωστόσο, από τη βιβλιογραφία γίνεται φανερό ότι τα συστήματα FL είναι ευάλωτα σε επιθέσεις όπως poisoning και inference, οι οποίες θα μπορούσαν να προέρχονται από οποιοδήποτε μέρος κατά τη διάρκεια της διαδικασίας ενοποιημένης βελτιστοποίησης.

Η έννοια του Federated Learning (FL) περιλαμβάνει την κατάρτιση στατιστικών μοντέλων σε κέντρο δεδομένων, όπως κινητά τηλέφωνα ή νοσοκομεία, διατηρώντας ταυτόχρονα τα δεδομένα τοπικά. Η εκπαίδευση σε ετερογενή και δυνητικά τεράστια δίκτυα εισάγει πολλές προκλήσεις που απορρέουν από την απόκλιση από το βασικό πρότυπο προσέγγισης για μεγάλης κλίμακας μηχανική μάθηση (machine learning), κατανεμημένη βελτιστοποίηση (Li, 2020).

Το FL περιλαμβάνει την αποκεντρωμένη μηχανική μάθηση σε ενιαίο στατιστικό μοντέλο από δεδομένα αποθηκευμένα σε δεκάδες έως δυνητικά εκατομμύρια απομακρυσμένες συσκευές. Η επικοινωνία είναι ένα κρίσιμο σημείο συμφόρησης για τα ενοποιημένα δίκτυα (Bonawitz, 2019) τα οποία, όταν συνδέονται με ζητήματα σχετικά με το απόρρητο ή με την ασφαλή αποστολή στα πρωτογενή δεδομένα απαιτεί την παραμονή τους στη συσκευή όπου δημιουργούνται.

Τα κινητά τηλέφωνα, οι φορητές συσκευές και τα αυτόνομα οχήματα είναι μερικά μόνο από τα σύγχρονα κατανεμημένα δίκτυα που δημιουργούν έναν πλούτο δεδομένων κάθε μέρα. Λόγω της αυξανόμενης υπολογιστικής δύναμης αυτών των συσκευών, σε συνδυασμό με τις ανησυχίες για τη μετάδοση ιδιωτικών πληροφοριών, άρχισε να γίνεται όλο και πιο ελκυστικό να αποθηκεύονται τα δεδομένα τοπικά. Οι απλοί υπολογισμοί σε κατανεμημένες, συσκευές χαμηλής ισχύος είναι ένας τομέας έρευνας εδώ και δεκαετίες (Bonomi, 2012). Πρόσφατες εργασίες έχουν εξετάσει κεντρικά μοντέλα μηχανικής μάθησης που αποθηκεύουν τοπικά (Kuflik, 2012).

Όσο οι υπολογιστικές δυνατότητες των συσκευών εντός των κατανεμημένων δικτύων μεγαλώνουν, είναι δυνατόν να αξιοποιηθούν βελτιωμένοι τοπικοί πόροι σε κάθε συσκευή. Επιπλέον, οι ανησυχίες σχετικά με το απόρρητο και τη μετάδοση ανεπεξέργαστων δεδομένων απαιτούν να παραμείνουν αυτά σε τοπικές συσκευές. Αυτό οδήγησε σε ένα αυξανόμενο ενδιαφέρον για το Federated Learning (McMahan, 2017), το οποίο διερευνά την εκπαίδευση στατιστικών μοντέλων απευθείας σε απομακρυσμένες συσκευές. Ο όρος FL έχει αναπτυχθεί στην πράξη από μεγάλες εταιρείες (Sheller, 2018) και παίζουν κρίσιμο ρόλο στην υποστήριξη εφαρμογών ευαίσθητων στην ιδιωτική ζωή όπου τα δεδομένα εκπαίδευσης διανέμονται με βάση τη θεωρία του “computing at the edge” (Brisimi, 2018).

Με βάση όλα τα παραπάνω οι κύριοι άξονες της μεταπτυχιακής διατριβής είναι:

α) η μελέτη της σχετικής με FL βιβλιογραφίας.

β) η κωδικοποίηση και αξιολόγηση των ζητημάτων ασφαλείας και ιδιωτικότητας των συστημάτων FL.

γ) η ανάπτυξη πειραματικής εφαρμογής για τη μελέτη στη βάση σεναρίων αντιμετώπισης του data poisoning με βάση το model averaging.

δ) η πρόταση λύσεων για την αντιμετώπιση των παραπάνω ζητημάτων.

Έτσι, τα βασικά ερευνητικά ερωτήματα που προκύπτουν είναι:

- Ποιες είναι οι απειλές για την ασφάλεια και την ιδιωτικότητα στα συστήματα FL;
- Ποια είναι τα κριτήρια αξιολόγησης των παραπάνω απειλών;
- Ποιά είναι τα χαρακτηριστικά των διαφόρων συστημάτων FL που επηρεάζουν τη διαμόρφωση του μοντέλου απειλών;
- Πώς μπορούν να αντιμετωπιστούν οι παραπάνω απειλές;

Η καινοτομία της μεταπτυχιακής διατριβής αφορά τη μελέτη των συστημάτων FL υπό το πρίσμα των ζητημάτων ασφάλειας και ιδιωτικότητας.

Κεφάλαιο 2: Νέα Δεδομένα στα συστήματα FL

2.1 Εισαγωγή

Την τελευταία δεκαετία η τεχνολογία της μηχανικής μάθησης (Machine Learning - ML) έχει ανθίσει και πολλές και ποικίλες εφαρμογές τεχνητής νοημοσύνης (Artificial Intelligence - AI) έχουν δημιουργηθεί, όπως η τεχνητή «όραση υπολογιστή», η αυτόματη αναγνώριση ομιλίας, η επεξεργασία φυσικής γλώσσας(Pouyanfar et al., 2019, Hatcher and Yu, 2018, Goodfellow et al., 2016). Η επιτυχία αυτών των τεχνολογιών μηχανικής μάθησης, ειδικότερα η «βαθιά μάθηση» (Deep Learning - DL), τροφοδοτήθηκε από τη διαθεσιμότητα τεράστιων ποσοτήτων δεδομένων (Trask, 2019, Pouyanfar et al., 2019, Hatcher and Yu, 2018). Χρησιμοποιώντας αυτά τα δεδομένα, τα συστήματα DL μπορούν να εκτελέσουν μια ποικιλία εργασιών που μπορεί μερικές φορές να υπερβαίνουν την ανθρώπινη απόδοση.

Για παράδειγμα, τα ενισχυμένα συστήματα αναγνώρισης προσώπου μπορούν να επιτύχουν εμπορικά αποδεκτά επίπεδα απόδοσης με την χρήση εκατομμυρίων εικόνων εκπαίδευσης. Αυτά τα συστήματα απαιτούν συνήθως έναν τεράστιο όγκο δεδομένων για να επιτευχθεί ικανοποιητικό επίπεδο απόδοσης. Για παράδειγμα, το σύστημα ανίχνευσης αντικειμένων του Facebook έχει αναφερθεί ότι έχει εκπαιδευτεί με 3,5 δισεκατομμύρια εικόνες από το Instagram (Hartmann, 2019).

Γενικά, τα δεδομένα που απαιτούνται για την αύξηση των εφαρμογών τεχνητής νοημοσύνης είναι συχνά μεγάλα σε μέγεθος. Ωστόσο, σε πολλούς τομείς εφαρμογών, οι άνθρωποι έχουν διαπιστώσει ότι είναι δύσκολο να βρεθεί δείγμα «Big Data». Τις περισσότερες φορές το δείγμα δεν είναι μόνο μικρού μεγέθους αλλά και δεν περιέχει ορισμένες σημαντικές πληροφορίες, όπως η έλλειψη τιμών ή ετικετών. Η ύπαρξη επαρκών ετικετών για τα δεδομένα απαιτεί συχνά μεγάλη προσπάθεια από ειδικούς του

τομέα. Για παράδειγμα, στην ιατρική ανάλυση εικόνας, οι γιατροί συχνά απασχολούνται στο να παρέχουν διάγνωση με βάση τις εικόνες σάρωσης των οργάνων του ασθενούς, το οποίο ως διαδικασία είναι χρονοβόρα. Ως αποτέλεσμα, συχνά δεν μπορούν να ληφθούν δεδομένα εκπαίδευσης υψηλής ποιότητας και μεγάλου όγκου.

Η σύγχρονη κοινωνία ενημερώνεται όλο και περισσότερο για ζητήματα που αφορούν την ιδιοκτησία δεδομένων. Όπως, ποιος έχει το δικαίωμα να χρησιμοποιήσει τα δεδομένα για την κατασκευή τεχνολογιών AI. Σε μια υπηρεσία προϊόντων που βασίζεται σε τεχνολογία τεχνητής νοημοσύνης, ο κάτοχος της υπηρεσίας ισχυρίζεται ότι είναι ιδιοκτήτης των δεδομένων σχετικά με τα προϊόντα και τις συναλλαγές αγοράς, αλλά η κυριότητα των δεδομένων σχετικά με τις συμπεριφορές των χρηστών και τις συνήθειες πληρωμής είναι ασαφής. Επιπλέον, δεδομένου ότι τα δεδομένα δημιουργούνται και ανήκουν σε διαφορετικά μέρη και οργανισμούς, μια παραδοσιακή προσέγγιση είναι ότι η συλλογή και μεταφορά των δεδομένων επιτελείται σε μια κεντρική τοποθεσία όπου ισχυροί υπολογιστές μπορούν να εκπαιδεύσουν και να κατασκευάσουν μοντέλα ML.

Ενώ η τεχνολογία AI εξαπλώνεται σε συνεχώς διευρυμένους τομείς εφαρμογών, οι ανησυχίες σχετικά με το απόρρητο του χρήστη και την εμπιστευτικότητα των δεδομένων επεκτείνονται. Οι χρήστες ανησυχούν όλο και περισσότερο ότι οι προσωπικές τους πληροφορίες χρησιμοποιούνται (ή ακόμη ότι γίνεται κατάχρηση αυτών) για εμπορικούς και πολιτικούς σκοπούς χωρίς την άδειά τους. Πρόσφατα, πολλές μεγάλες εταιρείες του διαδικτύου έχουν υποστεί για αυτό το λόγο την επιβολή προστίμων λόγω διαρροής προσωπικών δεδομένων χρηστών με δική τους υπαιτιότητα σε άλλες εμπορικές εταιρείες.

Στο νομικό μέτωπο, οι νομοθέτες και τα ρυθμιστικά όργανα επινοούν νέους νόμους για το πώς πρέπει να διαχειρίζονται και να χρησιμοποιούνται τα δεδομένα αυτά. Ένα σημαντικό παράδειγμα είναι η υιοθέτηση του Κανονισμού Προστασίας Προσωπικών Δεδομένων (GDPR) από την Ευρωπαϊκή Ένωση (ΕΕ) το 2018 (GDPR, 2018). Στις ΗΠΑ, ο νόμος περί απορρήτου των καταναλωτών της Καλιφόρνια (CCPA) είναι σε ισχύ από το (DLA Piper, 2019).

Σε αυτό το νέο νομοθετικό τοπίο, η συλλογή και η ανταλλαγή δεδομένων μεταξύ διαφορετικών οργανισμών καθίσταται όλο και πιο δύσκολη, αν όχι απολύτως αδύνατη, καθώς περνά ο καιρός. Επιπλέον, απαγορεύει την διέλευση ορισμένων ευαίσθητων δεδομένων (π.χ. χρηματοοικονομικές συναλλαγές και ιατρικά αρχεία) (Yang et al., 2019). Λόγω του ανταγωνισμού που επικρατεί στον κλάδο της ασφάλειας των χρηστών, της ασφάλειας δεδομένων και των περίπλοκων διοικητικών διαδικασιών, ακόμη και τον συνδυασμό των δεδομένων μεταξύ των διαφόρων τμημάτων οι εταιρείες αντιμετωπίζουν ποικίλα προβλήματα. Το απαγορευτικά υψηλό κόστος καθιστά σχεδόν αδύνατη την ενοποίηση δεδομένων διασκορπισμένων σε διαφορετικά ιδρύματα (WeBank AI, 2019). Τώρα που ο παραδοσιακός ιδιωτικός τρόπος συλλογής και κοινής χρήσης δεδομένων δεν καλύπτει τις ανάγκες της αγοράς, η ενοποίηση δεδομένων, που περιλαμβάνει διαφορετικούς κατόχους δεδομένων είναι αναγκαία αν και είναι εξαιρετικά δύσκολο να ενοποιηθούν.

Η επίλυση του προβλήματος του κατακερματισμού και της απομόνωσης των δεδομένων είναι μια μεγάλη πρόκληση για τους ερευνητές και τους επαγγελματίες του κλάδου. Η αποτυχία αντιμετώπισης αυτού του προβλήματος θα οδηγήσει πιθανότατα σε νέες εξελίξεις και εφαρμογές του AI (Yang et al., 2019).

Ένας άλλος λόγος για τον οποίο η βιομηχανία τεχνητής νοημοσύνης αντιμετωπίζει μια δυσκολία στη διαχείριση των δεδομένων είναι ότι το όφελος της συνεργασίας κατά την κοινή χρήση των μεγάλων δεδομένων δεν είναι σαφές. Αν υποθέσουμε ότι επιθυμούν δύο οργανισμοί να συνεργαστούν πάνω σε ιατρικά δεδομένα προκειμένου να εκπαιδεύσουν ένα κοινό μοντέλο ML, η παραδοσιακή μέθοδος μεταφοράς των δεδομένων από έναν οργανισμό σε έναν άλλο, συχνά σημαίνει ότι ο κάτοχος των αρχικών δεδομένων θα χάσει τον έλεγχο των δεδομένων που είχε αρχικά. Παράλληλα, η τιμή των δεδομένων μειώνεται μόλις αυτά αλλάξουν κάτοχο.

Επιπλέον, όταν γίνεται ενσωμάτωση των δεδομένων που αποκτήθηκαν, δεν είναι σαφές πώς κατανέμεται δίκαια το όφελος μεταξύ των συμμετεχόντων. Με τον υπολογισμό αιχμής μέσω του διαδικτύου των δεδομένων, τα Big Data συχνά δεν είναι μια μονολιθική οντότητα, αλλά μάλλον διανέμονται σε πολλά μέρη. Για παράδειγμα, οι δορυφόροι που

λαμβάνουν εικόνες από τη Γη δεν μπορούν να περιμένουν να μεταδώσουν όλα τα δεδομένα σε κέντρα δεδομένων στο έδαφος, αφού η απαιτούμενη μετάδοση θα είναι πολύ μεγάλη. Ομοίως, στα αυτόνομα αυτοκίνητα, κάθε αυτοκίνητο πρέπει να είναι σε θέση για να επεξεργαστεί πολλές πληροφορίες τοπικά με μοντέλα ML, ενώ συνεργάζεται παγκοσμίως με άλλα αυτοκίνητα και υπολογιστικά κέντρα. Το πώς γίνεται η κοινή χρήση μοντέλων μεταξύ των πολλαπλών χρηστών με ασφαλή και αποτελεσματικό τρόπο είναι μια νέα πρόκληση για τους ερευνητές.

2.2 Η Αναγκαιότητα του FL

Όπως αναφέρθηκε προηγουμένως, πολλοί λόγοι καθιστούν το πρόβλημα της ανταλλαγής των δεδομένων εμπόδιο στη χρήση των Big Data που απαιτούνται για την εκπαίδευση μοντέλων ML. Είναι λοιπόν φυσικό να αναζητά η επιστημονική κοινότητα λύσεις για την κατασκευή μοντέλων ML που δεν βασίζονται στη συλλογή όλων των δεδομένων σε μια κεντρική αποθήκευση όπου μπορεί να εκπαιδευτεί το μοντέλο. Μια ιδέα είναι να εκπαιδεύσει κανείς ένα μοντέλο σε κάθε τοποθεσία όπου βρίσκεται μια πηγή δεδομένων και στη συνέχεια οι ιστότοποι να επικοινωνούν μεταξύ τους τα αντίστοιχα μοντέλα τους, προκειμένου να επιτευχθεί συναίνεση για ένα κοινό μοντέλο.

Προκειμένου να διασφαλιστεί το απόρρητο των χρηστών και η εμπιστευτικότητα των δεδομένων, η διαδικασία επικοινωνίας είναι προσεκτικά κατασκευασμένη έτσι ώστε κανένας ιστότοπος να μην μπορεί να «μαντέψει» τα προσωπικά δεδομένα άλλων ιστότοπων. Αυτή είναι η ιδέα πίσω από το Federated Learning (FL). Η ομοσπονδιακή μάθηση (FL) ασκήθηκε για πρώτη φορά σε αρχιτεκτονική τύπου edge-server από τους McMahan *et al* στο πλαίσιο της ενημέρωσης γλωσσικών μοντέλων σε κινητά τηλέφωνα (McMahan et al., 2016a, b, Konecný et al., 2016a, b). Οι διακομιστές Edge εκτελούν υπολογισμό δεδομένων εκτός από αυτούς που εκτελούνται από το διακομιστή εφαρμογών. Συνήθως, οι διακομιστές edge αναπτύσσονται ακριβώς μπροστά από το διακομιστή εφαρμογών και εκτελούν απλές γενικές λειτουργίες, όπως είναι η εξισορρόπηση του φορτίου (Rooney, 2005).

Υπάρχουν πολλές φορητές συσκευές που κατέχουν ιδιωτικά δεδομένα. Για παράδειγμα, για την εκπαίδευση των μοντέλων πρόβλεψης στο σύστημα Gboard, το οποίο είναι το σύστημα πληκτρολογίου της Google για την αυτόματη συμπλήρωση των λέξεων, οι ερευνητές της Google ανέπτυξαν ένα Federated Learning σύστημα για την περιοδική ενημέρωση σε ένα συλλογικό μοντέλο.

Το μοντέλο πρόβλεψης λέξεων στο Gboard βελτιώνεται με βάση όχι μόνο τα συσσωρευμένα δεδομένα ενός κινητού τηλεφώνου, αλλά όλων των τηλεφώνων μέσω μιας τεχνικής γνωστής ως ομόσπονδος μέσος όρος. Ο συνδυασμός μέσων όρων δεν απαιτεί μεταφορά πρωτογενή δεδομένων (raw data) από οποιαδήποτε συσκευή σε μία κεντρική τοποθεσία. Αντ' αυτού, με την ομοσπονδιακή μάθηση, το μοντέλο σε κάθε κινητή συσκευή, η οποία μπορεί να είναι smartphone ή tablet, κρυπτογραφείται και αποστέλλεται στο cloud. Όλα τα κρυπτογραφημένα μοντέλα είναι ενσωματωμένα σε ένα παγκόσμιο μοντέλο, το οποίο είναι κρυπτογραφημένο, έτσι ώστε ο διακομιστής στο cloud δεν γνωρίζει τα δεδομένα σε κάθε συσκευή (Yang et al., 2019, McMahan et al., 2016a, b, Konecný et al., 2016a, b, Hartmann, 2018, Liu et al., 2019).

Το ενημερωμένο μοντέλο, το οποίο βρίσκεται υπό κρυπτογράφηση, στη συνέχεια μεταφορτώνεται σε όλες τις μεμονωμένες συσκευές στο cloud (Konecný et al., 2016b, Hartmann, 2018, Yang et al., 2018). Κατά τη διαδικασία αυτή, τα μεμονωμένα δεδομένα των χρηστών σε κάθε συσκευή δεν αποκαλύπτονται σε άλλους, ούτε στους διακομιστές στο cloud. Το ομοσπονδιακό σύστημα εκμάθησης της Google δείχνει ένα καλό παράδειγμα B2C (business-to-consumer), στο σχεδιασμό ενός ασφαλούς κατανεμημένου μαθησιακού περιβάλλοντος για εφαρμογές B2C. Στη ρύθμιση B2C (ο όρος «επιχείρηση προς καταναλωτή» αναφέρεται στη διαδικασία πώλησης προϊόντων και υπηρεσιών απευθείας μεταξύ μιας επιχείρησης και των καταναλωτών που είναι οι τελικοί χρήστες των προϊόντων ή των υπηρεσιών της. Οι περισσότερες εταιρείες που πωλούν απευθείας σε καταναλωτές μπορούν να αναφέρονται ως εταιρείες B2C, η ομοσπονδιακή μάθηση μπορεί να εξασφαλίσει προστασία της ιδιωτικής ζωής καθώς και αυξημένη απόδοση λόγω της επιτάχυνσης στη μετάδοση των πληροφοριών μεταξύ των συσκευών αιχμής και του κεντρικού διακομιστή. Εκτός από το μοντέλο B2C, η ομοσπονδιακή μάθηση μπορεί επίσης να υποστηρίξει το μοντέλο B2B (business-to-business). Στην

ομοσπονδιακή μάθηση, μια θεμελιώδης αλλαγή στη μεθοδολογία αλγοριθμικού σχεδιασμού είναι, αντί να μεταφέρει κανείς δεδομένα από ιστότοπους σε ιστότοπους, μεταφέρει παραμέτρους μοντέλου με τρόπο τέτοιο ώστε άλλοι χρήστες να μην μπορούν να «δεχτούν» το περιεχόμενο των δεδομένων άλλων χρηστών .

Η ομοσπονδιακή μάθηση στοχεύει στη δημιουργία ενός κοινού μοντέλου ML με βάση τα δεδομένα που βρίσκονται σε πολλές τοποθεσίες. Υπάρχουν δύο διαδικασίες στην ομοσπονδιακή μάθηση: α) η εκπαίδευση μοντέλου (model training) και β) τα συμπεράσματα μοντέλου (model inference). Κατά τη διάρκεια της εκπαίδευσης των μοντέλων, οι πληροφορίες μπορούν να ανταλλάσσονται μεταξύ των χρηστών-δεδομένων αλλά όχι των δεδομένων. Η ανταλλαγή δεν αποκαλύπτει ευαίσθητα ιδιωτικά τμήματα των δεδομένων σε κάθε ιστότοπο (Yang et al., 2019).

Κατά τη διαδικασία των συμπερασμάτων, το μοντέλο εφαρμόζεται σε περίπτωση νέας παρουσίας νέων δεδομένων. Για παράδειγμα, σε B2B ρύθμιση, ένα ομοσπονδιακό σύστημα ιατρικής απεικόνισης μπορεί να λάβει δεδομένα από νέους ασθενείς που έρχονται για διάγνωση σε διαφορετικά νοσοκομεία. Σε αυτήν την περίπτωση, τα μέρη συνεργάζονται για να κάνουν μια πρόβλεψη (Yang et al., 2019).

Σε γενικές γραμμές, η ομοσπονδιακή μάθηση είναι ένα αλγοριθμικό πλαίσιο για την κατασκευή μοντέλων ML που μπορεί να χαρακτηριστεί από τα ακόλουθα χαρακτηριστικά:

• Υπάρχουν δύο ή περισσότερα μέρη που ενδιαφέρονται να δημιουργήσουν από κοινού ένα μοντέλο ML. Κάθε ένα κατέχει ορισμένα δεδομένα που επιθυμεί να συμβάλει στην εκπαίδευση του μοντέλου.

• Στη διαδικασία εκπαίδευσης μοντέλου, τα δεδομένα που κατέχει κάθε μέρος δεν «αφήνουν» αυτό το μέρος.

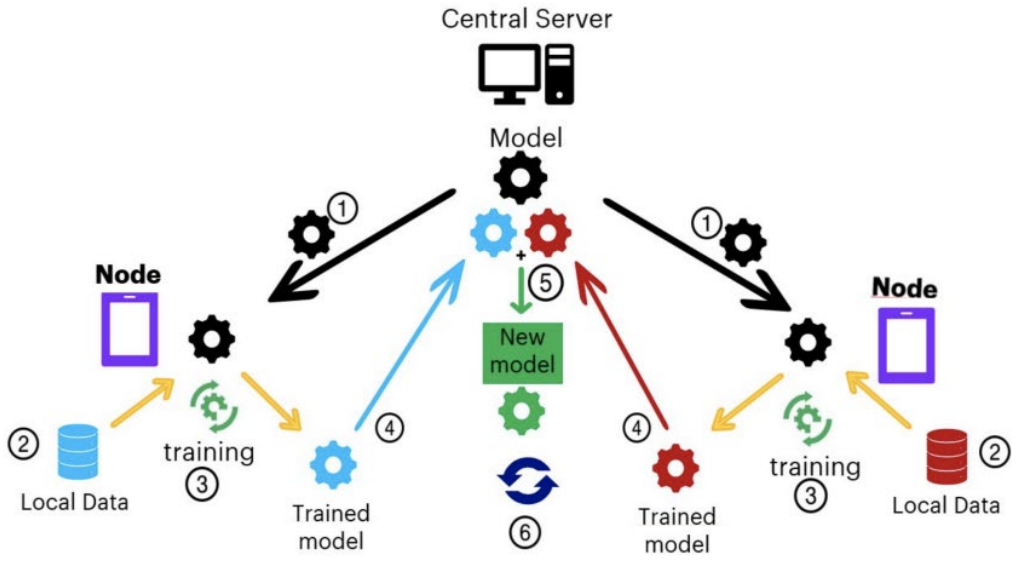
• Το μοντέλο μπορεί να μεταφερθεί εν μέρει από το ένα μέρος στο άλλο με σχέδιο κρυπτογράφησης, έτσι ώστε άλλα μέρη να μην μπορούν να τροποποιήσουν εκ νέου τα δεδομένα σε οποιοδήποτε μέρος.

• Η απόδοση του μοντέλου που προκύπτει είναι μια καλή προσέγγιση του ιδανικού μοντέλου όλων των δεδομένων που μεταφέρονται σε ένα μόνο μέρος (Yang et al., 2019).

Χαρακτηριστικό είναι το γεγονός ότι εάν χρησιμοποιεί κανείς ασφαλή ομοσπονδιακή μάθηση για να δημιουργήσει ένα μοντέλο ML σε κατανεμημένες πηγές δεδομένων, η απόδοση αυτού του μοντέλου σε μελλοντικά δεδομένα είναι περίπου η ίδια με το μοντέλο που βασίζεται στην ένωση όλων των πηγών δεδομένων (Yang et al., 2019).

Με άλλα λόγια, επιτρέπει κανείς στο ομοσπονδιακό σύστημα μάθησης να αποδίδει λίγο λιγότερο από ένα κοινό μοντέλο, γιατί στην ενοποιημένη μάθηση, οι κάτοχοι δεδομένων δεν εκθέτουν τα δεδομένα τους σε κεντρικό διακομιστή ή άλλους ιδιοκτήτες. Αυτή η πρόσθετη εγγύηση ασφάλειας και απορρήτου μπορεί να αξίζει πολύ περισσότερο από την απώλεια ακρίβεια (Yang et al., 2019).

Ένα ομοσπονδιακό σύστημα μάθησης μπορεί ή όχι να περιλαμβάνει έναν κεντρικό υπολογιστή συντονισμού ανάλογα με την εφαρμογή. Ένα παράδειγμα που περιλαμβάνει έναν συντονιστή σε μια ομοσπονδιακή αρχιτεκτονική μάθησης φαίνεται στην εικόνα 2.1. Σε αυτήν τη ρύθμιση, ο συντονιστής είναι ένας κεντρικός διακομιστής συγκέντρωσης (δηλαδή, ο διακομιστής παραμέτρων), ο οποίος στέλνει ένα αρχικό μοντέλο στους τοπικούς κατόχους δεδομένων. Οι τοπικοί κάτοχοι δεδομένων εκπαιδεύουν κάθε μοντέλο χρησιμοποιώντας το αντίστοιχο σύνολο δεδομένων και στείλουν τις ενημερώσεις του μοντέλου στον διακομιστή (Yang et al., 2019).



Εικόνα 2. 1 Τρόπος Λειτουργίας FL (Zaman., 2020)

Η συσσωμάτωση διακόπτεται και στη συνέχεια συνδυάζει τις ενημερώσεις του μοντέλου που λαμβάνονται από τους κατόχους δεδομένων (π.χ., χρησιμοποιώντας τον μέσο όρο του FL) (McMahan et al., 2016a) και στέλνει τις συνδυασμένες ενημερώσεις μοντέλου πίσω στους τοπικούς ιδιοκτήτες των δεδομένων. Αυτή η διαδικασία επαναλαμβάνεται έως ότου συγκλίνει το μοντέλο ή μέχρι τον μέγιστο αριθμό επιτυχών επαναλήψεων. Σύμφωνα με αυτήν την αρχιτεκτονική, τα πρωτογενή δεδομένα των τοπικών κατόχων δεδομένων δεν «απομακρύνονται» ποτέ από τους τοπικούς κατόχους (Yang et al., 2019).

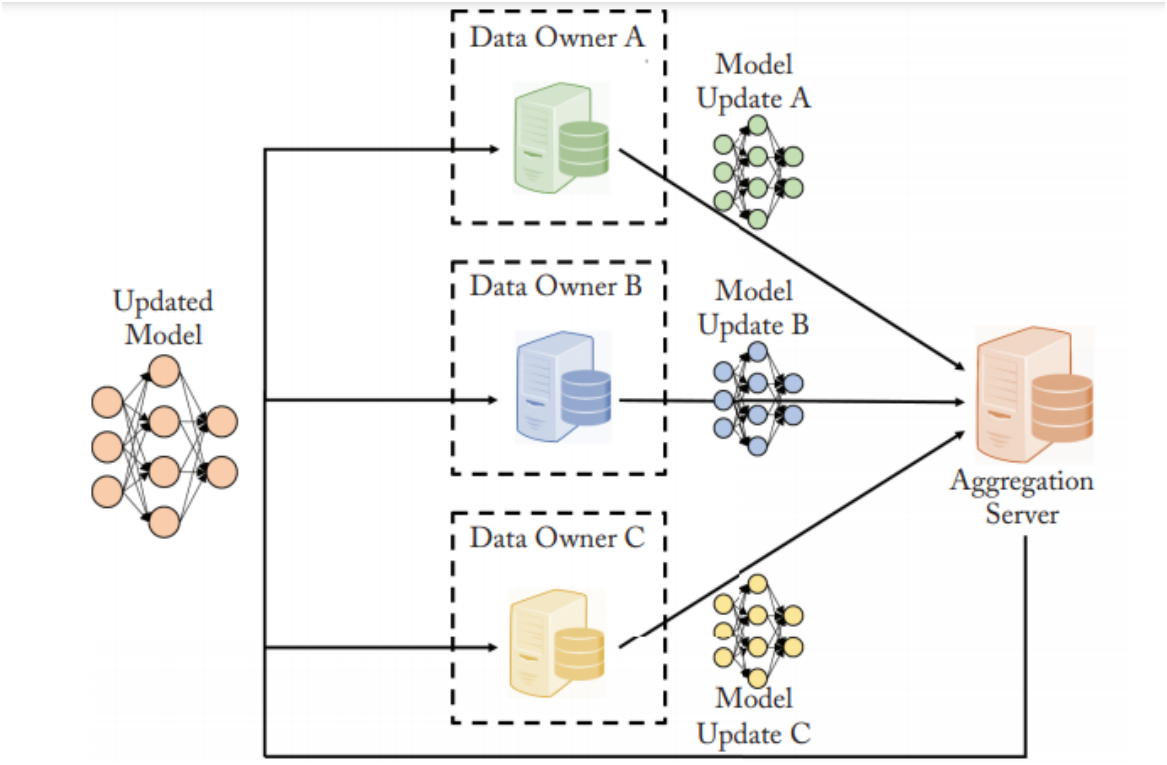
Αυτή η προσέγγιση όχι μόνο διασφαλίζει το απόρρητο των χρηστών και την ασφάλεια των δεδομένων, αλλά και εξοικονομεί τα έξοδα επικοινωνίας για την αποστολή ανεπεξέργαστων δεδομένων. Η επικοινωνία μεταξύ του κεντρικού διακομιστή συγκέντρωσης και των τοπικών κατόχων δεδομένων μπορεί να κρυπτογραφηθεί (π.χ., χρησιμοποιώντας ομομορφική κρυπτογράφηση (Yang et al., 2019, Liu et al., 2019) για την προστασία από τη διαρροή πληροφοριών)).

Η ομοσπονδιακή αρχιτεκτονική μάθησης μπορεί επίσης να σχεδιαστεί με τρόπο peer to peer, ο οποίος δεν απαιτεί την ύπαρξη συντονιστή. Αυτό εξασφαλίζει περαιτέρω εγγύηση ασφάλειας στην οποία τα μέρη επικοινωνούν απευθείας χωρίς τη βοήθεια τρίτου, όπως φαίνεται στην εικόνα 2.3. Το πλεονέκτημα της αρχιτεκτονικής αυτής είναι η αυξημένη

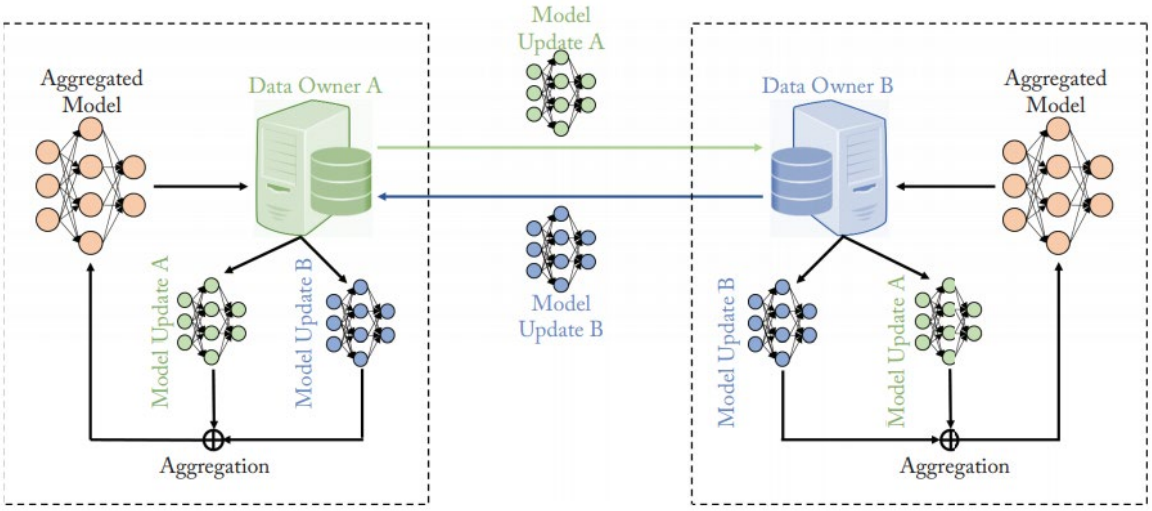
ασφάλεια, αλλά ένα μειονέκτημα είναι ο επιπλέον υπολογισμός που απαιτείται για την κρυπτογράφηση και την αποκρυπτογράφηση των μηνυμάτων.

Το σύστημα Federated Learning έχει πολλά οφέλη. Διατηρεί το απόρρητο των χρηστών και την ασφάλεια των δεδομένων λόγω του σχεδιασμού του. Επίσης δεν απαιτείται μεταφορά δεδομένων. Η ομοσπονδιακή μάθηση επιτρέπει σε πολλά μέρη να εκπαιδεύσουν συνεργατικά ένα μοντέλο ML, έτσι ώστε κάθε μέρος να μπορεί να δημιουργεί ένα καλύτερο μοντέλο από αυτό που μπορεί επιτύχει κάθε μέρος μόνο του. Για παράδειγμα, το Federated Learning μπορεί να χρησιμοποιηθεί από ιδιωτικές εμπορικές τράπεζες για τον εντοπισμό δανεισμού πολλαπλών μερών, ο οποίος ήταν πάντα ένα πρόβλημα στον τραπεζικό κλάδο, ειδικά στη βιομηχανία χρηματοδότησης μέσω διαδικτύου (WeBank AI, 2019).

Με την ομοσπονδιακή μάθηση, δεν χρειάζεται να δημιουργήσει κανείς μια κεντρική βάση δεδομένων, και κάθε χρηματοοικονομικό ίδρυμα που συμμετέχει στην ομοσπονδιακή μάθηση μπορεί να ξεκινήσει νέα ερωτήματα χρηστών σε άλλες εταιρείες εντός της ομοσπονδίας. Οι άλλες εταιρείες χρειάζονται μόνο να απαντήσουν σε ερωτήσεις σχετικά με τον τοπικό δανεισμό χωρίς να γνωρίζουν συγκεκριμένες πληροφορίες του χρήστη. Προστατεύει όχι μόνο το απόρρητο των χρηστών και την ακεραιότητα των δεδομένων, αλλά επιτυγχάνει επίσης έναν σημαντικό επιχειρηματικό στόχο του εντοπισμού του πολυμερούς δανεισμού.



Εικόνα 2. 2 Παράδειγμα αρχιτεκτονικής client-server model (Yang, 2019)



Εικόνα 2. 3 Παράδειγμα αρχιτεκτονικής peer-to-peer model (Yang, 2019)

Ενώ η ομοσπονδιακή μάθηση έχει μεγάλες δυνατότητες, αντιμετωπίζει επίσης πολλές προκλήσεις. Ο σύνδεσμος επικοινωνίας μεταξύ του τοπικού κατόχου δεδομένων και του διακομιστή συγκέντρωσης ενδέχεται να είναι αργός και ασταθής (Hartmann, 2018). Μπορεί να υπάρχει πολύ μεγάλος αριθμός τοπικών κατόχων δεδομένων (π.χ. χρήστες κινητών). Θεωρητικά, κάθε χρήστης κινητής τηλεφωνίας μπορεί να συμμετέχει στην ομοσπονδιακή μάθηση. Όμως, ενδέχεται να ακολουθούν δεδομένα από διαφορετικούς συμμετέχοντες στην ομοσπονδιακή μάθηση με μη πανομοιότυπες διανομές (Zhao et al., 2019, Sattler et al., 2019, van Engelen, 2018). Έτσι, οι συμμετέχοντες μπορεί να έχουν μη ισορροπημένο αριθμό δειγμάτων δεδομένων, το οποίο μπορεί να οδηγήσει σε ένα προκατειλημμένο μοντέλο ή ακόμη και σε αποτυχία εκπαίδευσης ενός μοντέλου.

2.3 Κατηγορίες του FL

Ταξινομούμε την ομόσπονδη μάθηση σε οριζόντια ομοσπονδιακή μάθηση (HFL), κάθετη ομοσπονδιακή μάθηση (VFL) και ομόσπονδη μάθηση μεταφοράς (FTL), σύμφωνα με τον τρόπο με τον οποίο τα δεδομένα κατανέμονται μεταξύ διαφόρων μερών στα χαρακτηριστικά και στους χώρους δειγματοληψίας. Οι εικόνες 2.4-2.6 δείχνουν τις τρεις ομοσπονδιακές κατηγορίες μάθησης για ένα σενάριο δύο μερών (Yang et al., 2019).

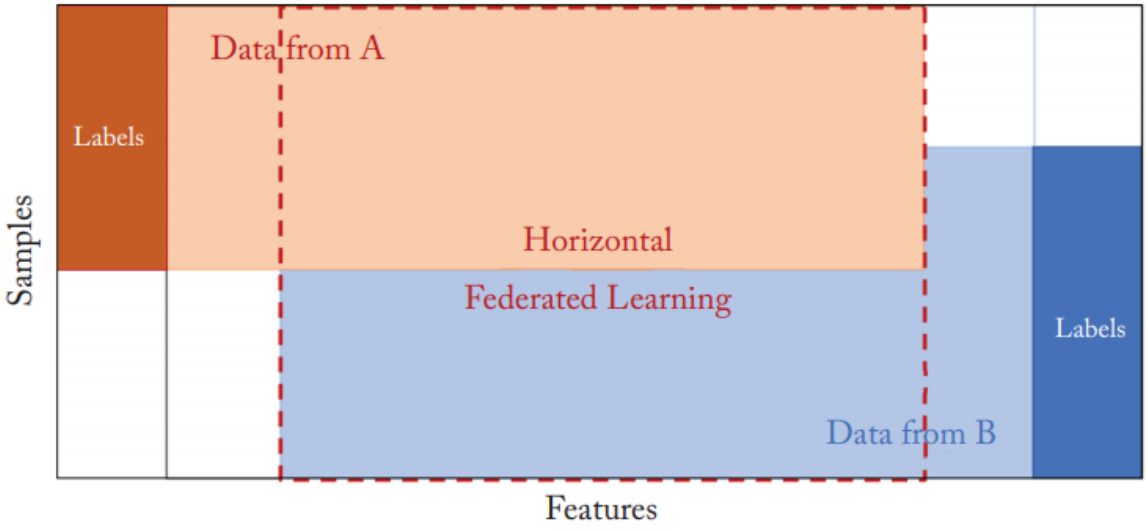
Η κατηγορία HFL αναφέρεται στην περίπτωση όπου οι συμμετέχοντες στην ομόσπονδη μάθηση μοιράζονται αλληλεπικαλυπτόμενα δεδομένα, όπως οι δυνατότητες των δεδομένων ευθυγραμμίζονται μεταξύ των συμμετεχόντων, αλλά διαφέρουν σε δείγματα δεδομένων. Μοιάζει με την κατάσταση κατά την οποία τα δεδομένα χωρίζονται οριζόντια μέσα σε προβολή πίνακα. Ως εκ τούτου, χαρακτηρίζεται ως ομόσπονδη μάθηση κατάτμησης δείγματος, ή ομόσπονδα παραδείγματα εκμάθηση (Kairouz et al., 2019).

Διαφορετική από την κατηγορία HFL, είναι η VFL για την οποία ισχύει το σενάριο όπου οι συμμετέχοντες στην ομοσπονδιακή μάθηση μοιράζονται αλληλεπικαλυπτόμενα δείγματα δεδομένων, δηλαδή τα δείγματα δεδομένων είναι ευθυγραμμισμένα μεταξύ των συμμετεχόντων, αλλά διαφέρουν στα χαρακτηριστικά των δεδομένων. Μοιάζει με την κατάσταση που είναι τα δεδομένα κατακόρυφα κατατμημένα μέσα σε προβολή πίνακα.

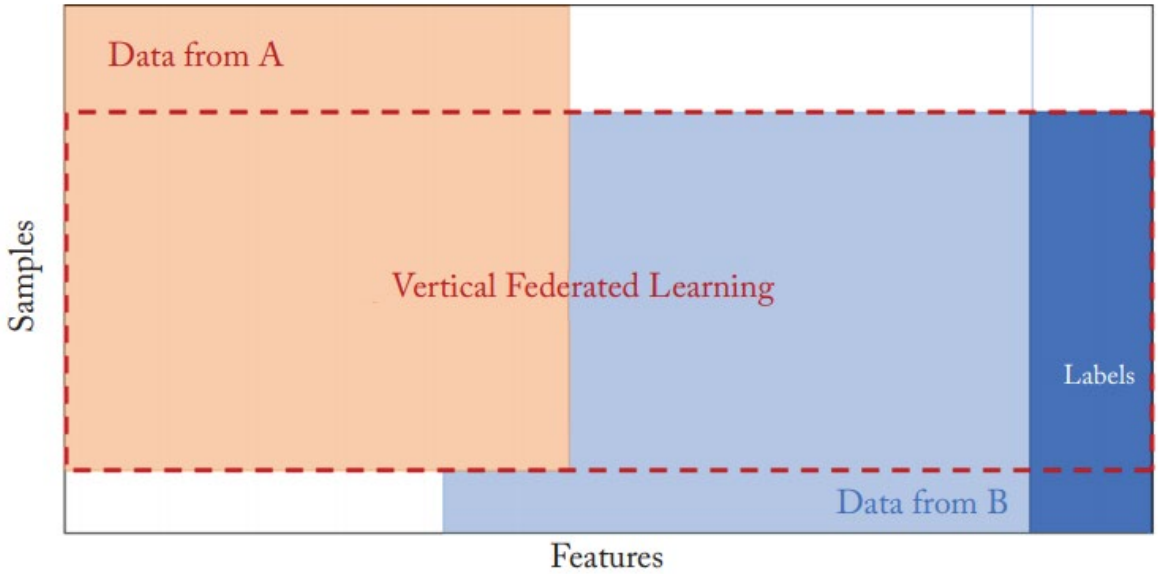
Έτσι, ονομάζουμε επίσης την κατηγορία VFL ως ομόσπονδη μάθηση με διαχωρισμό χαρακτηριστικών.

Τέλος, η κατηγορία FTL ισχύει για την περίπτωση που δεν υπάρχει αλληλεπικάλυψη σε δείγματα δεδομένων ούτε στα χαρακτηριστικά τους. Για παράδειγμα, όταν τα δύο μέρη είναι δύο τράπεζες που εξυπηρετούν δύο διαφορετικές περιφερειακές αγορές, ενδέχεται να μοιράζονται μόνο ελάχιστους χρήστες, αλλά τα δεδομένα τους μπορεί να έχουν παρόμοια κεντρικά χαρακτηριστικά λόγω παρόμοιων επιχειρηματικών μοντέλων. Δηλαδή, με περιορισμένη αλληλοεπικάλυψη στους χρήστες αλλά μεγάλη επικάλυψη στα χαρακτηριστικά των δεδομένων. Οι δύο τράπεζες μπορούν να συνεργαστούν στην οικοδόμηση μοντέλων ML μέσω οριζόντιας ομοσπονδίας μάθησης (Yang et al., 2019, Liu et al., 2019).

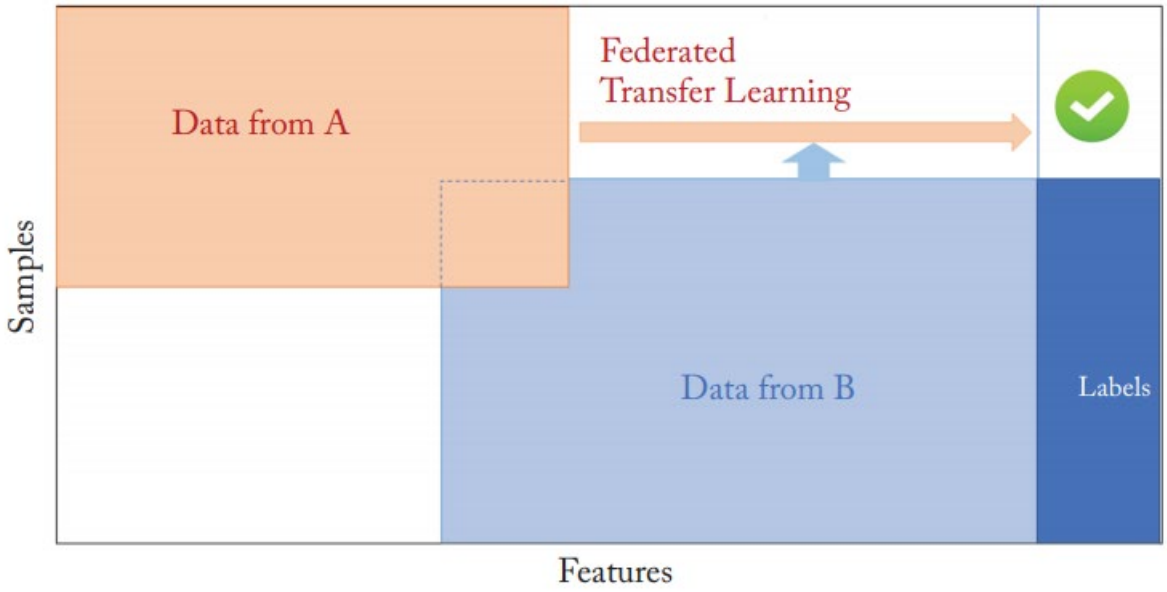
Όταν δύο μέρη παρέχουν διαφορετικές υπηρεσίες αλλά μοιράζονται μεγάλο αριθμό χρηστών (π.χ. μια τράπεζα και μια εταιρεία ηλεκτρονικού εμπορίου), μπορούν να συνεργαστούν στους διαφορετικούς χώρους δυνατοτήτων που κατέχουν, οδηγώντας σε ένα καλύτερο μοντέλο ML και για τις δύο εταιρείες. Δηλαδή, υπάρχει μεγάλη επικάλυψη στους χρήστες αλλά λίγα αλληλεπικαλύπτονται σε χαρακτηριστικά δεδομένων. Έτσι, οι δύο εταιρείες μπορούν να συνεργαστούν στην κατασκευή μοντέλων ML μέσω κάθετης ομοσπονδιακής μάθησης (Yang et al., 2019, Liu et al., 2019).



Εικόνα 2. 4 Κατηγορία HFL



Εικόνα 2. 5 Κατηγορία VFL



Εικόνα 2. 6 Κατηγορία FTL

Στην περίπτωση που τα συμμετέχοντα μέρη έχουν εξαιρετικά ετερογενή δεδομένα (π.χ. αναντιστοιχία διανομής, μετατόπιση τομέα, περιορισμένα επικαλυπτόμενα δείγματα και άλλα), οι κατηγορίες HFL και VFL μπορεί να μην είναι σε θέση να δημιουργήσουν αποτελεσματικά μοντέλα ML. Σε αυτά τα σενάρια, μπορεί κανείς να αξιοποιήσει τις τεχνικές μάθησης μεταφοράς για να γεφυρώσει το χάσμα μεταξύ ετερογενών δεδομένων που ανήκουν σε διαφορετικά μέρη.

Οι ερευνητές Pan και Yang (2010) διαιρούν τη μεταφορά μάθησης σε κυρίως τρεις κατηγορίες:

(i) μεταφορά βάσει παρουσίας,
(ii) μεταφορά βάσει χαρακτηριστικών και
(iii) μεταφορά βάσει μοντέλου.

Εδώ, παρέχονται σύντομες περιγραφές σχετικά με το πώς μπορούν να εφαρμοστούν αυτές οι τρεις κατηγορίες τεχνικών μάθησης μεταφοράς σε συστήματα FL. Αρχικά, στην κατηγορία «FTL-βάσει παρουσίας» τα συμμετέχοντα μέρη επιλέγουν επιλεκτικά ή επαναβαθμίζουν τα εκπαιδευτικά τους δεδομένα έτσι ώστε η απόσταση μεταξύ των

διανομών τομέα να ελαχιστοποιείται, ελαχιστοποιώντας τη λειτουργία αντικειμενικής απώλειας. Έπειτα, στην κατηγορία «FTL-μεταφορά βάσει χαρακτηριστικών» τα συμμετέχοντα μέρη μαθαίνουν συνεργατικά έναν κοινό χώρο αναπαράστασης χαρακτηριστικών, στον οποίο μπορεί να λειτουργήσει πιο εύκολα η κατανομή και η σημασιολογική διαφορά μεταξύ των αναπαραστάσεων χαρακτηριστικών που μεταμορφώνονται από ακατέργαστα δεδομένα. Στην κατηγορία «FTL-βάσει μοντέλου», τα συμμετέχοντα μέρη μαθαίνουν συνεργατικά μοντέλα που μπορούν να ωφεληθούν για τη μεταφορά μάθησης. Εναλλακτικά, τα συμμετέχοντα μέρη μπορούν να χρησιμοποιήσουν προ-εκπαιδευμένα μοντέλα ως σύνολο ή μέρος των αρχικών μοντέλων για μια ομοσπονδιακή εργασία μάθησης.

2.4 Εξελίξεις

Η ιδέα της ομοσπονδιακής μάθησης (Federated Learning) έχει εμφανιστεί σε διάφορες μορφές σε όλη την ιστορία της επιστήμη των υπολογιστών, όπως η προστασία της ιδιωτικής ζωής (Fang and Yang, 2008, Mohassel and Zhang, 2017, Vaidya and Clifton, 2004, Xu et al., 2015). Μελετήθηκε εκτεταμένα από την Google σε ένα ερευνητικό έγγραφο που δημοσιεύθηκε το 2016 στο arXiv. Από τότε, υπήρξε ένας τομέας ενεργής έρευνας στην κοινότητα της τεχνητής νοημοσύνης (AI) όπως αποδεικνύεται από τον ταχέως αναπτυσσόμενο όγκο των προτύπων που εμφανίζονται στο arXiv. Οι ερευνητές Yang et al. (2019) παρέχουν μια ολοκληρωμένη έρευνα για τις πρόσφατες εξελίξεις στο τομέα του Federated Learning.

Οι πρόσφατες ερευνητικές εργασίες για την ομοσπονδιακή μάθηση επικεντρώνονται κυρίως στη βελτίωση της ασφάλειας (Yang et al., 2019). Οι Cheng et al. (2019) προτείνουν το SecureBoost στο περιβάλλον της κάθετης ομοσπονδιακής μάθησης, η οποία είναι ένας νέος τρόπος διατήρησης της ιδιωτικής ζωής χωρίς απώλειες στο σύστημα ενίσχυσης τύπου δέντρων. Το SecureBoost παρέχει το ίδιο επίπεδο ακρίβειας με την προσέγγιση που δεν προστατεύει το απόρρητο. Θεωρητικά αποδεικνύεται ότι το πλαίσιο SecureBoost είναι τόσο ακριβές όσο άλλοι μη ομοσπονδιακοί αλγόριθμοι ενίσχυσης

τύπου δέντρων που βασίζονται σε συγκεντρωτικά σύνολα δεδομένων (Cheng et al., 2019).

Οι Liu et al. (2019) παρουσίασαν ένα ευέλικτο πλαίσιο ομαδικής μάθησης μεταφοράς που μπορεί να προσαρμοστεί αποτελεσματικά σε διάφορες ασφαλείς εργασίες πολλαπλών μερών ML. Σε αυτό το πλαίσιο, η ομοσπονδία επιτρέπει τη μεταφορά της γνώσης που πρέπει να μοιραστεί χωρίς να διακυβεύεται το απόρρητο των χρηστών και επιτρέπει την δωρεάν μεταφορά γνώσεων στο δίκτυο μέσω της τεχνικής μάθησης μεταφοράς.

Κεφάλαιο 3: Κωδικοποίηση των ζητημάτων ασφαλείας

Στο παρόν κεφάλαιο σκοπός είναι η κωδικοποίηση των ζητημάτων ασφαλείας. Αρχικά, γίνεται λόγος για τις ευπάθειες της ασφάλειας των συστημάτων Federated Learning και έπειτα αναφέρονται ζητήματα ιδιωτικότητας.

3.1 Ασφάλεια (Security)

Οι προγραμματιστές και οι ερευνητές της τεχνολογίας Federated Learning (FL) πρέπει να τηρούν τις βασικές αρχές ασφαλείας όπως η εμπιστευτικότητα, η ακεραιότητα και η διαθεσιμότητα. Η αποκεντρωμένη προσέγγιση της ύπαρξης τεράστιου αριθμού πελατών (προς συνεργασία) για την εκπαίδευση συστημάτων. Η έκθεση παραμέτρων μοντέλου καθιστά το σύστημα FL ευάλωτο σε διάφορες επιθέσεις και είναι ανοιχτό σε κινδύνους. Η τρέχουσα έρευνα για την εξερεύνηση τρωτών σημείων και για την πρόταση πλαισίων για την αντιμετώπιση των κινδύνων αυτών είναι πολύ περιορισμένη. Η διερεύνηση που ακολουθεί είναι ταξινομημένη με βάση την ασφάλεια σύμφωνα με το σύστημα FL8 (Mothukuri, 2020).

- Ποιά είναι η πηγή τρωτών σημείων στο οικοσύστημα FL;
- Ποιές είναι οι απειλές ασφαλείας / επιθέσεων στον τομέα FL;
- Ποιές είναι οι μοναδικές απειλές ασφαλείας για το FL σε σύγκριση με διανεμημένες λύσεις ML;
- Ποιές είναι οι αμυντικές τεχνικές για ευπάθειες ασφαλείας FL;

Στις ακόλουθες ενότητες, συζητάμε τα αποτελέσματα με βάση κάθε ερευνητική ερώτηση και παρέχουμε μια ανάλυση των δυνατοτήτων και των αδυναμιών.

3.1.1 Πηγή τρωτών σημείων στο οικοσύστημα FL

Μια ευπάθεια μπορεί να οριστεί ως αδυναμία σε ένα σύστημα που δίνει την ευκαιρία σε κάποιο κακόβουλο εισβολέα να αποκτήσει μη εξουσιοδοτημένη πρόσβαση (Owsap, n.d.). Η γνώση των τρωτών σημείων ενός συστήματος βοηθά στη διαχείριση και στην υπεράσπιση ενάντι στις πιθανές επιθέσεις. Ο εντοπισμός ευπαθειών θα βοηθήσει να χτιστεί ένα πιο ασφαλές περιβάλλον εφαρμόζοντας τις προϋποθέσεις για τα αμυντικά κενά. Η αποτυχία προστασίας της χρήσης και της έκθεσης των προσωπικών αναγνωρίσιμων πληροφοριών ή η μη τήρηση των νόμων περί προστασίας δεδομένων δεν θα προκαλέσει μόνο κακή δημοσιότητα, μπορεί επίσης να έχει πολύ περισσότερες συνέπειες από το νόμο.

Επιπλέον, είναι ένα υποχρεωτικό βήμα για τους προγραμματιστές των συστημάτων FL η σάρωση για όλες τις πηγές ευπάθειας και σύσφιξη της άμυνας για να διασφαλιστεί η ασφάλεια και το απόρρητο των δεδομένων. Για μια καλύτερη εικόνα των τρωτών σημείων, κατηγοριοποιούμε την πηγή τρωτών σημείων στη διαδικασία FL.

Τα αποτελέσματά δείχνουν ότι υπάρχουν πέντε διάφορες πηγές, που αναφέρονται παρακάτω, τα οποία μπορεί να θεωρηθούν αδύνατα σημεία εκμετάλλευσης.

- *Πρωτόκολλο επικοινωνίας:* Το σύστημα FL εφαρμόζει μια επαναληπτική διαδικασία μάθησης με τυχαία επιλεγμένους πελάτες. Το γεγονός αυτό συνεπάγεται σημαντική επικοινωνία μέσω ενός δεδομένου δικτύου. Η προσέγγιση του συστήματος FL προτείνει ένα μικτό δίκτυο (Chaum, 2003), το οποίο βασίζεται σε κρυπτογραφία δημόσιου κλειδιού. Το είδος αυτό της κρυπτογραφίας διατηρεί την πηγή και το περιεχόμενο μηνυμάτων ανώνυμο σε όλη την διάρκεια της επικοινωνίας.

- *Διαχείριση των δεδομένων των πελατών:* Το σύστημα FL σε ένα μεγαλύτερο τοπίο έχει πολλούς πελάτες. Τα δεδομένα είναι ανοιχτά για τους επιτιθέμενους να εκμεταλλευτούν παραμέτρους του μοντέλου και τα δεδομένα εκπαίδευσης. Η πρόσβαση στο παγκόσμιο μοντέλο μπορεί να είναι πιο ευάλωτη σε επιθέσεις ανασυγκρότησης δεδομένων.

- *Κεντρικός Διακομιστής:* Ο κεντρικός διακομιστής πρέπει να είναι εξασφαλισμένος. Αυτό είναι σημαντικό, γιατί είναι υπεύθυνος για την κοινή χρήση των αρχικών παραμέτρων του μοντέλου, τη συγκέντρωση των τοπικών μοντέλων και την κοινή χρήση των καθολικών ενημερώσεων του μοντέλου σε όλους τους πελάτες. Ο διακομιστής που βασίζεται σε cloud ή ο φυσικός διακομιστής πρέπει να ελέγχεται για να διασφαλιστεί ότι δεν υφίστανται εκμετάλλευση ανοιχτών τρωτών σημείων του διακομιστή από κακόβουλους εισβολείς.

- *Weaker Aggregation Algorithm:* Ο αλγόριθμος aggregation είναι κεντρικός στην λειτουργία του συστήματος. Με άλλα λόγια, ως ενημέρωση του τοπικού μοντέλου, θα πρέπει να είναι σε θέση να εντοπίζει τις ενημερώσεις των πελατών και θα πρέπει να απορρίπτει ενημερώσεις από ύποπτους πελάτες. Η αποτυχία απόρριψης των ενημερώσεων αυτών μπορεί να κάνει το παγκόσμιο μοντέλο ευάλωτο.

- *Implementer's of FL Environment:* Σκόπιμα ή ακούσια η ομάδα αρχιτεκτόνων - προγραμματιστών που συμμετέχουν στην υλοποίηση του συστήματος FL μπορεί να αποδειχθεί πηγή κινδύνου. Αυτό μπορεί να συμβεί είτε λόγω της σύγχυσης ή της έλλειψης κατανόησης των ευαίσθητων στοιχείων του χρήστη, που μπορεί να είναι και ο λόγος για την παραβίαση της ασφάλειας και της ιδιωτικής ζωής του ατόμου.

Ο κίνδυνος από τους ανθρώπους που συμμετέχουν στην υλοποίηση του συστήματος FL μπορεί να οφείλεται στο βασικό γεγονός ότι οι ίδιοι δεν έχουν λάβει τα κατάλληλα μέτρα για τη σάρωση ευαίσθητων δεδομένων (Mothukuria, 2020).

3.1.2 Απειλές ασφαλείας / επιθέσεις στον τομέα FL

Η απειλή ή η επίθεση είναι η πιθανότητα εκμετάλλευσης μιας ευπάθειας του συστήματος από κάποιο κακόβουλο ή περίεργο εισβολέα που επηρεάζει την ασφάλεια του συστήματος και παραβιάζει τις πολιτικές απορρήτου. Στο σύστημα FL, γενικά, ο κακόβουλος εισβολέας του συστήματος χρησιμοποιεί τις ευπάθειες (Men, 2019) με στόχο την κατάκτηση του ελέγχου ενός ή περισσοτέρων συμμετεχόντων (δηλαδή πελατών) και τελικά την χειραγώγηση του παγκόσμιου μοντέλου.

Σε ένα τέτοιο σενάριο, ο εισβολέας στοχεύει διαφορετικούς πελάτες με στόχο την πρόσβαση σε τοπικά δεδομένα σε κατάσταση ηρεμίας, ή κατά τη διαδικασία της εκπαίδευσης (Bagdasaryan, 2020). Οι απειλές ή οι επιθέσεις ασφαλείας ταξινομούνται και αντιστοιχούν στις περιγραφές που συζητούνται στις ακόλουθες υποενότητες.

3.1.3 Δηλητηρίαση (Poisoning)

Μια επίθεση με μεγάλη πιθανότητα εμφάνισης στα συστήματα FL είναι γνωστή ως δηλητηρίαση ή αλλιώς poisoning (Feng, 2019; Halawa, 2017), καθώς κάθε πελάτης στο σύστημα αυτό έχει πρόσβαση στα δεδομένα εκπαίδευσης και τη δυνατότητα προσθήκης βαρών δεδομένων που προσβλήθηκαν στο παγκόσμιο μοντέλο ML. Η «δηλητηρίαση» μπορεί να συμβεί κατά τη διάρκεια της εκπαίδευσης και μπορεί να επηρεάσει είτε την εκπαίδευση του συνόλου των δεδομένων ή μόνο το τοπικό μοντέλο που με τη σειρά του μπορεί να παραβιάσει έμμεσα το παγκόσμιο μοντέλο ML (σε παράγοντες όπως είναι η απόδοση ή η ακρίβεια).

Στα συστήματα FL, οι ενημερώσεις μοντέλων λαμβάνονται από μια μεγάλη ομάδα πελατών. Δηλαδή, η πιθανότητα δηλητηρίασης από έναν ή περισσότερους πελάτες όσο αφορά τα δεδομένα εκπαίδευσης είναι υψηλή, όπως και η σοβαρότητα της απειλής. Οι στόχοι της επίθεσης αυτού του είδους αφορούν διάφορα αντικείμενα στη διαδικασία του FL. Στη συνέχεια παρουσιάζονται οι κατηγορίες των επιθέσεων δηλητηρίασης:

- *Data Poisoning:* Η έννοια μιας επίθεσης δηλητηρίασης δεδομένων κατά αλγορίθμων ML παρουσιάστηκε για πρώτη φορά από συγγραφείς / ερευνητές

στη δημοσίευση Biggo (2012). Ο επιτιθέμενος στοχεύει στην ευπάθεια του αλγορίθμου διανυσματικών μηχανών υποστήριξης και προσπαθεί να ενσωματώσει σημεία κακόβουλων δεδομένων στη φάση της εκπαίδευσης με την ελπίδα να μεγιστοποιηθεί το σφάλμα ταξινόμησης. Από τότε, μια μεγάλη ποικιλία προσεγγίσεων έχει προταθεί για τον μετριασμό των επιθέσεων δηλητηρίασης δεδομένων στους αλγόριθμους ML με διαφορετικές ρυθμίσεις.

Ενώ το περιβάλλον του συστήματος FL επιτρέπει στους πελάτες να συμβάλλουν ενεργά στην εκπαίδευση δεδομένων και την αποστολή παραμέτρων μοντέλου στο διακομιστή, παρέχει αυτήν την ευκαιρία και σε κακόβουλους πελάτες. Οι ίδιοι μπορούν να δηλητηριάσουν το παγκόσμιο μοντέλο χειραγωγώντας τη διαδικασία εκπαίδευσης. Η δηλητηρίαση των δεδομένων σε συστήματα FL ορίζεται ως παραγωγή ακατάλληλων δειγμάτων για την εκπαίδευση του παγκόσμιου μοντέλου με ελπίδες παραγωγής παραποιημένων παραμέτρων μοντέλου και αποστολή τους στο διακομιστή.

Η έγχυση δεδομένων μπορεί επίσης να θεωρηθεί ως υποκατηγορία δηλητηρίασης δεδομένων όπου ο κακόβουλος πελάτης μπορεί να εισάγει κακόβουλα δεδομένα σε τοπικό επίπεδο - πελάτη για την επεξεργασία του μοντέλου. Ως αποτέλεσμα, ο κακόβουλος χρήστης μπορεί να αναλάβει τον έλεγχο τοπικών μοντέλων πολλαπλών πελατών και τελικά να καταφέρει να χειρίζεται το παγκόσμιο μοντέλο με κακόβουλα δεδομένα.

- *Model Poisoning:* Κατά τη δηλητηρίαση δεδομένων, ο κακόβουλος χρήστης στοχεύει την χειραγώγηση του παγκόσμιου μοντέλου χρησιμοποιώντας ψεύτικα δεδομένα στο μοντέλο δηλητηρίασης, από την άλλη ο κακόβουλος χρήστης στοχεύει άμεσα το παγκόσμιο μοντέλο. Πρότυπες επιθέσεις δηλητηρίασης έχουν αποδειχθεί πιο αποτελεσματικές σε σύγκριση με τις επιθέσεις δηλητηρίασης δεδομένων σε πρόσφατες έρευνες (Bagdasaryan, 2020).

- *Τροποποίηση δεδομένων:* Οι επιθέσεις παραβίασης ή τροποποίησης δεδομένων ενδέχεται να περιλαμβάνουν την αλλαγή του συνόλου των δεδομένων κατάρτισης, όπως είναι η σύγκρουση χαρακτηριστικών, η οποία συγχωνεύει δύο τάξεις στο σύνολο δεδομένων σε μια προσπάθεια να αλλάξει το μοντέλο ML για πάντα με εσφαλμένη ταξινόμηση της στοχευμένης τάξης. Ορισμένες τεχνικές περιλαμβάνουν απλώς την προσθήκη «σκιάς» ή μοτίβου μιας άλλης κλάσης σε μια στοχευμένη τάξη που μπορεί να προκαλέσει σύγχυση στο μοντέλο ML. Μια άλλη τεχνική περιλαμβάνει την τυχαία ανταλλαγή ετικετών της εκπαίδευσης σύνολο δεδομένων. Μπορούν να ληφθούν υπόψη και οι επιθέσεις με έγχυση και τροποποίηση δεδομένων ως τύπος επιθέσεων δηλητηρίασης δεδομένων ML στο σύστημα FL (Nasr, 2019).

3.1.4 Inference

Οι επιθέσεις Inference απειλούν περισσότερο την ιδιωτική ζωή όμως συμπεριλαμβάνονται και εδώ για τη συνολική σύγκριση των απειλών στο σύστημα FL. Η σοβαρότητα των επιθέσεων αυτών είναι παρόμοια με τις επιθέσεις δηλητηρίασης. Επίσης, υπάρχει μια υψηλή πιθανότητα επίθεσης είτε από τους συμμετέχοντες είτε από κάποιο κακόβουλο κεντρικό διακομιστή στη διαδικασία FL.

3.1.5 Επιθέσεις Backdoor

Οι επιθέσεις δηλητηρίασης και Inference είναι πιο διαφανείς σε σύγκριση με τις backdoor επιθέσεις. Μια επίθεση backdoor είναι ένας τρόπος να εισάγει κανείς μια κακόβουλη εργασία στο υπάρχον μοντέλο διατηρώντας παράλληλα την ακρίβεια της πραγματικής εργασίας. Είναι δύσκολο και χρονοβόρο να προσδιορίσει κανείς τις επιθέσεις backdoor αφού η ακρίβεια της πραγματικής εργασίας στο ML μπορεί να μην επηρεάσει άμεσα. Οι ερευνητές στο Bagdasaryan (2020) πειραματίζονται στο πώς εφαρμόζονται οι επιθέσεις backdoor. Επιπλέον, οι ερευνητές στο Liu (2018) προτείνουν το «κλάδεμα» των μοντέλων και τη βελτιστοποίηση της λύσης για τον μετριασμό των κινδύνων backdoor επιθέσεων.

Η σοβαρότητα των επιθέσεων backdoor είναι υψηλή καθώς χρειάζεται σημαντικός χρόνος για τον εντοπισμό της εμφάνισης της επίθεσης. Επιπλέον, ο αντίκτυπος της επίθεσης είναι υψηλός καθώς οι επιθέσεις backdoor έχουν τη δυνατότητα να συγχέουν τα μοντέλα ML και να προβλέπουν τα ψεύτικα θετικά με σιγουριά. Οι απειλές Trojan (Bagdasaryan, 2020) είναι παρόμοιας κατηγορίας επιθέσεων backdoor που προσπαθούν να διατηρήσουν την υπάρχουσα εργασία του ML μοντέλου κατά την εκτέλεση κακόβουλης εργασίας σε κατάσταση μυστικότητας.

3.1.6 GAN

Οι Generative Adversarial Network-based (GAN) επιθέσεις με βάση το δίκτυο Adversarial στο FL έχουν πειραματιστεί και αναλύθηκαν από πολλούς ερευνητές. Με την ικανότητά τους να ξεκινούν την δηλητηρίαση και Inference επιθέσεις, οι επιθέσεις με βάση το GAN αποτελούν απειλή τόσο για την ασφάλεια αλλά και για το απόρρητο ενός δεδομένου συστήματος. Ερευνητικό έργο δείχνει πώς οι επιθέσεις GAN μπορούν να χρησιμοποιηθούν για τη λήψη δεδομένων εκπαίδευσης μέσω Inference και τη χρήση GAN για να δηλητηριάσουν τα εκπαιδευτικά δεδομένα (Hitaj, 2017). Όπως όλες οι δυνατότητες μιας απειλής που βασίζεται στο GAN δεν μπορεί να προβλεφθεί η κατηγοριοποίηση του αντίκτυπου και η προτεραιότητα (Hitaj, 2017).

3.1.7 Διακοπή του συστήματος IT downtime

Ο χρόνος διακοπής του συστήματος παραγωγής αποτελεί αναπόφευκτη απειλή στις του Information Technology (IT). Συχνά παρατηρείται ότι οι πολύ διαμορφωμένες και ασφαλείς εφαρμογές πρέπει να έχουν φάση εκτός λειτουργίας λόγω μη προγραμματισμένων δραστηριοτήτων σε διακομιστές back-end. Στο σύστημα FL, η σοβαρότητα αυτής της απειλής είναι χαμηλή, αφού στην πραγματικότητα είναι ένα τοπικό-παγκόσμιο μοντέλο σε κάθε κόμβο πελάτη και η διαδικασία εκπαίδευσης μπορεί συνεχιστεί μετά τη διακοπή. Ακόμη, με χαμηλή σοβαρότητα, αυτή είναι μια σημαντική απειλή καθώς το downtime μπορεί να είναι μια καλά σχεδιασμένη επίθεση για κλοπή πληροφοριών από το περιβάλλον FL.

3.1.8 Κακόβουλος διακομιστής

Στο σύστημα FL μεταξύ των συσκευών, το μεγαλύτερο μέρος της εργασίας γίνεται στον κεντρικό διακομιστή, επιλέγοντας τις παραμέτρους του μοντέλου για την ανάπτυξη του καθολικού μοντέλου. Οι κακόβουλοι διακομιστές έχουν τεράστιο αντίκτυπο και μπορούν εύκολα να εξαγάγουν δεδομένα ιδιωτικών πελατών ή να χειριστούν το παγκόσμιο μοντέλο αξιοποιώντας την κοινή υπολογιστική ισχύ για τη δημιουργία κακόβουλων εργασιών στο παγκόσμιο μοντέλο ML.

3.1.9 Σημεία συμφόρησης επικοινωνίας (Communication bottlenecks)

Μία από τις προκλήσεις στην εκπαίδευση ενός μοντέλου ML από πολλαπλά ετερογενή δεδομένα της συσκευής είναι το εύρος ζώνης επικοινωνίας. Στην προσέγγιση FL, το κόστος επικοινωνίας μειώνεται μεταφέροντας εκπαιδευμένα μοντέλα αντί να στέλνει κανείς τεράστιο όγκο δεδομένων. Ωστόσο, έχουμε την ανάγκη να διατηρήσουμε το εύρος ζώνης επικοινωνίας. Υπάρχουν μερικοί αλγόριθμοι που βασίζονται στην ασύγχρονη συγκέντρωση μοντέλων και λίγες στρατηγικές για καλή απόδοση ακόμη και με χαμηλή επικοινωνία στο εύρος ζώνης. Υπάρχουν διάφορες ερευνητικές μελέτες σχετικά με τη διατήρηση του εύρους ζώνης επικοινωνίας στο περιβάλλον FL. Η σοβαρότητα αυτής της απειλής είναι υψηλή, καθώς τα σημεία συμφόρησης επικοινωνίας μπορούν να διαταράξουν το FL περιβάλλον σημαντικά (Mukherjee, 2020).

3.1.10 Επιθέσεις free-riding

Λίγοι πελάτες παίζουν παθητικό ρόλο και συνδέονται μόνο με το περιβάλλον για να αξιοποιήσουν τα οφέλη του παγκόσμιου μοντέλου ML χωρίς να συνεισφέρουν στην εκπαιδευτική διαδικασία. Τέτοιοι παθητικοί πελάτες μπορούν επίσης να εισάγουν εικονικές ενημερώσεις χωρίς εκπαίδευση του μοντέλου ML με τα τοπικά δεδομένα τους. Ερευνητικά έργα διερεύνησαν αυτού του είδους τις επιθέσεις σε περιβάλλον FL και πρότειναν μια βελτιωμένη έκδοση του στην τεχνική ανίχνευσης ανωμαλιών χρησιμοποιώντας αυτόματους κωδικοποιητές. Ο αντίκτυπος αυτής της επίθεσης θα ήταν περισσότερο σε ένα μικρότερο περιβάλλον FL καθώς η απουσία συμμετοχής πελατών

μπορεί να επηρεάσει αρνητικά την παγκόσμια εκπαίδευση μοντέλων. Η πιθανότητα αυτής της επίθεσης είναι χαμηλή και η σοβαρότητα της μέτρια (Mukherjee, 2020).

3.1.11 Μη διαθεσιμότητα (Unavailability)

Η μη διαθεσιμότητα ή η εγκατάλειψη πελατών μεταξύ των εκπαιδευτικών διαδικασιών μπορεί να αποφέρει αναποτελεσματικά αποτελέσματα στην εκπαίδευση του παγκόσμιου μοντέλου. Αυτή η απειλή είναι παρόμοια με την επίθεση free-riding, αλλά σε αυτό το σενάριο οι πελάτες χάνουν ακούσια τη συμμετοχή τους στη διαδικασία κατάρτισης λόγω προβλημάτων δικτύου. Η σοβαρότητα αυτής της απειλής είναι μέτρια καθώς η πιθανότητα της είναι χαμηλή (Mothukuri, 2020).

3.1.12 Υποκλοπές

Στο σύστημα FL, έχουμε μια επανάληψη της διαδικασίας μάθησης που περιλαμβάνει κύκλους επικοινωνίας από τους πελάτες στον κεντρικό διακομιστή. Οι επιτιθέμενοι μπορεί να παρακολουθούν και να εξάγουν δεδομένα μέσω ενός αδύναμου καναλιού επικοινωνίας εάν υπάρχει. Η υποκλοπή μπορεί να θεωρηθεί απειλή μέσης σοβαρότητας κατά της επίθεσης σε μοντέλα FL, δεδομένου ότι τα μοντέλα black-box γενικά είναι δύσκολο να επιτεθούν. Οι επιτιθέμενοι θα προτιμούσαν εξαγορά πελάτη με ασθενέστερη ασφάλεια που θα προσφέρει εύκολα παραμέτρους μοντέλου και συνεπώς του παγκόσμιου μοντέλου (Mothukuri, 2020).

3.1.13 Αλληλεπίδραση με τους νόμους περί προστασίας δεδομένων

Αυτή η απειλή έχει χαμηλή πιθανότητα εμφάνισης καθώς ένας επιστήμονας δεδομένων που διαμορφώνει το περιβάλλον FL διασφαλίζει ότι η ανάπτυξη του παγκόσμιου μοντέλου έχει αναλυθεί σωστά πριν τεθεί σε παραγωγή σε όλους τους πελάτες. Η σοβαρότητα της απειλής είναι χαμηλή, αλλά εξακολουθεί να αποτελεί σημαντική απειλή ως σκόπιμη ή ακούσια εσφαλμένη διαμόρφωση στο σύστημα FL και μπορεί να οδηγήσει σε παραβίαση της ασφάλειας (Mothukuri, 2020).

3.2 Μοναδικές απειλές ασφαλείας για τη FL σε σύγκριση με τις κατανεμημένες λύσεις ML

Οι λύσεις κατανεμημένης μηχανικής εκμάθησης (DML) που προτείνονται μέχρι τώρα στοχεύουν στην επίλυση προκλήσεις του Big Data και της υπολογιστικής ισχύς κατά την εκπαίδευση του μοντέλου ML. Από την άποψη της αρχιτεκτονικής, η DML μοιράζεται μερικές κοινές ιδιότητες με τη FL και υπάρχουν ερευνητικές μελέτες για την αντιμετώπιση προβλημάτων ασφάλειας και απορρήτου στο DML. Ωστόσο, το σύστημα FL είναι διαφορετικό σε σχέση με τις υπάρχουσες λύσεις DML και από προεπιλογή υπάρχει υψηλότερο επίπεδο ασφάλειας και διασφάλισης απορρήτου.

Αυτή η ενότητα έχει ως στόχο να συζητήσει τις μοναδικές απειλές στα συστήματα FL και τις κοινές απειλές στην κοινή χρήση μεταξύ FL και DML. Αυτό βοηθά στην κατανόηση της υπάρχουσας εργασίας στο DML και στην διερεύνηση προσαρμόσιμων ερευνητικών ιδεών από DML έως FL. Οι συγκεκριμένοι κίνδυνοι για την ασφάλεια και το απόρρητο για το DML είναι εκτός πεδίου σε αυτό το έγγραφο και δεν συζητούνται, εστιάζουμε μόνο σε κοινούς παράγοντες κινδύνου των FL και DML.

Η εικόνα 3.2 παρουσιάζει μια εμπεριστατωμένη περίληψη της ταξινόμησης των απειλών ασφαλείας και επιθέσεων. Η απειλή δηλητηρίασης και οι Backdoor επιθέσεις είναι τόσο για DML αλλά και για συστήματα FL. Τα δεδομένα στους κόμβους πελατών μπορούν να τροποποιηθούν σε κόμβους πελατών που είναι κοινά στο σύστημα FL και στην αρχιτεκτονική DML. Ο διακομιστής παραμέτρων του συστήματος DML και ο κεντρικός διακομιστής είναι επιρρεπής σε επιθέσεις που οδηγούν σε παραβίαση της ασφάλειας.

Το σημείο συμφόρησης επικοινωνίας είναι, επίσης, μια απειλή στα συστήματα DML και FL και απαιτεί μεγάλη προσοχή καθώς και τα δύο πλαίσια πρέπει να επικοινωνούν με τους αντίστοιχους κόμβους των πελατών τους. Επιπλέον, η εικόνα 3.3 απεικονίζει τη σοβαρότητα των απειλών. Η σοβαρότητα υπολογίζεται με βάση την πιθανότητα του αντιπάλου να εκμεταλλευτεί την ευπάθεια και να ξεκινήσει μια απειλή.

Όπως φαίνεται στην εικόνα 3.3, οι επιθέσεις που βασίζονται στην δηλητηρίαση έχουν τη μεγαλύτερη σοβαρότητα. Αυτό θα μπορούσε να αποδοθεί στο γεγονός ότι το παγκόσμιο μοντέλο μπορεί να δηλητηριαστεί από πολλές πηγές όπως τοπικές ενημερώσεις μοντέλων, κακόβουλους διακομιστές και πολλά άλλα. Όσο μεγαλύτερη είναι η πιθανότητα απειλής, τόσο υψηλότερος είναι ο αντίκτυπος της επίθεσης στο σύστημα FL.

Η απειλή των backdoor επιθέσεων είναι η πιο επικίνδυνη καθώς επίσης είναι πιο δύσκολο να προσδιοριστεί μια τέτοια επίθεση και ο αντίκτυπος της έχει την δυνατότητα καταστροφής της αυθεντικότητας του παγκόσμιου μοντέλου. Οι επιθέσεις που βασίζονται στο GAN έχουν επίσης μεγάλη βαρύτητα λόγω της απρόβλεπτης ικανότητάς τους να επηρεάζουν την ασφάλεια και το απόρρητο των δεδομένων του χρήστη. Η κακόβουλη απειλή διακομιστή είναι η πιο επικίνδυνη, λόγω της ανοιχτής ευπάθειας των φυσικών διακομιστών που βασίζονται σε cloud.

Η διακοπή του συστήματος καθώς και ο χρόνος διακοπής IT επισημαίνονται με χαμηλό επίπεδο σοβαρότητας καθώς ο αντίκτυπος θα ήταν μικρότερος με κάθε πελάτη που κατέχει το παγκόσμιο μοντέλο ξεχωριστά. Η συμφόρηση επικοινωνίας στο FL είναι ένα καλά ερευνημένο θέμα καθώς μπορεί να είναι showstopper στην εκπαίδευση. Ο αντίκτυπος και η προτεραιότητα αυτής της απειλής θεωρείται υψηλός στο περιβάλλον FL με τεράστιο αριθμό πελατών δημοσιεύοντας ενημερώσεις με κάθε επανάληψη της μαθησιακής διαδικασίας. Ο αντίκτυπος της απειλής free-riding ορίζεται ως μέσος, καθώς αυτό είναι δυνατό σε λιγότερα σενάρια και ο αντίκτυπός του δεν είναι σοβαρός καθώς η εκπαίδευση συνεχίζεται με άλλους πελάτες. Η μη διαθεσιμότητα μπορεί να σταματήσει τη διαδικασία εκμάθησης όπου ο αλγόριθμος συγκέντρωσης δεν είναι ισχυρός για τον χειρισμό των εγκαταλείψεων. Ο αντίκτυπος της απειλής αυτής επισημαίνεται ως μέσου επιπέδου καθώς υπάρχουν αλγόριθμοι, οι οποίοι έχουν σχεδιαστεί για τη διαχείριση των εγκαταλείψεων αυτών (Mukherjee, 2020).

3.3 Απειλές και επιθέσεις απορρήτου στον τομέα FL

Το σύστημα και το μοντέλο FL στοχεύει στη διασφάλιση της ιδιωτικότητας των συμμετεχόντων ζητώντας από τους συμμετέχοντες να μοιραστούν τοπικές παραμέτρους μοντέλου κατάρτισης αντί για τα πραγματικά δεδομένα τους. Ωστόσο, σύμφωνα με την πρόσφατη έρευνα (Mothukuri, 2020), το μοντέλο FL εξακολουθεί να έχει κάποιες απειλές για το απόρρητο, επειδή οι αντίπαλοι μπορούν να αποκαλύψουν εν μέρει τα εκπαιδευτικά δεδομένα κάθε συμμετέχοντα στο αρχικό σύνολο δεδομένων εκπαίδευσης βάσει της μεταφορτωμένης παραμέτρου τους. Τέτοιες κρίσιμες απειλές στο σύστημα FL μπορούν να γενικευτούν σε διαφορετικές κατηγορίες επιθέσεων βάσει των συμπερασμάτων.

3.3.1 Επιθέσεις μελών τύπου Inference

Όπως υποδηλώνει το όνομα, μια Inference επίθεση είναι ένας τρόπος για να συναγάγει κανείς δεδομένα εκπαίδευσης αλλά και λεπτομέρειες. Η επίθεση συνδρομής μέλους στοχεύει στη λήψη πληροφοριών ελέγχοντας εάν τα δεδομένα υπάρχουν σε ένα εκπαιδευτικό σύνολο. Ο εισβολέας κάνει κακή χρήση του παγκόσμιου μοντέλου για να λάβει πληροφορίες σχετικά με τα εκπαιδευτικά δεδομένα των άλλων χρηστών. Σε τέτοιες περιπτώσεις, οι πληροφορίες σχετικά με το σύνολο δεδομένων εκπαίδευσης συνάγονται μέσω εικασίας και εκπαίδευσης του προγνωστικού μοντέλου για την πρόβλεψη των αρχικών δεδομένων εκπαίδευσης. Οι ερευνητές στο (Nasr, 2019) διερεύνησαν την ευπάθεια του νευρικού δικτύου για να απομνημονεύσει τα εκπαιδευτικά δεδομένα που είναι επιρρεπές σε παθητικές και ενεργές επιθέσεις τύπου Inference.

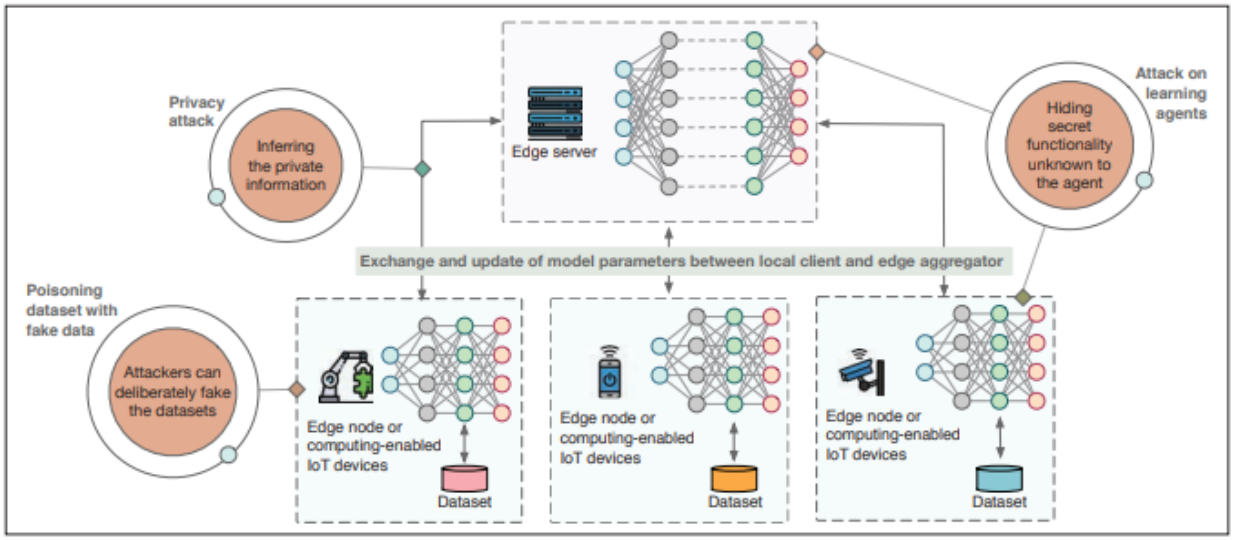
3.3.2 Αθέλητη διαρροή δεδομένων και ανακατασκευή μέσω συμπερασμάτων

Είναι ένα σενάριο όπου διαρρέουν με ακούσιο τρόπο ενημερώσεις ή πληροφορίες από πελάτες στον κεντρικό διακομιστή. Οι ερευνητές στο (Yang, 2019) εκμεταλλεύονται τα δεδομένα ευπάθειας διαρροής και ανακατασκευάζουν επιτυχώς τα δεδομένα των άλλων πελατών μέσω μιας επίθεσης Inference.

Το ερευνητικό έργο στη δημοσίευση Hitaj (2017) διερευνά πώς μπορούν τα ιδιωτικά δεδομένα από έναν έντιμο πελάτη να αποκαλυφθούν δεδομένα χρησιμοποιώντας επιθέσεις Inference που βασίζονται σε GAN. Ο πελάτης δημιουργεί δεδομένα παρόμοια με τα εκπαιδευτικά δεδομένα χρησιμοποιώντας GAN και ανακτά ευαίσθητες πληροφορίες από άλλους πελάτες στο σύστημα FL. Οι κακόβουλοι πελάτες έχει αναφερθεί ότι χρησιμοποιούν το παγκόσμιο μοντέλο και παραμέτρους για την ανακατασκευή των εκπαιδευτικών δεδομένων άλλων πελατών (Mukherjee, 2020).

3.3.3 Επιθέσεις Inference που βασίζονται σε GAN

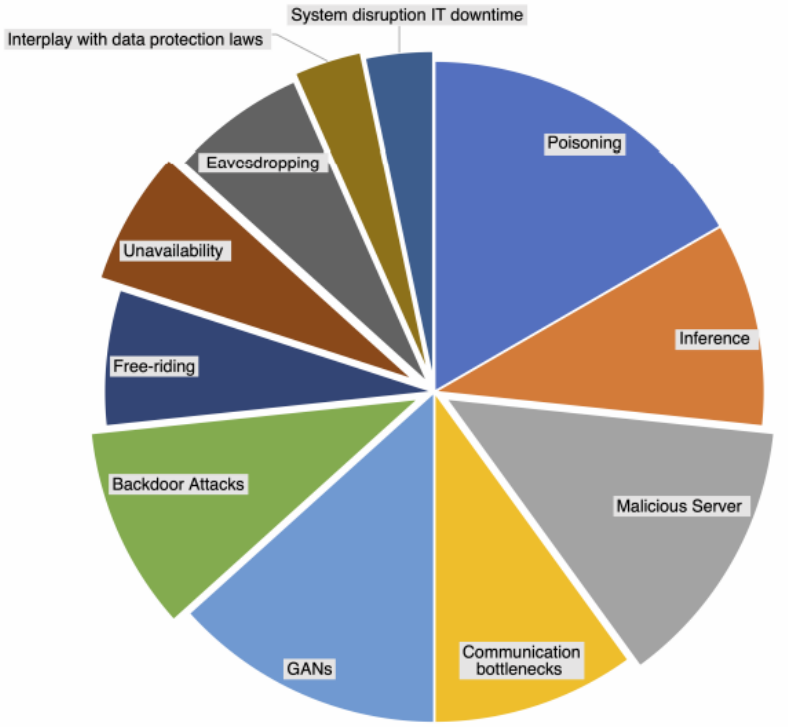
Τα GANs είναι εχθρικά δίκτυα που έχουν αποκτήσει μεγάλη δημοτικότητα σε μεγάλους τομείς δεδομένων τα τελευταία χρόνια και είναι επίσης εφαρμόσιμα σε FL προσεγγίσεις. Ειδικά για το σύστημα FL, οι συγγραφείς ή οι ερευνητές προτείνουν το mGAN-AI πλαίσιο για την εξερεύνηση επιθέσεων με βάση το GAN στο σύστημα FL. Οι επιθέσεις mGAN-AI πειραματίζονται σε έναν κακόβουλο κεντρικό διακομιστή του περιβάλλοντος FL. Η επίθεση Inference αποκτά την υψηλότερη ακρίβεια με το πλαίσιο mGAN-AI επειδή δεν παρεμβαίνει στην εκπαιδευτική διαδικασία (Mukherjee, 2020).



Εικόνα 3. 1 Ζητήματα ασφαλείας ενδέχεται να προκύψουν σε: α) δηλητηρίαση από σύνολο δεδομένων σε κόμβους ακρών, β) επίθεση απορρήτου κατά τη διάρκεια ανταλλαγής μηνυμάτων εκπαίδευσης μεταξύ τοπικού κόμβου άκρου και διακομιστή άκρων, γ) επίθεση από πράκτορα μάθησης (Mukherjee, 2020).

Scheme	Security goal	Cryptographic mechanism	Advantages (+) Limitations (–)
Data privacy scheme [11]	• Data privacy	• Based on Hash-Solomon code	+ Protection of fragmentary information – Authentication is not considered
Anonymous and secure aggregation scheme [9]	• Secure aggregation • User/identity privacy • User authentication	• Elliptic curves • Castagnos-Laguillaumie cryptosystem	+ Confidentiality and privacy – Limited threats are considered
Scheme for access control [12]	• Authorized access and control	• Attribute-based access control	+ Implements security as a service – Complexity
Storage privacy scheme [13]	• Multi-layer privacy	• Computational intelligence	+ Preserving data privacy and aggregation – Anonymity is not considered
Privacy preserving fog-enabled aggregation [14]	• Multi-layer privacy	• Stream ciphers and PKI • Homomorphic encryption	+ Privacy preserving – Restricted attack capability

Εικόνα 3. 2 Σύνοψη Ζητημάτων Ασφαλείας (Mukherjee, 2020)



Εικόνα 3. 3 Σοβαρότητα Απειλών (Mothukuri, 2020)

Κεφάλαιο 4: Τρόποι Αντιμετώπισης

Κατά τη διάρκεια της μαθησιακής διαδικασίας υπάρχουν αρκετά ζητήματα απορρήτου και ασφάλειας και μπορεί κανείς να διευκρινίσει τις αντίστοιχες μεθόδους προστασίας σε τρεις κατηγορίες:

- προστασία της ιδιωτικής ζωής στον πελάτη πλευρά,
- προστασία της ιδιωτικής ζωής στο διακομιστή και
- προστασία ασφαλείας για το FL.

4.1 Προστασία απορρήτου από την πλευρά του πελάτη

Στο σύστημα FL, οι πελάτες ανεβάζουν τα μαθησιακά τους αποτελέσματα συμπεριλαμβανομένων των τιμών των παραμέτρων στο διακομιστή, αλλά μπορεί να μην εμπιστεύονται τον διακομιστή. Αυτό συμβαίνει, καθώς ένας κακόβουλος διακομιστής μπορεί να εκμεταλλευτεί τα μεταφορτωμένα δεδομένα για να συμπεράνει ιδιωτικές πληροφορίες. Για την ανακούφιση αυτής της ανησυχίας, οι πελάτες μπορούν να χρησιμοποιήσουν ορισμένες τεχνολογίες διατήρησης της ιδιωτικής ζωής ως εξής:

Διαταραχή (Perturbation): Η ιδέα της διαταραχής προσθέτει θόρυβο στις παραμέτρους που ανεβάζουν οι πελάτες. Αυτή η γραμμή εργασίας συχνά χρησιμοποιεί διαφορικό απόρρητο για να αποκρύψει ορισμένα ευαίσθητα χαρακτηριστικά έως ότου το τρίτο μέρος δεν είναι σε θέση να διακρίνει το άτομο, κάνοντας έτσι τα δεδομένα αδύνατο να αποκατασταθούν έτσι ώστε να προστατευτεί το απόρρητο του χρήστη.

Στην έρευνα των Geyer et al (2017), οι συγγραφείς παρουσίασαν μια διαφορετική προσέγγιση απορρήτου στο FL με τη σειρά για να προσθέσουν προστασία σε δεδομένα από τον πελάτη. Ωστόσο, η ρίζα αυτών των μεθόδων εξακολουθεί να απαιτεί ότι τα

δεδομένα διαβιβάζονται αλλού. Αυτό συνήθως συνεπάγεται ανταλλαγή μεταξύ της ακρίβειας και του απορρήτου, το οποίο χρειάζεται προσαρμογές (Ma, 2020).

Dummy: Η έννοια της εικονικής μεθόδου πηγάζει από την προστασία της τοποθεσίας της ιδιωτικής ζωής (Kido, 2005). Παράμετροι εικονικού μοντέλου μαζί με το πραγματικό στέλνονται στον διακομιστή από πελάτες, οι οποίοι ενδέχεται να αποκρύψουν τη συνεισφορά του πελάτη κατά τη διάρκεια της εκπαίδευσης. Λόγω της αθροιστικής επεξεργασίας στο διακομιστή, η απόδοση μπορεί ακόμα να είναι εγγυημένη.

4.2 Προστασία απορρήτου στο διακομιστή

Μετά τη συλλογή ενημερωμένων παραμέτρων από πελάτες, ο διακομιστής εκτελεί έναν σταθμισμένο μέσο όρο σε αυτές τις παραμέτρους ανάλογα με το μέγεθος των δεδομένων. Ωστόσο, όταν ο διακομιστής μεταδίδει τις συγκεντρωτικές παραμέτρους σε πελάτες για τον συγχρονισμό μοντέλων, οι πληροφορίες μπορεί να διαρρεύσουν καθώς ενδέχεται να υπάρχουν υποκλοπές. Έτσι, η προστασία από την πλευρά του διακομιστή είναι επίσης σημαντική.

Συγκέντρωση: Η βασική ιδέα της συγκέντρωσης είναι η συλλογή δεδομένων ή παραμέτρων μοντέλου από διαφορετικούς πελάτες από την πλευρά του διακομιστή. Μετά συγκεντρωτικά, οι αντίπαλοι ή ο μη αξιόπιστος διακομιστής δεν μπορούν να επιθεωρήσουν τις πληροφορίες του πελάτη σύμφωνα με αυτές τις συγκεντρωτικές παραμέτρους. Επιπλέον, σε ορισμένα σενάρια, ο διακομιστής έχει την ελευθερία επιλογής πελατών με υψηλή ποιότητα παραμέτρων ή με μη ευαίσθητες απαιτήσεις. Ωστόσο, το ερώτημα για το πώς να σχεδιάσει κανείς έναν κατάλληλος μηχανισμό συνάθροισης εξακολουθεί να είναι πρόκληση για τα τρέχουσα συστήματα FL.

Ασφαλής υπολογισμός πολλαπλών μερών (SMC): Το root του SMC χρησιμοποιεί κρυπτογράφηση για να κάνει μη αναμενόμενες ενημερώσεις μεμονωμένων συσκευών από έναν διακομιστή, αντί να αποκαλύπτει μόνο το άθροισμα μετά από επαρκή αριθμό ενημερώσεων (Rosulek, 2017). Αναλυτικά, το SMC είναι ένα τετρακύλινδρο δια δραστικό πρωτόκολλο που προαιρετικά ενεργοποιείται κατά τη διάρκεια της φάσης αναφοράς

μιας δεδομένης επικοινωνίας. Σε κάθε γύρο πρωτοκόλλου, ο διακομιστής συλλέγει μηνύματα από όλες τις συσκευές και, στη συνέχεια, χρησιμοποιεί το σύνολο των μηνυμάτων της συσκευής για τον υπολογισμό μιας ανεξάρτητης απάντησης και επιστροφής σε κάθε συσκευή. Έπειτα, οι συσκευές μεταφορτώνουν κρυπτογραφικά καλυμμένες ενημερώσεις μοντέλων στο διακομιστή. Τέλος, υπάρχει μια φάση οριστικοποίησης όπου οι συσκευές αποκαλύπτουν επαρκή κρυπτογραφικά στοιχεία για να επιτρέψουν στον διακομιστή να ξεκαθαρίσει τη συγκεντρωτική ενημέρωση του μοντέλου.

4.3 Προστασία ασφαλείας για το πλαίσιο FL

Όσο για την ασφάλεια ολόκληρου του πλαισίου του συστήματος FL, εξετάζονται κυρίως οι επιθέσεις «κλοπής» μοντέλων. Συγκεκριμένα, οποιοσδήποτε συμμετέχων στο σύστημα FL μπορεί να παρουσιάσει κρυφή λειτουργικότητα backdoor στο κοινό παγκόσμιο μοντέλο. Κατά συνέπεια, εκεί είναι επίσης ορισμένα προστατευτικά μέτρα για την ασφάλεια του σχεδιασμού του συστήματος FL.

Ομομορφική Κρυπτογράφηση: Αυτού του είδους η κρυπτογράφηση (Papernot, 2016) υιοθετείται για την προστασία του χρήστη στην ανταλλαγή δεδομένων μέσω παραμέτρων υπό τον μηχανισμό κρυπτογράφησης. Δηλαδή, οι παράμετροι κωδικοποιούνται πριν από τη μεταφόρτωση, και απαιτούνται επίσης κλειδιά αποκωδικοποίησης δημόσιου-ιδιωτικού τομέα, τα οποία ενδέχεται να προκαλέσουν επιπλέον κόστος στην επικοινωνία.

Back-door Defender: Υφιστάμενες άμυνες κατά των επιθέσεων backdoor δεν είναι αποτελεσματικές καθώς τα περισσότερα από αυτά απαιτούν πρόσβαση στα δεδομένα εκπαίδευσης. Επιπλέον, το σύστημα FL δεν μπορεί να βεβαιώσει ότι όλοι οι πελάτες δεν είναι κακόβουλοι και ότι δεν έχουν καμία ορατότητα στο τι κάνουν οι συμμετέχοντες τοπικά και έτσι δεν μπορεί να εμποδίζει κανέναν από τον έλεγχο ενημερώσεων των συμμετεχόντων στο κοινό μοντέλο.

Έτσι, τα ακόλουθα πρέπει να εξεταστούν:

• Θα πρέπει να παρέχονται θεωρητικά αποτελέσματα της σύγκλιση του συστήματος FL που διατηρεί την προστασία της ιδιωτικής ζωής.

• Μαθησιακές επιδόσεις, δηλαδή η ακρίβεια μάθησης, οι κύκλοι επικοινωνίας και οι παραλλαγές στις λειτουργίες απώλειας, πρέπει να διερευνηθούν.

• Ο αλγόριθμος προστασίας της ιδιωτικής ζωής, τόσο θεωρητικά όσο και εμπειρικά, πρέπει να επινοηθεί. Επιπλέον, η ανταλλαγή μεταξύ του απορρήτου και η ταχύτητα σύγκλισης χρειάζεται επίσης περαιτέρω έρευνα.

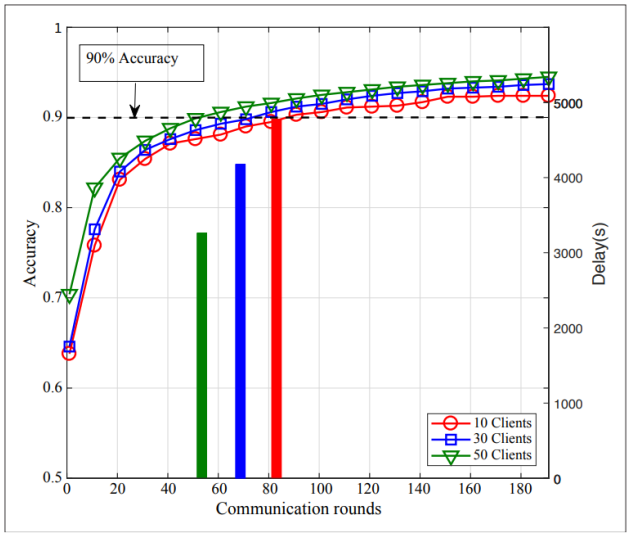
Υπάρχουν τρεις βασικοί τρόποι για να αποτρέψει κανείς τη δηλητηρίαση δεδομένων σε ένα σύστημα FL με γνώμονα την προστασία της ιδιωτικής ζωής. Το πρώτο είναι η αναγνώριση κακόβουλων πελατών όταν ρυθμίζεται το σύστημα. Σε αυτό το σενάριο, μπορούν να χρησιμοποιηθούν τεχνικές μάθησης. Για παράδειγμα, μπορεί να χρησιμοποιηθεί ένας εποπτευόμενος αλγόριθμος μάθησης για να βρεθούν οι κακόβουλοι πελάτες κατά τη διάρκεια κάθε γύρου επικοινωνίας.

Μια άλλη τεχνική επικεντρώνεται στη διαδικασία συνάθροισης. Μετά από κάθε συγκέντρωση, σύμφωνα με την ποιότητα των μεταφορτωμένων παραμέτρων μάθησης από τους πελάτες, ο διακομιστής μπορεί να προσαρμόσει το βάρος της συνάθροισης για κάθε πελάτη. Με αυτόν τον τρόπο, ο διακομιστής είναι σε θέση να αποκτήσει περισσότερη εμπιστοσύνη σε πελάτες που είναι πιο χρήσιμοι για την επίτευξη γρήγορης σύγκλισης και καλής μαθησιακής απόδοσης. Ο τρίτος και τελευταίος τρόπος είναι να εφαρμοστούν έννοιες από κοινωνικά δίκτυα ενημέρωσης σε κάθε γύρο επικοινωνίας αξιοποιώντας την κοινωνική σχέση κάθε πελάτη στη συνολική απόδοση του συστήματος (Ma, 2020).

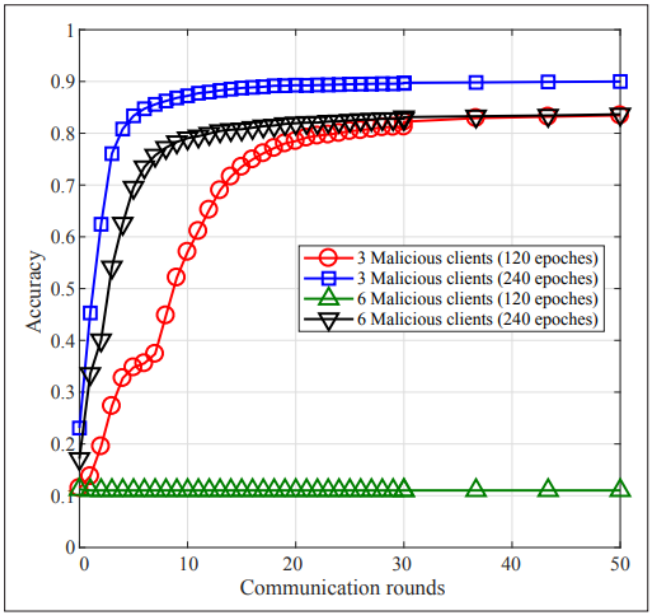
Όσο αφορά το πρόβλημα της κλιμάκωσης, μια πολλά υποσχόμενη μέθοδος για την αντιμετώπιση του μεγάλου χρόνου αναμονής είναι η ρύθμιση προθεσμίας καθυστέρησης μεταφόρτωσης για κάθε πελάτη. Σε κάθε εποχή εκμάθησης, ο διακομιστής θα συλλέγει τουλάχιστον τις απαιτούμενες παραμέτρους των πελατών πριν γίνει η εκτέλεση για τον επόμενο γύρο της FL. Εάν ο χρόνος αναμονής υπερβαίνει αυτήν την προθεσμία, ο τρέχων γύρος μάθησης εγκαταλείπεται (Ma, 2020).

Επιπλέον, για την αντιμετώπιση του μεγάλου αριθμού πελατών, μπορεί κανείς να χρησιμοποιήσει την έννοια της ομαδοποίησης των χρηστών. Με διαχωρισμό των πελατών σε διαφορετικά σμήνη (ομάδες) πλασματικά, σε κάθε σύμπλεγμα οι πελάτες ανταγωνίζονται μεταξύ τους για να ολοκληρώσουν τον στόχο. Ο διακομιστής θα παρέχει επίσης οφέλη σε αντάλλαγμα. Σε αυτόν τον νέο σχεδιασμό δομών, ο μεγάλος αριθμός πελατών θα διαχωριστεί από τα κοινά τους ενδιαφέροντα, με παρόμοια φυσική τοποθεσία ή τις ίδιες μεθόδους μεταφόρτωσης (Ma, 2020).

Στις εικόνες 4.1 και 4.2 οι Ma et al προτείνουν μια έξυπνη μέθοδο συγκέντρωσης για την αντιμετώπιση του προβλήματος των κακόβουλων πελατών. Ο προτεινόμενος αλγόριθμος περιλαμβάνει δύο μέρη. Το πρώτο μέρος περιλαμβάνει την προσθήκη μιας διαδικασίας δοκιμής από την πλευρά του διακομιστή, και ενημερώνει το βάρος συνάθροισης σύμφωνα με την απόδοση της δοκιμής για τις παραμέτρους που ανεβάζει ο κάθε πελάτης. Το δεύτερο μέρος περιλαμβάνει την αύξηση των τοπικών epochs για κάθε πελάτη. Όπως φαίνεται στο σχήμα, ο προτεινόμενος αλγόριθμος μπορεί να μετριάσει την απόδοση υποβάθμισης που προκαλείται από τους κακόβουλους πελάτες. Επιπλέον, χρειάζονται περισσότερα epochs όταν υπάρχουν περισσότεροι κακόβουλοι πελάτες στο σύστημα FL (Ma, 2020).



Εικόνα 4. 1 Σύγκριση απόδοσης με διαφορετικό αριθμός πελατών στο CNN (Ma, 2020)



Εικόνα 4. 2 Σύγκριση απόδοσης με διαφορετικό αριθμός κακόβουλων πελατών βάσει της προτεινόμενης μεθόδου συγκέντρωσης στο CNN (Ma, 2020)

Κεφάλαιο 5: Πειραματική Εφαρμογή στη βάση σεναρίου

Στην παρούσα ενότητα γίνεται λόγος για την πειραματική εφαρμογή που αποτελεί μέρος της παρούσας διπλωματικής εργασίας. Αρχικά, αναφέρεται ο σκοπός, η περιγραφή του σεναρίου, η μεθοδολογία και τέλος τα συμπεράσματα.

5.1 Σκοπός

Σκοπός της πειραματικής εφαρμογής είναι η επίλυση ενός θέματος ασφαλείας πάνω στο σύστημα Federated Learning (FL). Το ζήτημα που επιλέχθηκε είναι το data poisoning. Όπως ήδη αναφέρθηκε και με βάση τους Biggo et al (2012) το data poisoning είναι μια επίθεση δηλητηρίασης δεδομένων κατά αλγορίθμων ML. Ο επιτιθέμενος στοχεύει στην ευπάθεια του αλγορίθμου διανυσματικών μηχανών υποστήριξης και προσπαθεί να ενσωματώσει κακόβουλα δεδομένα στη φάση της εκπαίδευσης με την ελπίδα να μεγιστοποιηθεί το σφάλμα ταξινόμησης. Από τότε, μια μεγάλη ποικιλία προσεγγίσεων έχει προταθεί για τον μετριασμό των επιθέσεων δηλητηρίασης δεδομένων στους αλγόριθμους ML με διαφορετικές ρυθμίσεις.

Σε ένα σύστημα FL επιτρέπεται στους πελάτες (clients) να έχουν ενεργό ρόλο στην εκπαίδευση (training) δεδομένων και στην αποστολή παραμέτρων μοντέλου στο διακομιστή. Παράλληλα, όμως, δίνει αυτή την ευκαιρία και σε κακόβουλους πελάτες. Αυτοί μπορούν να δηλητηριάσουν το παγκόσμιο μοντέλο χειραγωγώντας τη διαδικασία εκπαίδευσης.

5.2 Μεθοδολογία

Για την επίλυση του προβλήματος αυτού, που μπορεί να επηρεάσει δραματικά την απόδοση και την αποτελεσματικότητα του μοντέλου, οι Muñoz-González et al (2019) σε μια δημοσίευση που αφορά το machine learning (ML) γενικότερα αναφέρουν την τεχνική του model averaging για την εξάλειψη των κακόβουλων χρηστών.

Ο απλούστερος τρόπος εφαρμογής της ομοσπονδιακής εκπαίδευσης είναι η τοπική εκπαίδευση και, στη συνέχεια, η μέση μέτρηση των μοντέλων. Αυτό χρησιμοποιεί τα ίδια δομικά στοιχεία. Σημειώστε ότι στην πλήρη λειτουργία του Federated Averaging που παρέχεται από το tff.learning, αντί να υπολογίζεται ο μέσος όρος των μοντέλων, προτιμούμε το μέσο όρο των delta των μοντέλων, για διάφορους λόγους, π.χ. για συμπίεση.

Η μέθοδος του model averaging του μοντέλου είναι μια τεχνική μάθησης που μειώνει τη διακύμανση σε ένα τελικό μοντέλο νευρωνικού δικτύου, θυσιάζοντας την απόδοση του μοντέλου για την αύξηση της ασφάλειας, την οποία σε άλλη περίπτωση δεν θα περιμέναμε από το μοντέλο.

5.3 Υλοποίηση πειραματικής εφαρμογής

Για την υλοποίηση έγινε χρήση του TensorFlow, το οποίο είναι μια δωρεάν βιβλιοθήκη λογισμικού ανοιχτού κώδικα για εφαρμογές μηχανικής μάθησης. Μπορεί να χρησιμοποιηθεί σε ένα ευρύ φάσμα εργασιών, αλλά δίνει ιδιαίτερη έμφαση στην εκπαίδευση και την εξαγωγή συμπερασμάτων βαθιών νευρωνικών δικτύων.

Το πρώτο βήμα είναι η δημιουργία του virtual environment, το οποίο είναι είναι ένα εργαλείο που βοηθά στη διατήρηση των πακέτων που απαιτούνται από διαφορετικά έργα-projects ξεχωριστά δημιουργώντας απομονωμένα εικονικά περιβάλλοντα python για αυτά.

```
sudo apt update

sudo apt install python3-dev python3-pip # Python 3

sudo pip3 install --user --upgrade virtualenv

virtualenv --python python3 "venv"

source "venv/bin/activate"

pip install --upgrade pip
```

Ακολουθεί η εγκατάσταση του πακέτου 2.4.1 στο Tensorflow:

```
pip install --upgrade tensorflow_federated
```

Στην διαδικασία αυτή γίνεται χρήση του Federated Core (FC), το οποίο είναι ένα πακέτο που καθιστά δυνατή τη συμπαγή έκφραση λογικής προγράμματος που συνδυάζει τον κώδικα TensorFlow με κατανεμημένους χειριστές επικοινωνίας. Ο κύριος στόχος του tf.contrib.distribute είναι να επιτραπεί στους χρήστες να χρησιμοποιούν τα υπάρχοντα μοντέλα και τον κώδικα εκπαίδευσης με ελάχιστες αλλαγές για να καταστήσουν δυνατή την κατανεμημένη εκπαίδευση.

Για να καταστήσει τον υπάρχοντα εκπαιδευτικό κώδικα πιο αποτελεσματικό, ο στόχος του Federated Core της TFF (TensorFlow Federated) είναι να δώσει στους ερευνητές και τους επαγγελματίες σαφή έλεγχο των συγκεκριμένων μοτίβων της κατανεμημένης επικοινωνίας που θα χρησιμοποιούν στα συστήματά τους. Η εστίαση στο FC είναι στην παροχή μιας ευέλικτης και επεκτάσιμης γλώσσας για την έκφραση αλγορίθμων ροής κατανεμημένων δεδομένων.

Το TFF στο σύνολό του στοχεύει σενάρια στα οποία διανέμονται δεδομένα και πρέπει να παραμείνουν τέτοια, όπου η συλλογή όλων των δεδομένων σε κεντρική τοποθεσία ενδέχεται να μην είναι βιώσιμη επιλογή. Αυτό έχει επιπτώσεις στην εφαρμογή αλγορίθμων μηχανικής μάθησης που απαιτούν αυξημένο βαθμό σαφούς ελέγχου, σε σύγκριση με σενάρια στα οποία όλα τα δεδομένα μπορούν να συσσωρευτούν σε μια κεντρική τοποθεσία σε ένα κέντρο δεδομένων.

Για την εκπαίδευση του συστήματος χρησιμοποιείται η βιβλιοθήκη MNIST. Η βάση δεδομένων MNIST περιέχει χειρόγραφα ψηφία, 60.000 παραδείγματα και ένα δοκιμαστικό σύνολο 10.000 παραδειγμάτων. Είναι ένα υποσύνολο ενός μεγαλύτερου σετ που διατίθεται από το NIST. Τα ψηφία έχουν κανονικοποιηθεί στο μέγεθος και έχουν κεντραριστεί σε μια εικόνα σταθερού μεγέθους. Έτσι, σκοπός είναι η μετάφραση ή μετατροπή των εικόνων αυτών σε ψηφία μέσω του συστήματος του Federated Learning. Επιπλέον, η προσομοίωση περιλαμβάνει ένα σενάριο στο οποίο υπάρχουν δεδομένα από 10 χρήστες και καθένας από τους χρήστες συμβάλλει στη γνώση του τρόπου αναγνώρισης ενός διαφορετικού ψηφίου.

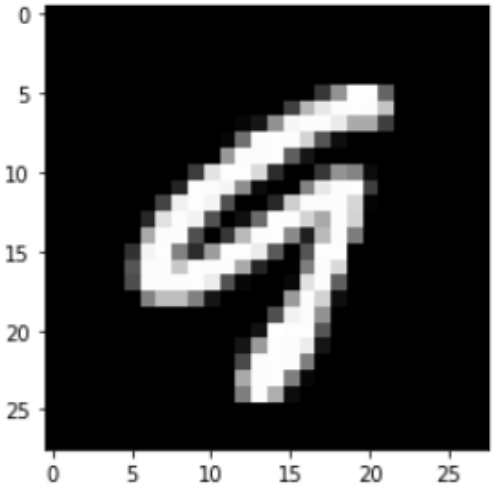
Αρχικά, γίνεται φόρτωση των δεδομένων MNIST:

```
mnist_train, mnist_test = tf.keras.datasets.mnist.load_data()

[(x.dtype, x.shape) for x in mnist_train]
```

Τα δεδομένα επιστρέφουν ως πίνακες, ένας με εικόνες και ένας άλλος με ετικέτες ψηφίων. Κάθε εικόνα ενσωματώνεται σε έναν φορέα 784 στοιχείων, αφού κάθε εικόνα

έχει 28*28 pixels (όπως φαίνεται στην εικόνα 5.1) και είναι ασπρόμαυρες εικόνες άρα το εύρος είναι από 0 εώς 255 (όπως εξάγεται από το σύστημα). Για να είναι πολύ πιο εύκολο να επεξεργαστούν τα δεδομένα γίνεται μετατροπή του εύρος από 0 εώς 255 σε 0 εώς 1.



Εικόνα 5. 1 Χειρόγραφο ψηφίο

Επιπρόσθετα, μετατρέπουμε τον δισδιάστατο αυτό πίνακα σε μονοδιάστατο μέσω της μεθόδου flatten() του πακέτου numpy. Και αυτή η μέθοδος βοηθά στην ευκολότερη διαχείριση των δεδομένων.

```
def get_data_for_digit(source, digit):

  output_sequence = []

  all_samples = [i for i, d in enumerate(source[1]) if d == digit]

  for i in range(0, min(len(all_samples), NUM_EXAMPLES_PER_USER),
BATCH_SIZE):

    batch_samples = all_samples[i:i + BATCH_SIZE]

    output_sequence.append({

      'x':

        np.array([source[0][i].flatten() / 255.0 for i in batch_samples],

                 dtype=np.float32),
```

```
NUM_EXAMPLES_PER_USER = 1000

BATCH_SIZE = 100

def get_data_for_digit(source, digit):

  output_sequence = []

  all_samples = [i for i, d in enumerate(source[1]) if d == digit]

  for i in range(0, min(len(all_samples), NUM_EXAMPLES_PER_USER), BATCH_SIZE):

    batch_samples = all_samples[i:i + BATCH_SIZE]

    output_sequence.append({

        'x':

            np.array([source[0][i].flatten() / 255.0 for i in batch_samples],

                     dtype=np.float32),

        'y':

            np.array([source[1][i] for i in batch_samples], dtype=np.int32)

    })

  return output_sequence

federated_train_data = [get_data_for_digit(mnist_train, d) for d in range(10)]

federated_test_data = [get_data_for_digit(mnist_test, d) for d in range(10)]
```

Έχοντας τα δεδομένα, καθορίζεται μια συνάρτηση απώλειας (loss function) όπου μπορεί κανείς να την χρησιμοποιήσει για εκπαίδευση. Είναι μια μέθοδος αξιολόγησης του πόσο καλά ένας συγκεκριμένος αλγόριθμος μοντελοποιεί τα δεδομένα. Εάν οι προβλέψεις αποκλίνουν πάρα πολύ από τα πραγματικά αποτελέσματα, η λειτουργία απώλειας θα έχει ως αποτέλεσμα έναν πολύ μεγάλο αριθμό. Σταδιακά, με τη βοήθεια μιας λειτουργίας βελτιστοποίησης, η λειτουργία απώλειας μαθαίνει να μειώνει το σφάλμα στην πρόβλεψη.

```
BATCH_SPEC = collections.OrderedDict(

    x=tf.TensorSpec(shape=[None, 784], dtype=tf.float32),

    y=tf.TensorSpec(shape=[None], dtype=tf.int32))

BATCH_TYPE = tff.to_type(BATCH_SPEC)

str(BATCH_TYPE)
```

Όταν καλείτε μια συνάρτηση Python διακοσμημένη με tff.tf_computation μέσα στο σώμα μιας άλλης τέτοιας λειτουργίας, η λογική του εσωτερικού TFF υπολογισμού ενσωματώνεται στη λογική του εξωτερικού. Εάν γράφει κανείς και τους δύο υπολογισμούς, είναι προτιμότερο να κάνει την εσωτερική λειτουργία μια κανονική λειτουργία Python ή tf από ότι μια tff.tf_.

```
@tf.function
def forward_pass(model, batch):
  predicted_y = tf.nn.softmax(
     tf.matmul(batch['x'], model['weights']) + model['bias'])
  return -tf.reduce_mean(
     tf.reduce_sum(
        tf.one_hot(batch['y'], 10) * tf.math.log(predicted_y),
axis=[1]))


@tff.tf_computation(MODEL_TYPE, BATCH_TYPE)
def batch_loss(model, batch):
  return forward_pass(model, batch)
```

Συνοπτικά, ο τρόπος σκέψης που ακολουθήθηκε είναι ο ακόλουθως. Πρώτα, δίνεται ένα ψηφίο σε κάθε client, για παράδειγμα στον client X το ψηφίο 5. Κάνοντας evaluation βρίσκει κανείς πως το loss είναι πολύ μικρό.

```
print('initial_model loss =', local_eval(initial_model,
                                         federated_train_data[5]))
print('locally_trained_model loss =',
      local_eval(locally_trained_model, federated_train_data[5]))
```

```
initial_model loss = 23.025854
locally_trained_model loss = 0.43484682
```

Στην συνέχεια κάνοντας evaluation για ένα άλλο ψηφίο πχ το ψηφίο 0, βλέπει κανείς πως το loss αυξάνεται. Αυτό είναι λογικό, αφού ο συγκεκριμένος client X δεν έχει «ξαναδεί» το ψηφίο αυτό.

```
print('initial_model loss =', local_eval(initial_model,
                                         federated_train_data[0]))
print('locally_trained_model loss =',
      local_eval(locally_trained_model, federated_train_data[0]))
```

```
initial_model loss = 23.025854
locally_trained_model loss = 74.50075
```

Όταν αυτό γίνει για όλους τους clients, υπολογίζεται το average loss. Η απόδοση του συστήματος προφανώς είναι καλύτερη, αλλά όχι η επιθυμητή. Για να βελτιωθεί το μοντέλο πρέπει όλοι οι clients να στείλουν τα αποτελέσματά τους στον server και αυτός με την σειρά του αναβαθμίζει το μοντέλο και δημιουργεί μια ανανεωμένη έκδοση του παλαιού μοντέλου. Για να εξαλειφθεί το παραπάνω υπολογίζεται ο μέσος όρος των μοντέλων.

```
print('initial_model loss =', federated_eval(initial_model,
                                              federated_train_data))
print('locally_trained_model loss =',
      federated_eval(locally_trained_model, federated_train_data))
```

```
initial_model loss = 23.025852
locally_trained_model loss = 54.432625
```

Η παραπάνω μέτρηση δεν είναι προφανώς αρκετή. Το επόμενο βήμα είναι ο server να στείλει το updated model στους clients και ξαναυπολογίζεται από τα outputs ο μέσος όρος. Στο παράδειγμα της παρούσας εργασίας αυτό γίνεται 10 φορές, δηλαδή έχουμε 10 φορές averaging του μοντέλου.

```
model = initial_model
learning_rate = 0.1
for round_num in range(10):
  model = federated_train(model, learning_rate, federated_train_data)
  learning_rate = learning_rate * 0.9
  loss = federated_eval(model, federated_train_data)
  print('round {}, loss={}'.format(round_num, loss))
```

```
round 0, loss=21.60552406311035
round 1, loss=20.365676879882812
round 2, loss=19.274803161621094
round 3, loss=18.311105728149414
round 4, loss=17.45724868774414
round 5, loss=16.69875717163086
round 6, loss=16.023344039916992
round 7, loss=15.420503616333008
round 8, loss=14.881217956542969
round 9, loss=14.397723197937012
```

```
print('initial_model test loss =',
      federated_eval(initial_model, federated_test_data))
print('trained_model test loss =', federated_eval(model, federated_test_data))
```

```
initial_model test loss = 22.795593
trained_model test loss = 14.24283
```

5.4 Αποτελέσματα - Συζήτηση

Οι Blanchard et al. (2017) έδειξαν ότι το πρότυπο του Federated Learning (FL) αποτυγχάνει με την παρουσία ελαττωματικών και κακόβουλων πελατών. Έτσι, μόνο ένας κακός πελάτης μπορεί να θέσει σε κίνδυνο το ολόκληρη την απόδοση και τη σύγκλιση του κοινόχρηστου μοντέλου. Για τον μετριασμό αυτού του περιορισμού, δηλαδή του data poisoning, έχουν ήδη προταθεί στη βιβλιογραφία διαφορετικές ισχυρές ομόσπονδες στρατηγικές μάθησης (Blanchard et al. 2017; Mhamdi, Guerraoui and Rouault 2018).

Μερικές από αυτές τις τεχνικές βασίζονται σε ισχυρά στατιστικά στοιχεία (π.χ. διάμεσοι εκτιμητές) για ενημέρωση του συγκεντρωτικού μοντέλου (Blanchard et al. 2017; Mhamdi, Guerraoui and Rouault 2018; Yin et al. 2018), τα οποία μπορεί να είναι υπολογιστικά ακριβά για μεγάλα μοντέλα και τον αριθμό των πελατών σε σύγκριση σε τυπικούς κανόνες συγκέντρωσης, όπως ομόσπονδος μέσος όρος (McMahan et al. 2017).

Αυτές οι τεχνικές αγνοούν επίσης το κλάσμα των σημείων εκπαίδευσης που παρέχονται από κάθε πελάτη, κάτι που μπορεί να είναι περιοριστικό σε περιπτώσεις όπου οι πελάτες παρέχουν σημαντικά διαφορετική ποσότητα σημείων δεδομένων για την εκπαίδευση του μοντέλου.

Συμπεράσματα – Επίλογος

Η ομοσπονδιακή μάθηση είναι μια νέα τεχνολογία που υποστηρίζει την τεχνολογία AI σε συσκευές αποκεντρωμένης μάθησης. Το σύστημα FL προτάθηκε να επεκτείνει τα οφέλη της μηχανικής μάθησης σε τομείς με ευαίσθητα δεδομένα. Η παρούσα εργασία παρέχει μια ολοκληρωμένη μελέτη σχετικά με τα επιτεύγματα, τα ζητήματα και τις επιπτώσεις στην ασφάλεια και την προστασία της ιδιωτικής ζωής στο περιβάλλον των συστημάτων FL. Με την αξιολόγηση τα αποτελέσματα για τα ζητήματα ασφάλειας και το απόρρητο, σκοπός είναι να δοθούν νέες προοπτικές.

Το σύστημα FL είναι ένα σχετικά νέο πλαίσιο που κυκλοφόρησε στην αγορά και χρειάζεται περαιτέρω έρευνα για την διερεύνηση τρόπων βελτίωσης που ταιριάζει σε διαφορετικά στυλ περιβάλλοντος FL. Η ομοσπονδιακή μάθηση έχει ένα σύνολο προκλήσεων που χρειάζονται περαιτέρω έρευνα. Με βάση τις σχετικές παρατηρήσεις και τα αποτελέσματά της παρούσας εργασίας, εντοπίστηκαν τα ακόλουθα ζητήματα που θα μπορούσαν να αποτελούν μελλοντικούς δρόμους έρευνας.

Οι τρέχουσες αμυντικές προσπάθειες στο σύστημα FL έχουν σχεδιαστεί για την προστασία από γνωστές ευπάθειες και συγκεκριμένες προκαθορισμένες κακόβουλες δραστηριότητες, καθιστώντας τις λιγότερο χρήσιμες κατά τον εντοπισμό επιθέσεων εκτός των παραμέτρων σχεδιασμού τους όταν δοκιμάζονται. Παρόλο που αυτό το φαινόμενο ισχύει για σχεδόν όλους τους μηχανισμούς άμυνας της εφαρμογής ML, η πιθανότητα είναι μεγαλύτερη στο σύστημα FL καθώς δεν υπάρχουν πολλές εκδόσεις στην παραγωγή που θα είχαν δείξει την πιθανότητα διαφόρων επιθέσεων. Τα επιτεύγματα που χρησιμοποιούν προηγμένη «deep» μάθηση έχουν δείξει πολλά υποσχόμενες λύσεις στην καταπολέμηση τέτοιων επιθέσεων.

Μια μεγάλη πρόκληση του FL είναι η ανιχνευσιμότητα του παγκόσμιου μοντέλου ML σε ολόκληρο τον κύκλο ζωής της υποκείμενης διαδικασίας ML. Για παράδειγμα, εάν μια τιμή

πρόβλεψης αλλάξει στο παγκόσμιο μοντέλο ML, θα πρέπει να υπάρχει ικανότητα παρακολούθησης για να προσδιοριστεί ποιες τιμές συγκέντρωσης πελατών οδήγησαν σε αυτήν την αλλαγή. Εάν η λογική πίσω από τη συμπεριφορά του μοντέλου ML είναι ένα «μαύρο κουτί», τότε είναι κανείς αναγκασμένος να χάσει την λογική και βασίζεται τυφλά στην ανθρώπινη τεχνητή νοημοσύνη.

Το σύστημα FL είναι μια αρκετά νέα προσέγγιση που απαιτεί λεπτομερή ανάλυση όλων των επαγγελματιών και των μειονεκτημάτων όλων των διαφορετικών προσεγγίσεων. Οι τυποποιημένες τεχνικές πρέπει να οριστούν για να υποστηρίζουν τις αναδυόμενες απαιτήσεις του FL σε διαφορετικούς τομείς. Δεδομένου ότι η προστασία της ιδιωτικής ζωής είναι βασικός παράγοντας στο σύστημα FL, πρέπει να γίνει εστίαση σε περαιτέρω έρευνα σχετικά με την ενίσχυση της ιδιωτικής ζωής και την τυποποίηση προσεγγίσεων για κάθε απαίτηση.

Οι τρέχουσες ερευνητικές εργασίες δείχνουν τον τρόπο ενίσχυσης της προστασίας της ιδιωτικής ζωής στο σύστημα FL με κόστος της θυσίας της αποτελεσματικότητας ή της ακρίβειας. Ωστόσο, δεν υπάρχει έρευνα που εργάζεται για την εύρεση του κατάλληλου επιπέδου κρυπτογράφησης για SMC και την ποσότητα που προστέθηκε ο θόρυβος. Εάν το επίπεδο κρυπτογράφησης ή η ποσότητα θορύβου δεν είναι αρκετή, οι συμμετέχοντες εξακολουθούν να υποφέρουν από τον κίνδυνο διαρροής απορρήτου. Αντίθετα, εάν το επίπεδο κρυπτογράφησης είναι πολύ υψηλό ή έχει προστεθεί πολύς θόρυβος στις παραμέτρους, το μοντέλο υποφέρει σοβαρά από την χαμηλή ακρίβεια.

Υπάρχουν επί του παρόντος ορισμένα πλαίσια FL που μπορούν να χρησιμοποιηθούν για την εφαρμογή συστημάτων που βασίζονται σε FL όπως το TensorFlow Federated, το PySyft και το FATE. Εκτός από το PySyft, δεν υπάρχουν πλαίσια, βιβλιοθήκες, ή εργαλειοθήκες που μπορούν να ενσωματώσουν και να εκτελέσουν SMC ή DP αυτή την στιγμή. Έτσι, το σχέδιο κατάρτισης και η στρατηγική για την επιλογή πελατών για την εκπαίδευση του συστήματος είναι ζωτικής σημασίας στο FL.

Βιβλιογραφία

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., & Van Overveldt, T. (2019). Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046.

Bonomi F., Milito R., Zhu J., and Addepalli S., (2012). Fog computing and its role in the Internet of Things, Proc. SIGCOMM Workshop on Mobile Cloud Computing, doi: 10.1145/2342509.2342513

Blanchard, P., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems (pp. 119-129).

Brisimi T.S., Chen R., Mela T., Olshevsky A., I. C. Paschalidis, and W. Shi, (2018) Federated learning of predictive models from federated electronic health records, Int. J. Medical Informatics, vol. 112, pp. 59–67. doi: 10.1016/j.ijmedinf.2018.01.007.

Cheng, M., Singh, S., Chen, P., Chen, P. Y., Liu, S., & Hsieh, C. J. (2019). Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*.

Fang, W., & Yang, B. (2008, December). Privacy preserving decision tree learning over vertically partitioned data. In *2008 International Conference on Computer Science and Software Engineering* (Vol. 3, pp. 1049-1052). IEEE.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

GDPR (2018) Info. https://gdpr-info.eu/ 145, 148, 149

Hartmann, T., Moawad, A., Fouquet, F., & Le Traon, Y. (2019). The next evolution of MDE: a seamless integration of machine learning into domain modeling. Software & Systems Modeling, 18(2), 1285-1304.

Hatcher, W. G., & Yu, W. (2018). A survey of deep learning: Platforms, applications and emerging research trends. IEEE Access, 6, 24411-24432.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & d'Oliveira, R. G. (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

Kuflik T., Kay J., and Kummerfeld B., Challenges and solutions of ubiquitous user modeling, (2012), Ubiquitous Display Environments, A. Krüger and T. Kuflik, Eds. Berlin: Springer-Verlag, pp. 7–30.

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016a). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016b). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

McMahan H.B., Moore E., Ramage D., Hampson S., and Aguera y Arcas B., (2017), Communication-efficient learning of deep networks from decentralized data, Proc. 20th Int. Conf. Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

McMahan, R. B., Mackay, B. C., & Schmeckpeper, D. E. (2016a). U.S. Patent Application No. 29/507,877.

McMahan, R. B., Mackay, B. C., & Schmeckpeper, D. E. (2016b). *U.S. Patent Application No. 29/506,862*.

Mohassel, P., & Zhang, Y. (2017, May). Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 19-38). IEEE.

Muñoz-González, L., Co, K. T., & Lupu, E. C. (2019). Byzantine-robust federated machine learning through adaptive model averaging. arXiv preprint arXiv:1909.05125.

Mhamdi, E. M. E., Guerraoui, R., & Rouault, S. (2018). The hidden vulnerability of distributed learning in byzantium. arXiv preprint arXiv:1802.07927.

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, *22*(2), 199-210.

Pouyanfar, S., Tao, Y., Tian, H., Chen, S. C., & Shyu, M. L. (2019). Multimodal deep learning based on multiple correspondence analysis for disaster management. World Wide Web, 22(5), 1893-1911.

Piper, D. L. A. (2019). Data protection Laws of the world. 2019.

Rooney, S., Bauer, D., & Scotton, P. (2005, February). Edge server software architecture for sensor applications. In *The 2005 Symposium on Applications and the Internet* (pp. 64-71). IEEE.

Sattler, F., Wiedemann, S., Müller, K. R., & Samek, W. (2019, July). Sparse binary compression: Towards distributed deep learning with minimal communication. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Sheller M.J., Reina G.A., Edwards B., Martin J., and Bakas S., (2018), Multiinstitutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation, Proc. Int. MICCAI Brainlesion Workshop, pp. 92–104. doi: 10.1007/978-3-030-11723-8_9.

Trask, N. A., & Huang, A. (2019). Physics informed machine learning at SNL (No. SAND2019-14338PE). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Vaidya, J., & Clifton, C. (2004, April). Privacy preserving naive bayes classifier for vertically partitioned data. In Proceedings of the 2004 SIAM international conference on data mining (pp. 522-526). Society for Industrial and Applied Mathematics.

van Engelen, J. E., van Lier, J. J., Takes, F. W., & Trautmann, H. (2018, September). Accurate WiFi-Based Indoor Positioning with Continuous Location Sampling. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 524-540). Springer, Cham.

WeBank AI (2019) Department, Federated AI technology enabler (FATE). https://github.com/F ederatedAI/FATE xv, 12, 14, 144.

Xu, K., Yue, H., Guo, L., Guo, Y., & Fang, Y. (2015, June). Privacy-preserving machine learning algorithms for big data systems. In *2015 IEEE 35th international conference on distributed computing systems* (pp. 318-327). IEEE.

Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature methods*, *16*(8), 687-694.

Zaman (2020) Retrieved from: https://medium.com/accenture-the-dock/instilling-responsible-and-reliable-ai-development-with-federated-learning-d23c366c5efd

Zhao, G., Pang, B., Xu, Z., Peng, D., & Xu, L. (2019). Assessment of urban flood susceptibility using semi-supervised machine learning model. *Science of The Total Environment*, *659*, 940-949.