

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Πτυχιακή Εργασία

**Εντοπισμός επιθέσεων και άμυνες στην τεχνητή
νοημοσύνη του διαδικτύου των πραγμάτων**

Άγγελος Σπυρίδων Κουρτέσης Π2020154

Επιβλέπων Καθηγητής: Χριστόφορος Νταντογιάν

Οκτώβριος 2024

Επιβλέπων

Νταντογιάν Χριστόφορος - Επίκουρος Καθηγητής

Τριμελής Επιτροπή

Μάγκος Εμμανουήλ - Καθηγητής
Νταντογιάν Χριστόφορος - Επίκουρος Καθηγητής
Τσώχου Αγγελική - Αναπληρώτρια Καθηγήτρια

Περίληψη

Η παρούσα εργασία ασχολείται με τη ραγδαία ανάπτυξη της τεχνητής νοημοσύνης (AI) και του Διαδικτύου των Πραγμάτων (IoT), αναδεικνύοντας τις δυνατότητές τους, αλλά και τους κινδύνους που εγκυμονούν. Η τεχνητή νοημοσύνη χρησιμοποιείται ευρέως σε διάφορους τομείς για την εκτέλεση σύνθετων εργασιών, όπως η αναγνώριση προτύπων και η λήψη αποφάσεων, ενώ το IoT συνδέει φυσικά αντικείμενα με το διαδίκτυο, επιτρέποντάς τους να «επικοινωνούν» μεταξύ τους. Ωστόσο, οι τεχνολογίες αυτές δεν είναι απαλλαγμένες από κινδύνους, όπως η "δηλητηρίαση" δεδομένων (data poisoning), που μπορεί να επηρεάσει την αξιοπιστία των συστημάτων AI. Στη συγκεκριμένη μελέτη, εξετάζεται η επίθεση "label flipping", η οποία στοχεύει στην παραποίηση των δεδομένων ενός Intrusion Detection System (IDS) για IoT συσκευές, μειώνοντας την ακρίβειά του και κλονίζοντας την απόδοσή του σε κρίσιμες προβλέψεις, παρά το γεγονός ότι η γενική ακρίβεια του συστήματος παραμένει υψηλή.

Περιεχόμενα

1.) Εισαγωγή.....	6
2.) Τεχνητή Νοημοσύνη (AI).....	8
2.1) Μηχανική Μάθηση.....	8
2.2) Βαθιά μάθηση.....	9
2.3) Νευρωνικά δίκτυα.....	10
2.4) Συναρτήσεις ενεργοποίησης.....	12
2.5) Πίνακας Σύγχυσης.....	15
2.6) Binary Cross-Entropy.....	18
3.) Internet of Things (IoT).....	19
3.1) Perception Layer.....	20
3.2) Network Layer.....	20
3.3) Application Layer.....	21
4.) Intrusion Detection Systems (IDS).....	22
4.1) Anomaly models.....	22
4.2) Misuse models.....	23
4.3) Specification-based models.....	23
4.4) Αρχιτεκτονική IDS.....	24
5.) Data poisoning.....	25
5.1) Backdoor Attack.....	25
5.2) Label Flipping Attack.....	26
5.3) Targeted Poisoning Attack.....	27
5.4) Availability Attack.....	28
5.5) Clean-label Poisoning Attack.....	29

6.) Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review.....	31
7.) Σκοπός της έρευνας.....	37
7.1) Το σύνολο των δεδομένων.....	38
7.2) Υλοποίηση του Intrusion Detection System & επίθεση.....	39
7.3) Αποτελέσματα.....	42
7.4) Εμπόδια και αντιμετώπιση αυτών.....	43
8.) Συμπεράσματα Έρευνας.....	45
Βιβλιογραφία.....	46
Εικόνες.....	48

1. Εισαγωγή

Τα τελευταία χρόνια, ολοένα και περισσότεροι κλάδοι υιοθετούν την τεχνητή νοημοσύνη με σκοπό βελτιστοποίηση διαδικασιών, την υποστήριξη αποφάσεων και την γενικότερη διευκόλυνση της καθημερινότητας. Η εξέλιξη της τεχνητής νοημοσύνης είναι ραγδαία το τελευταίο διάστημα και την συναντάμε όλοι μας κατά κόρον. Από το chatbot που μπορεί να γράφει ποιήματα και να υλοποιήσει εργασίες ημερών σε ελάχιστα δευτερόλεπτα, τα smartphones που συμπεριλαμβάνουν πλέον υπηρεσίες AI, μέχρι τα έξυπνα αυτοκίνητα που οδηγούν μόνα τους και τα ρομπότ που είναι σε θέση να δώσουν συνέντευξη στην τηλεόραση και να εκφωνήσουν μία ομιλία. Όμως, πόσο καλά γνωρίζουμε τι είναι η τεχνητή νοημοσύνη; Πόσο αξιόπιστη είναι; Ένας μεγάλος αριθμός του πληθυσμού είναι αρκετά διστακτικός ή και αρνητικός με την διείσδυση της τεχνητής νοημοσύνης στην ζωή μας. Σε μία συνέντευξη του ο επικεφαλής του τμήματος Επιστήμης Δεδομένων και Βαθιάς Μάθησης της Treelogic και επίκουρος καθηγητής στο τμήμα Θεωρίας Σήματος και Επικοινωνιών του Πανεπιστημίου Carlos III της Μαδρίτης Roberto Diaz Morales είπε:

«Ας σκεφτούμε με ένα μαχαίρι. Είναι απλώς ένα εργαλείο που μπορεί να χρησιμοποιηθεί για πολλούς διαφορετικούς σκοπούς. Κανείς δεν φοβάται ένα μαχαίρι ή είναι απαισιόδοξος για την ύπαρξη μαχαιριών, οι άνθρωποι τα χρησιμοποιούν καθημερινά όταν τρώνε μεσημεριανό γεύμα.

Με τον ίδιο τρόπο. Η τεχνητή νοημοσύνη είναι ένα καλό εργαλείο που κάνει χρήση των δεδομένων για την επίτευξη ενός συγκεκριμένου στόχου. Αν οι στόχοι αυτοί συμβαδίζουν με τα δικά μας οφέλη, μπορούμε να οικοδομήσουμε έναν πιο βιώσιμο πλανήτη (π.χ. μείωση των βιομηχανικών αποβλήτων με τη βελτιστοποίηση της κατασκευαστικής παραγωγής ή μείωση της ρύπανσης με τη βελτιστοποίηση της κυκλοφορίας).

Το πρόβλημα εδώ είναι ότι όλοι καταλαβαίνουν τι είναι ένα μαχαίρι και έχουμε νομοθεσία για τη σωστή χρήση τους. Αλλά η τεχνητή νοημοσύνη είναι μια πιο σύνθετη έννοια. Πρέπει λοιπόν πρώτα να διδάξουμε στον πληθυσμό τι ακριβώς είναι και τι μπορεί να κάνει η «τεχνητή νοημοσύνη». Και πρέπει επίσης να δημιουργήσουμε τη σωστή νομοθεσία, διότι επί του παρόντος οι νομικοί περιορισμοί δημιουργούνται από πολιτικούς που δεν έχουν απολύτως καμία ιδέα για την «τεχνητή νοημοσύνη» και είναι συνήθως πολύ χαλαροί σε ορισμένες πτυχές ή πολύ περιοριστικοί σε άλλες.»

Ουσιαστικά πρέπει κατανοήσουμε τι είναι η τεχνητή νοημοσύνη πριν την καταδικάσουμε ή την εκθειάσουμε αντίστοιχα, για όπως θα φανεί παρακάτω ακόμα και αυτή μπορεί να κάνει λάθη.

Από την άλλη πλευρά, άλλος ένας κλάδος που βρίσκει την ίδια απήχηση ακούει στο όνομα Διαδίκτυο των Πραγμάτων (Internet of Things). Όλοι έχουμε σπίτι μας κάποια “έξυπνη” συσκευή όπως, ένα ερκοντίσιον το οποίο μπορούμε να ελέγξουμε από το “έξυπνο” κινητό ή μία “έξυπνη” τηλεόραση που επικοινωνεί με το κινητό για να συνεχίσουμε το επεισόδιο από το σημείο που το αφήσαμε. Είναι αλήθεια πως το IoT μας κάνει την ζωή πιο εύκολη και όχι μόνο σε οικιακό επίπεδο. Για παράδειγμα στην υγεία, ένας ασθενής με καρδιακά προβλήματα μπορεί να φοράει έναν έξυπνο καρδιολογικό αισθητήρα ή μια συσκευή, όπως ένα έξυπνο ρολόι, που είναι συνδεδεμένο μέσω IoT. Αυτή η συσκευή παρακολουθεί συνεχώς τα ζωτικά του σημεία, όπως τον καρδιακό ρυθμό, την αρτηριακή πίεση και τα επίπεδα οξυγόνου στο αίμα. Τα δεδομένα από αυτές τις συσκευές μεταδίδονται αυτόματα σε πραγματικό χρόνο σε έναν κεντρικό υπολογιστή ή στο σύστημα παρακολούθησης του ιατρικού προσωπικού. Εάν ανιχνευτεί κάποια ανωμαλία, όπως αρρυθμίες ή αυξημένη πίεση, το σύστημα μπορεί να ειδοποιήσει αμέσως τον γιατρό ή να ενεργοποιήσει ένα αυτόματο σήμα έκτακτης ανάγκης προς τον ασθενή για να λάβει άμεσα ιατρική φροντίδα. Όμως, τι θα γινόταν αν κάποιος κακόβουλος χρήστης αποκτούσε μη εξουσιοδοτημένη πρόσβαση και παραποιούσε τα δεδομένα με αποτέλεσμα να μην είναι έγκυρα; Ή σε ένα άλλο σενάριο, να τα κρυπτογραφούσε ζητώντας κάποιο αντίτιμο για την αποκρυπτογράφησή τους; Τόσο το διαδίκτυο των πραγμάτων όσο και η τεχνητή νοημοσύνη υιοθετούνται συνεχώς από όλο και περισσότερους κλάδους. Αυτή η ενσωμάτωση εκτός από πολλά ωφέλει εγκυμονεί ακόμα περισσότερους κινδύνους. Σε αυτή την εργασία, θα αναλυθεί μία τεχνική επίθεσης της κατηγορίας της “δηλητηρίασης” των δεδομένων (data poisoning) γνωστή ως αντιστροφή ετικετών (label flipping) και πώς αυτή μπορεί να “μπερδέψει” ένα σύστημα τεχνητής νοημοσύνης που σκοπός του είναι η ανίχνευση επιθέσεων σε συσκευές (IoT).

2. Τεχνητή νοημοσύνη

Είναι δύσκολο να δοθεί ένας ακριβής ορισμός για την τεχνητή νοημοσύνη καθώς υπάρχουν διαφορετικές προσεγγίσεις της. Μερικοί την έχουν ορίσει με βάση την πιστότητά της με την ανθρώπινη απόδοση, ενώ άλλοι προτιμούν έναν ορισμό από την πλευρά της ορθολογικότητας, δηλαδή να κάνει το “σωστό”. Ένας ορισμός που θα μπορούσε να δοθεί είναι πως η τεχνητή νοημοσύνη (TN) είναι μια τεχνολογία που μελετά και αναπτύσσει θεωρίες, μεθόδους, τεχνολογίες και συστήματα εφαρμογών που χρησιμοποιούνται για την προσομοίωση, την επέκταση και την εξάπλωση της ανθρώπινης νοημοσύνης. Άλλωστε, η τεχνητή νοημοσύνη, κλάδος της επιστήμης των υπολογιστών, στοχεύει στην αποκάλυψη της ουσίας της νοημοσύνης και στην παραγωγή μιας νέας ευφυούς μηχανής που θα μπορεί να αντιδρά με τρόπο παρόμοιο με την ανθρώπινη νοημοσύνη. Επιπλέον, η έρευνα στον τομέα αυτό περιλαμβάνει τη ρομποτική, την αναγνώριση γλώσσας, την αναγνώριση εικόνας, την επεξεργασία φυσικής γλώσσας και τα εξειδικευμένα συστήματα, την ασαφή λογική, τον έλεγχο νευρωνικών δικτύων, τον έλεγχο μάθησης και την ιεραρχία.[1],[2]

2.1. Μηχανική μάθηση

Η μηχανική μάθηση (machine learning) είναι μία υποκατηγορία της τεχνητής νοημοσύνης, που επιδιώκει να κατασκευάσει αλγόριθμους οι οποίοι επιτρέπουν στα συστήματα να μαθαίνουν από δεδομένα, χωρίς να απαιτείται ο προγραμματισμός τους για κάθε συγκεκριμένη εργασία. Στον πυρήνα της μηχανικής μάθησης, βρίσκεται η ικανότητα προσαρμογής και βελτίωσης της απόδοσης μέσω εμπειρίας. Η βασική ιδέα πίσω από τη μηχανική μάθηση είναι ότι αντί να γράφονται σαφείς εντολές και κανόνες για το πώς θα εκτελεστεί μια εργασία, ο αλγόριθμος μαθαίνει από παραδείγματα ή δεδομένα που του παρέχονται. Με άλλα λόγια, το σύστημα δεν είναι προγραμματισμένο για κάθε συγκεκριμένο σενάριο που θα αντιμετωπίσει, αλλά μαθαίνει μοτίβα από τα δεδομένα και στη συνέχεια τα χρησιμοποιεί για να πάρει αποφάσεις ή να κάνει προβλέψεις για νέα δεδομένα. Η μετάβαση από ένα συγκεκριμένο σύνολο παρατηρήσεων σε έναν γενικό κανόνα ονομάζεται επαγωγή. Όταν η έξοδος του μοντέλου είναι ένα πεπερασμένο σύνολο τιμών (ΝΑΙ-ΟΧΙ, ΑΛΗΘΕΣ-ΨΕΥΔΕΣ κ.α.), το πρόβλημα της μάθησης ονομάζεται ταξινόμηση. Αν, είναι ένας αριθμός (π.χ. θερμοκρασία) το πρόβλημα λέγεται παλινδρόμηση. Υπάρχουν τρεις τύποι μάθησης που είναι οι εξής:

- **Επιβλεπόμενη μάθηση (supervised learning):** Το σύστημα παρατηρεί ζεύγυι εισόδου-εξόδου και μαθαίνει μία συνάρτηση που αντιστοιχίζει εισόδους με εξόδους, ώστε όταν έρθει αντιμέτωπο με νέα δεδομένα να μπορεί να προβλέψει τις κατάλληλες εξόδους. Στην περίπτωση αναγνώρισης εικόνας, οι εισοδοί είναι φωτογραφίες που η κάθε μία συνοδεύεται από μία έξοδο(π.χ. γάτα - σκύλος) την ονομαζόμενη ετικέτα.
- **Μη επιβλεπόμενη μάθηση (unsupervised learning):** Σε αυτόν τον τύπο μάθησης δεν υπάρχουν ετικέτες για τις εξόδους στα δεδομένα. Το σύστημα πρέπει να μάθει τις κρυμμένες δομές ή μοτίβα από τα ίδια τα δεδομένα. Η πιο συνηθισμένη εργασία αυτής της κατηγορίας είναι η ομαδοποίηση. Δηλαδή, όταν το σύστημα “βλέπει” πολλές ίδιες εικόνες (π.χ. γάτα) μπορεί να προσδιορίσει μία ομάδα παρόμοιων εικόνων.
- **Ενισχυτική μάθηση (reinforcement learning):** Το μοντέλο μαθαίνει δοκιμάζοντας, μέσα από τιμωρίες και ανταμοιβές. Για παράδειγμα, σε μία παρτίδα σκακιού αν κερδίσει θα έχει ανταμοιβή, ειδικά αν θα έχει τιμωρία. Είναι ζήτημα του συστήματος να κρίνει ποιές ενέργειες οδήγησαν στο αποτέλεσμα και να τις κρατήσει ή να τις αλλάξει, στοχεύοντας σε περισσότερες ανταμοιβές.[1]

2.2. Βαθιά μάθηση

Η πιο σημαντική υποκατηγορία τεχνικών μηχανικής μάθησης είναι η βαθιά μάθηση (deep learning). Σε αυτή την φάση οι υποθέσεις μετατρέπονται σε πολύπλοκα αλγεβρικά κυκλώματα με ρυθμιζόμενες συνδυαστικές δυνάμεις. Ο όρος “βαθιά” προκύπτει από το γεγονός ότι συνήθως αυτά τα κυκλώματα είναι οργανωμένα σε επίπεδα, δηλαδή οι διαδρομές των υπολογισμών από τις εισόδους στις εξόδους έχουν πολλά βήματα. Στις μέρες μας χρησιμοποιείται κυρίως για εργασίες όπως οπτική αναγνώριση, μηχανική μετάφραση, αναγνώριση ομιλίας κ.α. Επίσης έχει σημαντικό ρόλο στις εφαρμογές ενισχυτικής μάθησης. Η βασική ιδέα της μηχανικής μάθησης είναι η εκπαίδευση κυκλωμάτων τέτοιων ώστε οι διαδρομές υπολογισμών να είναι μεγάλες, με αποτέλεσμα όλες οι μεταβλητές εισόδου να αλληλεπιδρούν με πολύπλοκους τρόπους. Αυτού που είδους τα κυκλώματα τείνουν να είναι επαρκώς εκφραστικά ώστε να αποτυπώσουν την πολυπλοκότητα των δεδομένων του πραγματικού κόσμου για έναν μεγάλο αριθμό προβλημάτων. Παρόλο που οι μέθοδοι της γραμμικής και λογιστικής παλινδρόμησης μπορούν να διαχειρίζονται μεγάλα πλήθη μεταβλητών εισόδου, η διαδρομή των υπολογισμών από κάθε είσοδο προς την έξοδο είναι πολύ μικρή. Γίνεται πολλαπλασιασμός με μόνο ένα βάρος και μετά προστίθεται το αποτέλεσμα την συνολική έξοδο. Ακόμα, οι διαφορετικές μεταβλητές συμβάλλουν ανεξάρτητα στην έξοδο χωρίς να αλληλεπιδρούν μεταξύ τους. Αυτό

συμβάλλει αρνητικά στην στην εκφραστική ισχύ αυτών των μοντέλων. Δηλαδή, μπορούν να αναπαραστήσουν μόνο γραμμικές συναρτήσεις και σύνορα στον χώρο εισόδου αλλά στον πραγματικό κόσμο τα πράγματα είναι αρκετά πιο σύνθετα. Τα δίκτυα που είναι εκπαιδευμένα με μεθόδους βαθιάς μάθησης συχνά ονομάζονται νευρωνικά δίκτυα.[1]

2.3. Νευρωνικά δίκτυα

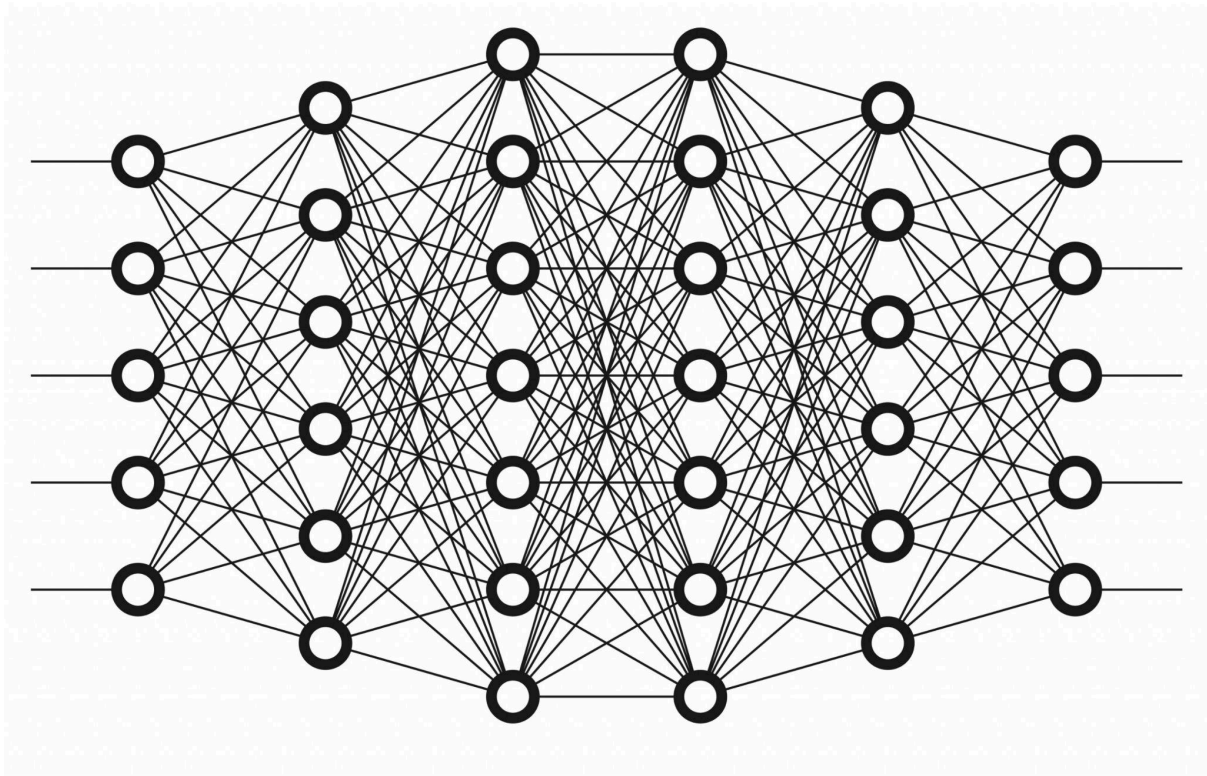
Τα νευρωνικά δίκτυα είναι υπολογιστικά μοντέλα εμπνευσμένα από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου, σχεδιασμένα για την αναγνώριση προτύπων και την εκμάθηση από δεδομένα. Ένα νευρωνικό δίκτυο αποτελείται από πολλαπλά στρώματα νευρώνων (ή κόμβων), τα οποία είναι διασυνδεδεμένα και μετασχηματίζουν τα δεδομένα από τα στρώματα εισόδου στα στρώματα εξόδου μέσω ζυγισμένων συνδέσεων. Κάθε νευρώνας λαμβάνει ένα σύνολο εισόδων, τις οποίες συνδυάζει και εφαρμόζει μια μη γραμμική συνάρτηση ενεργοποίησης για να παραχθεί η έξοδος. Τα βάρη των συνδέσεων προσαρμόζονται κατά τη διάρκεια της διαδικασίας εκπαίδευσης του δικτύου μέσω αλγορίθμων όπως η ανάδρομη διάδοση (backpropagation), με στόχο τη βελτιστοποίηση της απόδοσης του δικτύου σε ένα συγκεκριμένο πρόβλημα. Τα νευρωνικά δίκτυα χρησιμοποιούνται εκτενώς σε εφαρμογές όπως η αναγνώριση εικόνων, η επεξεργασία φυσικής γλώσσας και η πρόβλεψη τάσεων. Αποτελούν τη βάση της βαθιάς μάθησης (deep learning), η οποία περιλαμβάνει πολυεπίπεδα δίκτυα, γνωστά ως βαθιά νευρωνικά δίκτυα (deep neural networks). Υπάρχουν πολλές κατηγορίες νευρωνικών δικτύων. Οι δύο κυριότερες από αυτές είναι τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks) και τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks).

Τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης (FFN) είναι μία από τις πιο απλές μορφές νευρωνικού δικτύου, είναι ένα νευρωνικό δίκτυο με περισσότερες εφαρμογές στην πρόβλεψη και μοντελοποίηση των χαρακτηριστικών των τηλεπικοινωνιακών δικτύων και των προτύπων κίνησης από οποιοδήποτε άλλο νευρωνικό δίκτυο. Κάθε νευρώνας σε ένα επίπεδο συνδέεται προς μία κατεύθυνση. Αυτό σημαίνει σχηματίζει έναν κατευθυνόμενο μη-κυκλικό γράφο με συγκεκριμένους κόμβους εισόδου και εξόδου. Κάθε νευρώνας υπολογίζει μία συνάρτηση των δεδομένων εισόδου του και στέλνει το αποτέλεσμα στους διαδοχικούς του. Ουσιαστικά οι πληροφορίες “ρέουν” από τους κόμβους εισόδου, στους κόμβους εξόδου χωρίς την παρεμβολή βρόγχων. Τα δίκτυα αυτού του είδους δεν εκμεταλλεύονται τη χωρική δομή των δεδομένων (π.χ., τη γεωμετρία των εικόνων).

Στα συνελικτικά νευρωνικά δίκτυα (CNN), οι νευρώνες δεν συνδέονται πλήρως με κάθε νευρώνα του επόμενου επιπέδου. Αντίθετα επεξεργάζονται μικρές περιοχές (τοπικά patches) της εισόδου, όπως μικρές περιοχές μιας εικόνας, τουλάχιστον στα πρώτα επίπεδα. Αυτό γίνεται γιατί έχουν πρότυπα μοτίβα βαρών που αναπαράγονται σε όλες τις μονάδες (κόμβους) του κάθε επιπέδου. Ένα τέτοιο μοτίβο που αναπαράγεται σε πολλές τοπικές περιοχές ονομάζεται πυρήνας. Αυτό επιτρέπει στο δίκτυο να διατηρεί πληροφορίες για τη χωρική δομή των δεδομένων (π.χ., το πώς τα pixel είναι οργανωμένα σε μια εικόνα). Τα CNN είναι εξαιρετικά επιτυχημένα σε πρακτικές εφαρμογές. Το όνομα “συνελικτικό νευρωνικό δίκτυο” υποδηλώνει ότι το δίκτυο χρησιμοποιεί μια μαθηματική πράξη που ονομάζεται συνέλιξη. Η συνέλιξη είναι ένα εξειδικευμένο είδος γραμμικής πράξης. Στα πλαίσια των νευρωνικών δικτύων πρόκειται για την διαδικασία εφαρμογής του πυρήνα στα pixel μιας εικόνας. Η πράξη της συνέλιξης γράφεται με την χρήση του συμβόλου “ * ” Για παράδειγμα, έστω ότι διάνυσμα εισόδου x με μέγεθος n (pixels μονοδιάστατης εικόνας) και διάνυσμα πυρήνα k με μέγεθος l η εξίσωση γράφεται ως $z = x * k$ και μπορεί να οριστεί ως εξής [1].[3]:

$$Z_i = \frac{\sum_{j=1}^i k_j x_{j+i} - (l+1)}{2}$$

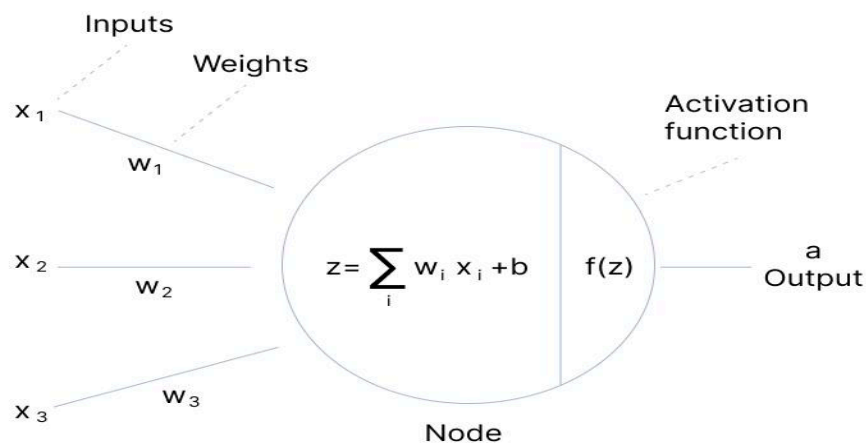
Παρακάτω, στην εικόνα 1 υπάρχει μία οπτική απεικόνιση ενός πλήρως συνδεδεμένου νευρωνικού δικτύου καθώς όλοι οι κόμβοι είναι συνδεδεμένοι μεταξύ τους. Η πρώτη στήλη είναι το επίπεδο εισόδου, οι επόμενες τέσσερις είναι τα “κρυφά” επίπεδα και τέλος το επίπεδο εξόδου.



Εικόνα 1

2.4. Συναρτήσεις ενεργοποίησης

Η συνάρτηση ενεργοποίησης (activation function) περιγράφεται ως μια μαθηματική συνάρτηση που καθορίζει την έξοδο ενός τεχνητού νευρώνα σε ένα νευρωνικό δίκτυο, δεδομένου του συνδυασμένου αθροίσματος των εισόδων του και των αντίστοιχων βαρών, όπως φαίνεται και στην εικόνα 2. Ο ρόλος της συνάρτησης ενεργοποίησης είναι να εισάγει μη γραμμικότητα στο σύστημα, επιτρέποντας στο νευρωνικό δίκτυο να μάθει και να αναπαραστήσει πιο σύνθετες σχέσεις στα δεδομένα. Χωρίς αυτήν τη μη γραμμικότητα, το δίκτυο θα μπορούσε μόνο να υπολογίσει γραμμικούς μετασχηματισμούς, κάτι που θα το καθιστούσε λιγότερο ισχυρό. Υπάρχουν πολλές συναρτήσεις ενεργοποίησης.[1]



V7 Labs

Εικόνα 2

Οι πιο συνήθεις είναι οι εξής:

- Η λογιστική ή σιγμοειδής (sigmoid) συνάρτηση, η οποία μετατρέπει τις εξόδους σε ένα εύρος μεταξύ 0 και 1. Επίσης, χρησιμοποιείται και στην λογιστική παλινδρόμηση. Ο τύπος της έχει ως εξής:

$$\sigma(x) = 1 / (1 + e^{-x})$$

- Η συνάρτηση ReLU (rectified linear unit) η οποία επιστρέφει μηδενική τιμή για αρνητικές εισόδους και την τιμή εισόδου για θετικές. Ο τύπος της έχει ως εξής:

$$ReLU(x) = \max(0, x)$$

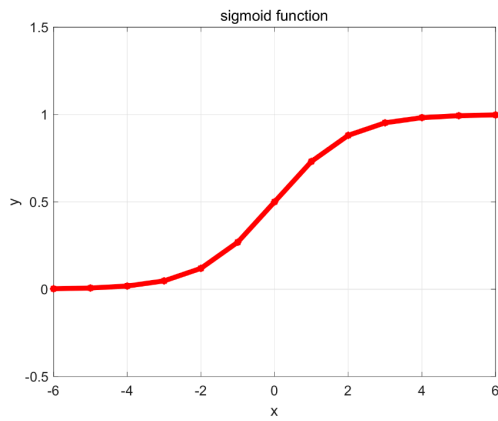
- Η συνάρτηση softplus, όπου πρόκειται για μία πιο ομαλή εκδοχή την ReLU και όπου η παραγωγός της είναι η σιγμοειδής συνάρτηση:

$$softplus(x) = \log(1 + e^x)$$

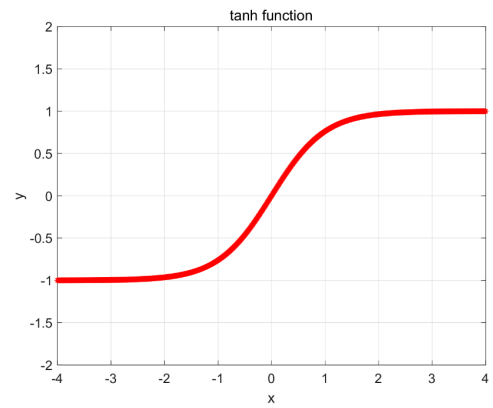
- Η συνάρτηση tanh που είναι μία μετατοπισμένη παραλλαγή της σιγμοειδούς συνάρτησης η οποία μετατρέπει τις εξόδους σε ένα εύρος μεταξύ -1 και 1:

$$\frac{e^{2x} - 1}{e^{2x} + 1}$$

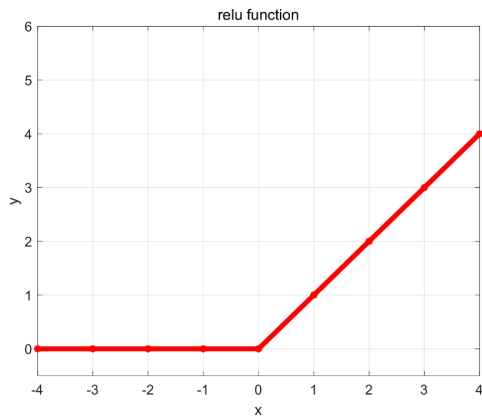
Στην εικόνα 3 υπάρχει η γραφική αναπαράσταση της κάθε συνάρτησης



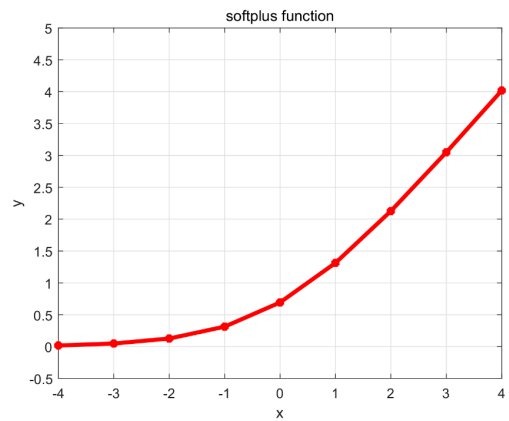
(a) The curve of sigmoid function



(b) The curve of tanh function



(c) The curve of ReLU function



(d) The curve of softplus function

Εικόνα 3

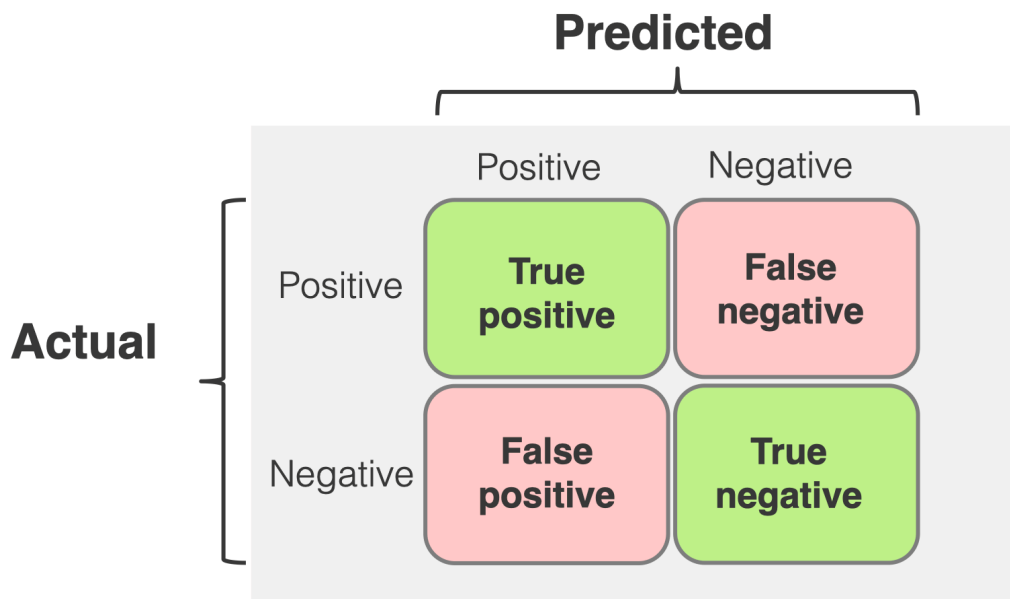
2.5. Πίνακας Σύγχυσης

Ο πίνακας σύγχυσης (confusion matrix) είναι ένα εργαλείο που χρησιμοποιείται στην αξιολόγηση της απόδοσης ταξινομητικών αλγορίθμων στην τεχνητή νοημοσύνη και τη μηχανική μάθηση. Πρόκειται για έναν πίνακα δισδιάστατης διάταξης (2X2), στον οποίο οι γραμμές αντιστοιχούν στις πραγματικές κατηγορίες και οι στήλες στις προβλεπόμενες κατηγορίες (εικ. 4). Ο πίνακας περιλαμβάνει τέσσερις βασικές τιμές:

1. True Positives (TP): Οι περιπτώσεις όπου η κατηγορία προβλέφθηκε από τον αλγόριθμο σωστά ως θετική.
2. True Negatives (TN): Οι περιπτώσεις όπου η κατηγορία προβλέφθηκε σωστά ως αρνητική.
3. False Positives (FP): Οι περιπτώσεις όπου η κατηγορία προβλέφθηκε ως θετική, αλλά στην πραγματικότητα ήταν αρνητική (επίσης γνωστό ως σφάλμα τύπου I).
4. False Negatives (FN): Οι περιπτώσεις όπου η κατηγορία προβλέφθηκε ως αρνητική, αλλά στην πραγματικότητα ήταν θετική (επίσης γνωστό ως σφάλμα τύπου II).

Ο πίνακας σύγχυσης παρέχει μια αναλυτική εικόνα των επιδόσεων του αλγορίθμου, επιτρέποντας τον υπολογισμό μέτρων όπως η ακρίβεια (accuracy), η ευαισθησία (recall), η ειδικότητα (specificity), και η ακρίβεια θετικής πρόβλεψης (precision) (εικ. 5).[4]

Στην συνέχεια ακολουθούν εικόνες για την δομή ενός πίνακα σύγχυσης καθώς και το πώς υπολογίζονται οι δείκτες που προαναφέρθηκαν.



Εικόνα 4



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

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

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

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

$$F1\ score = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Εικόνα 5

Κάθε μία από τις παραπάνω μετρικές εστιάζει σε διαφορετική πτυχή της απόδοσης του αλγορίθμου. Η ακρίβεια (accuracy) είναι μια μετρική που μετρά πόσο συχνά ένα μοντέλο προβλέπει σωστά το αποτέλεσμα. Μπορεί να υπολογιστεί διαιρώντας τον αριθμό των σωστών προβλέψεων (True Positives) με τον συνολικό αριθμό των προβλέψεων.

- Πλεονεκτήματα:

Η ακρίβεια είναι μια χρήσιμη μέτρηση όταν υπάρχουν ισορροπημένες κλάσεις και εστιάζουμε στην συνολική «ορθότητα» του μοντέλου και όχι στην ικανότητα πρόβλεψης μιας συγκεκριμένης κλάσης.

Η ακρίβεια είναι εύκολη στην κατανόηση.

- Μειονεκτήματα:

Εάν οι κλάσεις στο σύνολο των δεδομένων δεν είναι κατανεμημένες ομοιόμορφα, η ακρίβεια είναι λιγότερο χρήσιμη, καθώς δίνει την ίδια βαρύτητα στην ικανότητα του μοντέλου να προβλέπει όλες τις κατηγορίες.

Η ακρίβεια σε τέτοιες περιπτώσεις μπορεί να είναι παραπλανητική και να συγκαλύψει τη χαμηλή απόδοση στην κατηγορία-στόχο.

Γι αυτό, χρησιμοποιούμε και άλλους δείκτες όπως η ακρίβεια θετικής πρόβλεψης (precision)

η οποία μετρά πόσο συχνά ένα μοντέλο προβλέπει σωστά τη θετική κλάση. Μπορεί να υπολογιστεί από το πηλίκο του αριθμού των σωστών θετικών προβλέψεων (True Positives) με τον συνολικό αριθμό των περιπτώσεων που το μοντέλο πρόέβλεψε ως θετικές (True positives και False positives).

- Πλεονεκτήματα:

Λειτουργεί καλά για προβλήματα με δυσαναλογίες στις κλάσεις, καθώς δείχνει την ορθότητα του μοντέλου στον εντοπισμό της κλάσης-στόχου.

Η ακρίβεια είναι χρήσιμη όταν το κόστος ενός False Positive αποτελέσματος είναι υψηλό.

- Μειονεκτήματα:

Το precision δεν λαμβάνει υπόψη τα False Negatives, με αποτέλεσμα να μην υπολογίζει τις περιπτώσεις που χάνουμε τη κλάση-στόχο

Άλλος ένας αξιοσημείωτος δείκτης μέτρησης είναι η ευαισθησία (recall). Είναι μια μετρική που μετρά πόσο συχνά ένα μοντέλο μηχανικής μάθησης αναγνωρίζει σωστά θετικές περιπτώσεις (True Positives) από όλα τα πραγματικά θετικά δείγματα στο σύνολο δεδομένων. Υπολογίζεται από την διαίρεση του αριθμού των True Positives με τον αριθμό των θετικών περιπτώσεων. Δηλαδή, τα True Positives (επιτυχώς αναγνωρισμένες περιπτώσεις) και τα False Negatives (περιπτώσεις που δεν εντοπίστηκαν).

- **Πλεονεκτήματα:**
Λειτουργεί καλά για προβλήματα με ανισόρροπες κλάσεις (όπως το precision), καθώς επικεντρώνεται στην ικανότητα του μοντέλου να βρίσκει αντικείμενα της ετικέτας-στόχου.
Το recall είναι χρήσιμη όταν το κόστος των False Negatives είναι υψηλό.
- **Μειονεκτήματα:**
Δεν συνυπολογίζει το κόστος αυτών των False Positives

2.6. Binary Cross-Entropy

Η Δυαδική Διασταυρούμενη Εντροπία (Binary Cross-Entropy) είναι μια συνάρτηση μέτρησης απώλειας που χρησιμοποιείται κυρίως σε μοντέλα ταξινόμησης για δυαδικά προβλήματα, δηλαδή προβλήματα όπου η έξοδος είναι μία από δύο κατηγορίες (0 ή 1). Είναι ένας μετρικός τρόπος σύγκρισης της απόδοσης του μοντέλου στις προβλέψεις του σε σχέση με τις πραγματικές τιμές. Ο ορισμός της βασίζεται στην έννοια της εντροπίας και μετράει τη διαφορά μεταξύ των προβλεπόμενων πιθανοτήτων του μοντέλου και των πραγματικών τιμών (ετικετών).

Η μαθηματική της εξίσωση έχει ως εξής:

$$BCE = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)]$$

όπου:

- N είναι το πλήθος των παρατηρήσεων (δεδομένα εκπαίδευσης),
- y_i είναι η πραγματική ετικέτα (0 ή 1),
- p_i είναι η προβλεπόμενη πιθανότητα ότι η ετικέτα είναι 1.

Η Δυαδική Διασταυρούμενη Εντροπία (Binary Cross-Entropy) είναι ευαίσθητη στις λανθασμένες προβλέψεις, ειδικά όταν το μοντέλο έχει δείγματα “επικίνδυνα” για μια λάθος πρόβλεψη (δηλαδή, όταν η προβλεπόμενη πιθανότητα είναι πολύ κοντά στο 1 ή στο 0). [2]

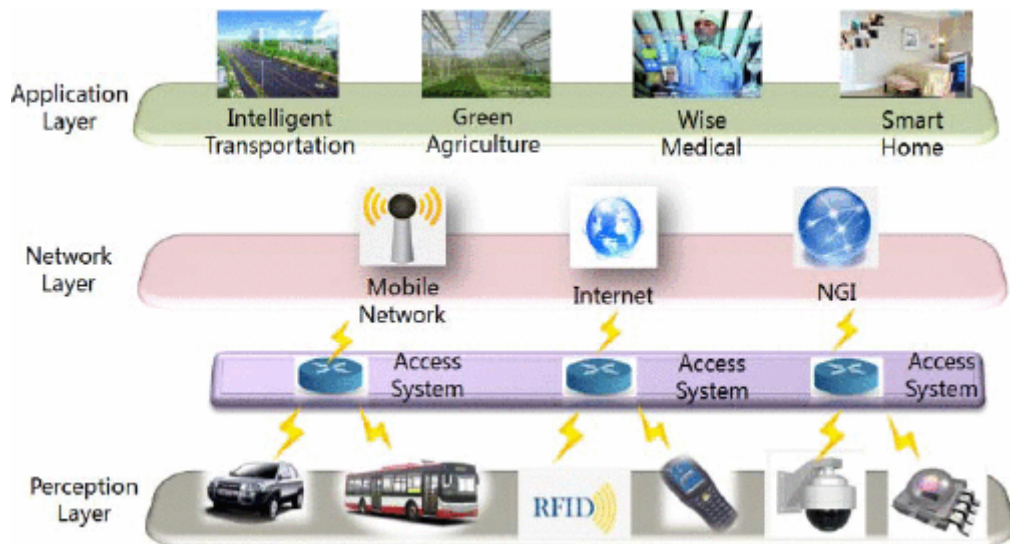
3. Internet of Things (IoT)

Η έννοια του Διαδικτύου των Πραγμάτων (Internet of Things) υλοποιείται, ως ένας αυξανόμενος αριθμός φυσικών αντικειμένων που συνδέεται με το Διαδίκτυο με ένα πρωτοφανή ρυθμό.

Οι θερμοστάτες και τα συστήματα παρακολούθησης και ελέγχου HVAC (θέρμανσης, αερισμού και κλιματισμού), τα οποία διέπουν τα έξυπνα σπίτια, είναι ένα απλό παράδειγμα τέτοιων αντικειμένων. Υπάρχουν περαιτέρω κλάδοι και περιβάλλοντα στα οποία το IoT μπορεί να έχει σημαντικό αντίκτυπο και να αναβαθμίσει την ποιότητα της ζωής μας.

Οι μεταφορές, η υγειονομική περίθαλψη, ο βιομηχανικός αυτοματισμός και η αντιμετώπιση εκτάκτων αναγκών σε φυσικές και ανθρωπογενείς καταστροφές, όταν η λήψη αποφάσεων από τον άνθρωπο είναι δύσκολη, είναι μερικές από αυτές τις χρήσεις. Επιτρέποντας στα φυσικά πράγματα να «μιλούν» μεταξύ τους, να μοιράζονται πληροφορίες και να συντονίζουν ενέργειες, το IoT τους δίνει τη δυνατότητα να βλέπουν, να ακούν, να σκέφτονται και να εκτελούν εργασίες.[5] Το Διαδίκτυο των Πραγμάτων (IoT) είναι μια ανερχόμενη τεχνολογία που προτάθηκε για πρώτη φορά για τη μελέτη του RFID από τον Ashton, καθηγητή του Κέντρου Αυτόματης Αναγνώρισης του MIT το 1999. Το IoT είναι μια σημαντική κίνηση για την υποστήριξη της σύνθεσης υπηρεσιών με διάφορες εφαρμογές. Δίνει τη δυνατότητα στα αντικείμενα που μας περιβάλλουν να επικοινωνούν μεταξύ τους μέσω του Διαδικτύου. Η δημοφιλής αρχιτεκτονική του IoT απεικονίζεται στην παρακάτω εικόνα.

Αποτελείται από τρία επίπεδα (εικ. 6): επίπεδο αντίληψης, επίπεδο δικτύου και επίπεδο εφαρμογής. Οι αισθητήρες, οι ετικέτες RFID και άλλα έξυπνα τερματικά συνδέονται στο IoT από το στρώμα αντίληψης. Το στρώμα δικτύου είναι υπεύθυνο για την επικοινωνία μεταξύ των «πραγμάτων» και των ανθρώπων. Πληθώρα εφαρμογών παρέχεται από το στρώμα εφαρμογών.[6]



Εικόνα 6

3.1. Perception Layer

Το κύριο καθήκον του perception layer (επίπεδο αντίληψης) είναι η συλλογή των πληροφοριών των πραγμάτων γύρω μας που αποτελούν μέρος του IoT. Αυτή η διαδικασία συλλογής βασίζεται σε διάφορες τεχνολογίες ανίχνευσης (π.χ. RFID, WSN, GPS, NFC κ.λπ.). Επιπλέον, το επίπεδο αυτό είναι υπεύθυνο για τη μετατροπή των πληροφοριών σε ψηφιακά σήματα, ώστε να είναι πιο εύκολη η μετάδοση μέσω δικτύου. Σε ορισμένες περιπτώσεις που η συλλογή δεν μπορεί να γίνει άμεσα από τα αντικείμενα, ενσωματώνονται μικροτσίπ με σκοπό την ενίσχυση τους με δυνατότητες ανίχνευσης και επεξεργασίας.

3.2. Network Layer

Το network layer (επίπεδο δικτύου) είναι υπεύθυνο για την επεξεργασία των δεδομένων που λαμβάνονται από το perception layer (επίπεδο αντίληψης). Επιπλέον, είναι υπεύθυνο για τη μετάδοση δεδομένων στο application layer (επίπεδο εφαρμογής) μέσω διαφόρων τεχνολογιών δικτύου, όπως ασύρματα/ενσύρματα δίκτυα και τοπικά δίκτυα (LAN). Τα κύρια μέσα μετάδοσης περιλαμβάνουν τα FTTx, 3G/4G, Wifi, bluetooth, Zigbee, UWB, υπερύθρες κ.ο.κ. Καθώς οι ποσότητες των δεδομένων που μεταφέρονται είναι μεγάλες και η χρήση ενός αξιόπιστου ενδιάμεσου λογισμικού για την αποθήκευση και την επεξεργασία τους είναι απαραίτητη, σε αυτό το επίπεδο χρησιμοποιείται η υπολογιστική νέφος (cloud computing).

3.3. Application Layer

Το επίπεδο εφαρμογής λαμβάνει και χρησιμοποιεί τα επεξεργασμένα δεδομένα από το προηγούμενο επίπεδο. Στην πραγματικότητα, το στρώμα αυτό αποτελεί το front end ολόκληρης της αρχιτεκτονικής του IoT. Επιπλέον, το στρώμα αυτό παρέχει τα απαιτούμενα εργαλεία (π.χ. συσκευές ενεργοποίησης) στους προγραμματιστές για την υλοποίηση του IoT. Για να ανταποκριθεί στις ιδιαιτερότητες του IoT, η αρχιτεκτονική τριών επιπέδων παρέχει ένα πλαίσιο υψηλού επιπέδου μέσω του οποίου μπορούν να εφαρμοστούν διαφορετικές προσεγγίσεις.[7]

4. Intrusion Detection System

Η ανίχνευση εισβολών είναι η διαδικασία παρακολούθησης των συμβάντων που λαμβάνουν χώρα σε ένα σύστημα ή δίκτυο υπολογιστών και η ανάλυσή τους για ενδείξεις πιθανών συμβάντων, τα οποία είναι παραβιάσεις ή άμεσες απειλές παραβίασης των πολιτικών ασφάλειας υπολογιστών, των πολιτικών αποδεκτής χρήσης ή των τυποποιημένων πρακτικών ασφάλειας. Τα Intrusion Detection Systems (IDS) είναι συστήματα ασφαλείας υπολογιστών που χρησιμοποιούνται για την ανίχνευση ανεπιθύμητων ή κακόβουλων δραστηριοτήτων εντός ενός δικτύου ή συστήματος. Ο βασικός τους στόχος είναι να παρακολουθούν συνεχώς τη ροή των δεδομένων σε ένα δίκτυο ή τις ενέργειες σε ένα σύστημα και να εντοπίζουν ενδείξεις επιθέσεων ή παραβιάσεων ασφαλείας με την χρήση ενός ή περισσότερων μοντέλων εισβολών.

4.1. Anomaly models

Τα μοντέλα ανίχνευσης διαταραχών (anomaly detection) αποφαινόνται με βάση στατιστικά στοιχεία και ταξινομούν τις ενέργειες ή τις καταστάσεις που είναι στατιστικά ασυνήθιστες ως “κακές”, καθώς σε αυτά τα μοντέλα η απροσδόκητη συμπεριφορά του συστήματος αποτελεί στοιχείο εισβολής. Έχει υπάρξει εξ αρχής κάποιο μέτρο για τον χαρακτηρισμό μιας ενέργειας ως κανονική. Υπάρχουν διάφορα μοντέλα αυτής της κατηγορίας, όπως:

- Το μοντέλο τιμών Κατωφλίου (threshold metric) στο οποίο, υπάρχει ένα όριο από γεγονότα που μπορούν να συμβούν μία συγκεκριμένη χρονική στιγμή. Έστω m είναι η ελάχιστη τιμή και n η μέγιστη τότε, σε μια συγκεκριμένη χρονική περίοδο αν υπάρχουν λιγότερα από m γεγονότα ή περισσότερα από n , τότε η συμπεριφορά χαρακτηρίζεται ως διαταραγμένη. Για παράδειγμα, σε μία εφαρμογή που πληκτρολογείται λάθος ο κωδικός πρόσβασης πάνω από n φορές τότε γίνεται αποτροπή πρόσβασης στο σύστημα.
- Το Μοντέλο Στατιστικών Ροπών κάνει χρήση του μέσου (Ροπή 1ης τάξης), της τυπικής απόκλισης (Ροπή 2ης τάξης) και ενδεχομένως άλλων μέτρων συσχέτισης όπως ροπές υψηλότερων τάξεων. Αν οι τιμές βρίσκονται εκτός του διαστήματος της συγκεκριμένης ροπής, η συμπεριφορά θεωρείται διαταραγμένη. Τα συγκεκριμένα μοντέλα είναι πιο ευέλικτα από αυτά των τιμών κατωφλίου.
- Το Μοντέλο Markov εξετάζει ένα σύστημα σε μία συγκεκριμένη χρονική στιγμή. Το σύστημα βρίσκεται σε μία κατάσταση με βάση τα γεγονότα που έχουν προηγηθεί και μόλις συμβεί ένα καινούριο

γεγονός, τότε το σύστημα μεταβαίνει σε μία καινούργια κατάσταση. Με βάση τον χρόνο μπορούν να δημιουργηθούν πιθανότητες μετάβασης σε μία κατάσταση. Όταν ένα γεγονός προκαλή μία μετάβαση με μικρή πιθανότητα, τότε αυτό το γεγονός κρίνεται διαταραγμένο.

4.2. Misuse models

Τα μοντέλα κακής συμπεριφοράς (misuse models) συγκρίνουν ενέργειες και καταστάσεις με προϋπάρχουσες ακολουθίες που είναι ήδη γνωστό ότι αποτελούν εισβολές και τις ταξινομούν ως “κακές”. Τέτοιου είδους μοντέλα για την ανίχνευση της “κακής” συμπεριφοράς χρειάζεται να γνωρίζουν όλες τις ευπάθειες των συστημάτων ή τις δυνητικές ευπάθειες που προσπαθεί να εκμεταλλευτεί ένας επιτιθέμενος. Το μοντέλο μετατρέπει όλη αυτή την γνώση σε ένα σύνολο από κανόνες που ορίζουν αν βρίσκεται σε εξέλιξη μία πιθανή εισβολή. Το αρνητικό σε τέτοιου είδους συστήματα είναι δεν μπορούν να ανιχνεύσουν επιθέσεις που είναι άγνωστες στους δημιουργούς των κανόνων. Για να βελτιωθούν οι δυνατότητες ανίχνευσης των μοντέλων, αξιοποιήθηκαν προσαρμοστικές μέθοδοι με την χρήση νευρωνικών δικτύων και δικτύων Petri.

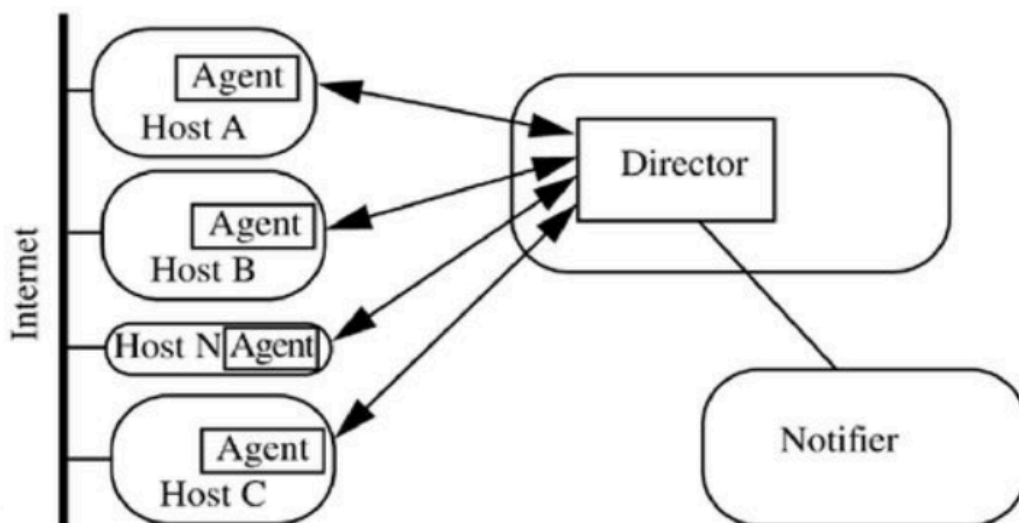
4.3. Specification-based models

Τα μοντέλα που βασίζονται στις προδιαγραφές (specification-based models) ταξινομούν τις καταστάσεις που παραβιάζουν τις προδιαγραφές ως “κακές”. Δηλαδή, αναζητά της καταστάσεις που είναι γνωστό ότι δεν είναι επιθυμητές μέσω κάποιων οδηγιών και όταν το σύστημα βρεθεί σε μία τέτοια κατάσταση που οι οδηγίες παραβιάζουν μία προδιαγραφή αναφέρεται ως πιθανή εισβολή. Για λόγους ασφάλειας, προσδιορίζονται και ελέγχονται μόνο τα προγράμματα που αλλάζουν την κατάσταση προστασίας των συστημάτων. Η ανίχνευση σε τέτοια μοντέλα στηρίζεται είτε στα ίχνη είτε στις ακολουθίες γεγονότων. Το ίχνος του συστήματος (system trace) ορίζεται ως μία ακολουθία γεγονότων κατά την διάρκεια εκτέλεσης ενός συστήματος διεργασιών.

4.4. Αρχιτεκτονική IDS

Τα IDS δεν εμποδίζουν απαραίτητα τις επιθέσεις, αλλά ειδοποιούν τους διαχειριστές συστημάτων όταν εντοπίζεται ύποπτη συμπεριφορά, ώστε να ληφθούν άμεσα μέτρα αντιμετώπισης. Ένα σύστημα ανίχνευσης εισβολών είναι παράλληλα και ένας μηχανισμός παρακολούθησης και ελέγχου. Αποτελείται από τρία μέρη (εικ. 6):

- Τον αντιπρόσωπο (agent) που αντιστοιχεί στον logger και αποκτά τις πληροφορίες από έναν στόχο.
- Τον διευθυντή (director) που έχει τον ρόλο του αναλυτή, δηλαδή αναλύει τα δεδομένα που λαμβάνει από τους αντιπροσώπους για να προσδιορίσει αν μία επίθεση είναι σε εξέλιξη ή αν έχει ήδη συμβεί.
- Τον αγγελιοφόρο (notifier), ο οποίος λαμβάνει την πληροφορία από τον διευθυντή με σκοπό να αποφασίσει πότε και πως να ειδοποιήσει την αναγκαία οντότητα. Επίσης, ο αγγελιοφόρος είναι σε επικοινωνία και με τους αντιπροσώπους ώστε να ρυθμιστούν θέματα εισαγωγής στοιχείων αν αυτό χρειαστεί.



Εικόνα 7

Τέλος, αυτά τα μοντέλα μπορούν να λειτουργούν με δύο διαφορετικούς τρόπους. Μπορούν να είναι προσαρμοστικά (adaptive) δηλαδή να αλλάζουν την συμπεριφορά τους ανάλογα με τις καταστάσεις και τις ενέργειες των συστημάτων. Εναλλακτικά, μπορούν να είναι στατικά (static) δηλαδή, έχουν αρχικοποιηθεί από δεδομένα που έχουν συλλεγεί και δεν τροποποιούνται κατά την διάρκεια εκτέλεσης του συστήματος.[8].[9]

5. Data poisoning

Μία από τις μεγαλύτερες απειλές για την τεχνητή νοημοσύνη είναι η “δηλητηρίαση” των δεδομένων (data poisoning). Αυτός ο τύπος επίθεσης στοχεύει στην τροποποίηση των δεδομένων εκπαίδευσης (training dataset) με σκοπό την υποβάθμιση της απόδοσης του μοντέλου. Υπάρχουν πολλοί τρόποι να πετύχει κανείς μία τέτοιου είδους επίθεση, οι οποίοι μπορούν να διαχωριστούν σε επιθέσεις “white-box” και “black-box” με βάση το περιβάλλον της επίθεσης. Ένας επιτιθέμενος που διαθέτει πλήρη γνώση του μοντέλου ανήκει στην κατηγορία των επιθέσεων “white-box”. Από την άλλη, όταν ο επιτιθέμενος δεν έχει καμία γνώση του μοντέλου και αλληλεπιδρά μόνο μέσω των εισόδων και εξόδων του, ανήκει στην κατηγορία των επιθέσεων “black-box”.[10]

5.1. Backdoor Attack

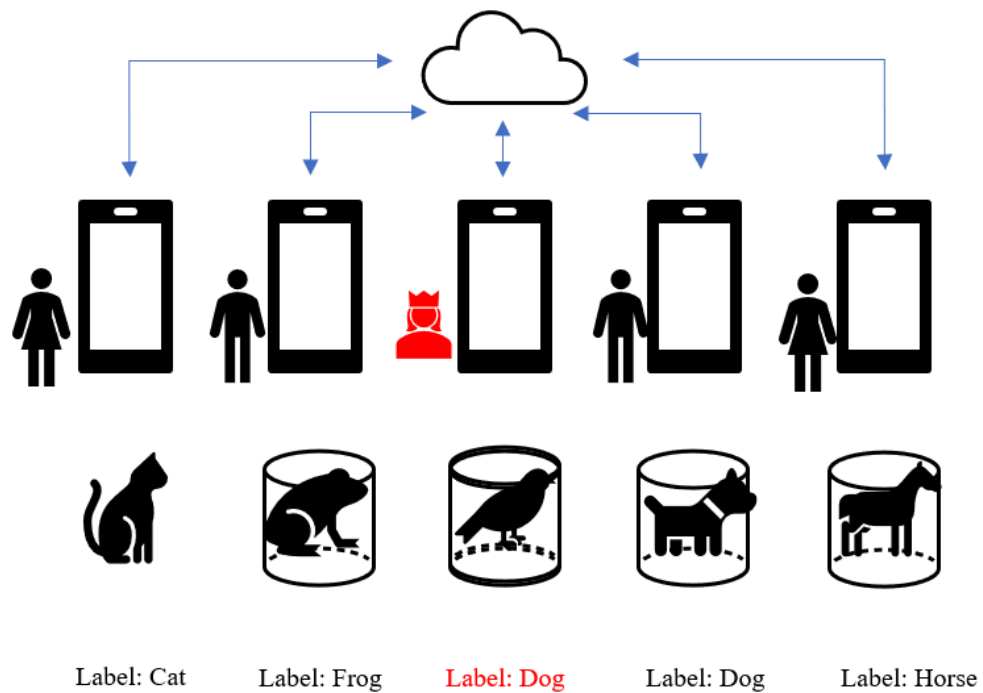
Η επίθεση αυτή, στο πλαίσιο της “δηλητηρίασης” δεδομένων (data poisoning), αναφέρεται σε μία τεχνική όπου ο επιτιθέμενος εισάγει ένα κρυφό μοτίβο ή « trigger » στα δεδομένα εκπαίδευσης (training dataset), το οποίο του επιτρέπει να χειραγωγήσει τη συμπεριφορά του μοντέλου κατά τη στιγμή της εξαγωγής συμπερασμάτων (εικ. 7). Κατά τη διάρκεια της εκπαίδευσης, ο επιτιθέμενος “δηλητηριάζει” ένα μικρό υποσύνολο των δεδομένων ενσωματώνοντας αυτό το trigger (το οποίο θα μπορούσε να είναι ένα συγκεκριμένο μοτίβο από pixel σε εικόνες ή μια ακολουθία σε κείμενο) και το συνδέει με μια κλάση-στόχο. Το μοντέλο μαθαίνει να ταξινομεί κάθε είσοδο που περιέχει αυτό το trigger ως ότι ανήκει σε αυτή τη συγκεκριμένη κλάση, ενώ εκτελείται κανονικά σε μη τροποποιημένα δεδομένα. Το βασικό χαρακτηριστικό μιας επίθεσης backdoor είναι ότι το μοντέλο συμπεριφέρεται σωστά σε κανονικές, καθαρές εισόδους, οπότε η επίθεση δεν είναι εύκολα ανιχνεύσιμη. Ωστόσο, όταν ο επιτιθέμενος εισάγει ένα δείγμα με το trigger κατά τη διάρκεια της εξαγωγής συμπερασμάτων, το μοντέλο εξάγει την επιθυμητή για τον ίδιο πρόβλεψη. Αυτός ο τύπος επίθεσης είναι ιδιαίτερα ύπουλος επειδή μπορεί να παρακάμψει τις τυπικές διαδικασίες δοκιμής (test) και επικύρωσης (validation), καθώς η “πίσω πόρτα” παραμένει κρυφή εκτός αν ενεργοποιηθεί από το συγκεκριμένο trigger.[11]

$$\begin{array}{ccc}
 \begin{array}{c} x \\ \text{"panda"} \\ 57.7\% \text{ confidence} \end{array} & + .007 \times \begin{array}{c} \text{sign}(\nabla_x J(\theta, x, y)) \\ \text{"nematode"} \\ 8.2\% \text{ confidence} \end{array} & = \begin{array}{c} x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \\ \text{"gibbon"} \\ 99.3\% \text{ confidence} \end{array}
 \end{array}$$

Εικόνα 8

5.2. Label Flipping Attack

Η επίθεση αντιστροφής ετικετών (label flipping) είναι μια μορφή επίθεσης “δηλητηρίασης” δεδομένων (data poisoning) όπου ο επιτιθέμενος μεταβάλλει κακόβουλα τις ετικέτες ορισμένων δειγμάτων, αφήνοντας τα χαρακτηριστικά εισόδου τους αμετάβλητα (εικ. 8). Αυτή η τροποποίηση οδηγεί σε εσφαλμένη ταξινόμηση από το εκπαιδευμένο μοντέλο. Συγκεκριμένα, οι ετικέτες μιας κλάσης αντιστρέφονται σε μια άλλη, με αποτέλεσμα το μοντέλο να μαθαίνει εσφαλμένες αντιστοιχίσεις μεταξύ των εισόδων και των εξόδων. Η επίθεση αυτή μπορεί να υποβαθμίσει σημαντικά την απόδοση του μοντέλου, ιδίως σε εργασίες όπως η ταξινόμηση. Για παράδειγμα, σε ένα πρόβλημα δυαδικής ταξινόμησης, ένας επιτιθέμενος μπορεί να αντιστρέψει θετικά παραδείγματα σε αρνητικά (και αντίστροφα), προκαλώντας σύγχυση στο μοντέλο κατά τη διάρκεια της εκπαίδευσης. Η επίθεση αυτή μπορεί να εφαρμοστεί είτε τυχαία είτε στοχευμένα, ανάλογα με τους στόχους του επιτιθέμενου. Σε ορισμένες περιπτώσεις, η αντιστροφή ετικετών μπορεί να στοχεύει σε συγκεκριμένα παραδείγματα ή να εστιάζει σε συγκεκριμένες περιοχές του χώρου εισόδου για να μεγιστοποιήσει τον αντίκτυπο στην ακρίβεια του μοντέλου.[12],[13]



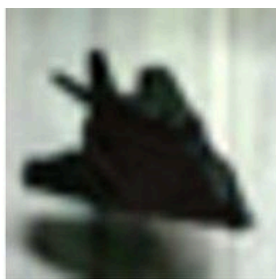
Εικόνα 9

5.3. Targeted Poisoning Attack

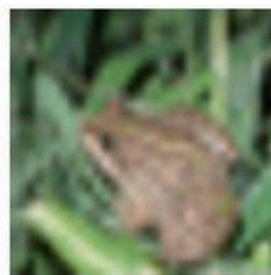
Ο επιτιθέμενος προσθέτει δεδομένα που στοχεύουν συγκεκριμένα σημεία του χώρου χαρακτηριστικών του μοντέλου (εικ. 9). Ο στόχος είναι να χειραγωγηθεί η πρόβλεψη του μοντέλου για συγκεκριμένες εισόδους, χωρίς να αλλοιωθεί η γενική απόδοση του μοντέλου. Σε μια στοχευμένη επίθεση “δηλητηρίασης”, ο επιτιθέμενος προκαλεί μια λανθασμένη πρόβλεψη από τον ταξινομητή δημιουργώντας ψευδή κακόβουλα δεδομένα για να διαστρεβλώσει την πραγματική κλάση ορισμένων αντικειμένων (που ονομάζονται αντικείμενα-στόχοι) σε μια συγκεκριμένη απάντηση-στόχο, η οποία είναι μια πιο διακριτική επίθεση. Η απάντηση-στόχος είναι συνήθως προκαθορισμένη από τον επιτιθέμενο, και όταν εκτελείται η στοχευμένη επίθεση, εάν το τελικό αποτέλεσμα ταξινόμησης του αντικειμένου-στόχου γίνει η απάντηση-στόχος μετά την επίθεση, η επίθεση σε αυτό το αντικείμενο είναι επιτυχής. Διαφορετικά, η επίθεση σε αυτό το αντικείμενο αποτυγχάνει.[10]



Base instance
(airplane)



Poison instance
(airplane)



Targeted instance
(frog)

Εικόνα 10

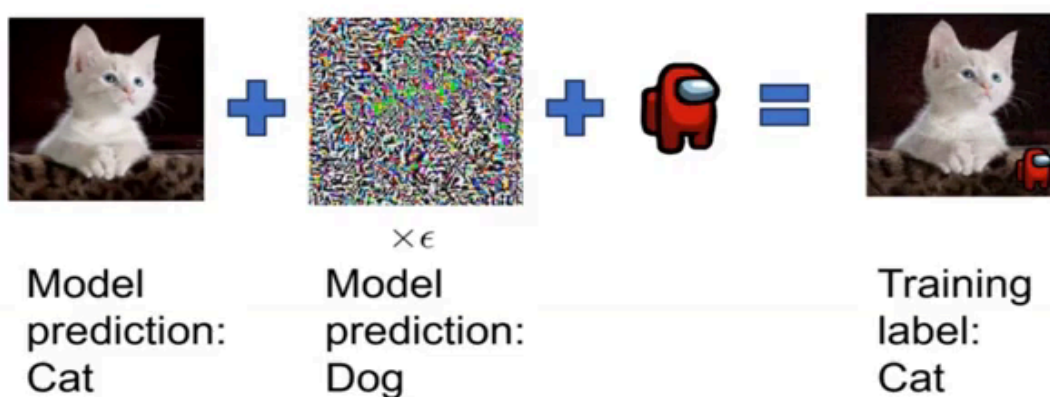
5.4. Availability Attack

Οι επιθέσεις “δηλητηρίασης” της διαθεσιμότητας (availability attack), γνωστές και ως επιθέσεις δηλητηρίασης δεδομένων χωρίς διάκριση, αποσκοπούν στην υπονόμευση της συνολικής απόδοσης του μοντέλου με κακόβουλη χειραγώγηση των δεδομένων εκπαίδευσης. Μια επιτυχημένη επίθεση διαθεσιμότητας μειώνει σημαντικά την ακρίβεια του μοντέλου, καθιστώντας το άχρηστο σε ρεαλιστικά σενάρια. Οι επιθέσεις αυτές εστιάζουν στη μείωση της λειτουργικότητας του συστήματος, προκαλώντας υπερβολικά ψευδή θετικά ή αρνητικά αποτελέσματα, καθυστερήσεις στη διαδικασία ή ακόμα και πλήρη κατάρρευση του συστήματος. Αυτές οι επιθέσεις είναι ιδιαίτερα επικίνδυνες σε εφαρμογές κρίσιμης σημασίας, όπως η υγειονομική περίθαλψη, τα αυτόνομα οχήματα και τα χρηματοοικονομικά συστήματα.

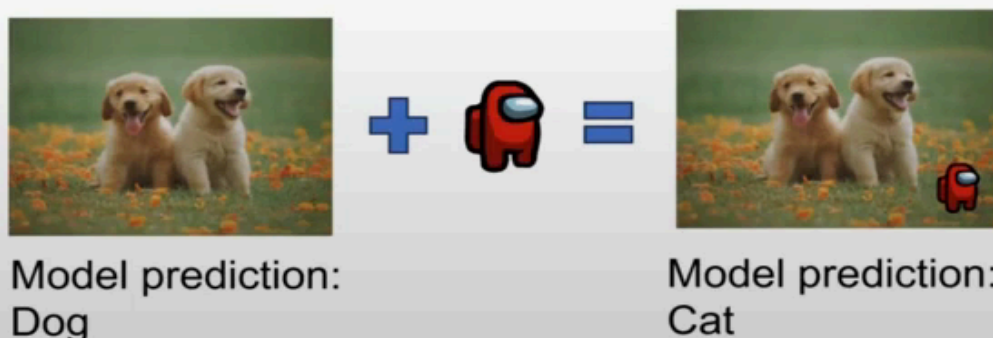
5.5. Clean-label Poisoning Attack

Αυτή είναι μια πιο εκλεπτυσμένη μορφή επίθεσης όπου ο επιτιθέμενος εισάγει κακόβουλα δείγματα χωρίς να αλλάζει τις ετικέτες τους (εικ. 10). Αυτό δυσκολεύει την ανίχνευση της επίθεσης, καθώς τα “δηλητηριασμένα” δεδομένα μοιάζουν έγκυρα, αλλά επιφέρουν καταστροφικά αποτελέσματα κατά την πρόβλεψη. Τα δηλητηριασμένα δεδομένα φαίνονται να έχουν “καθαρές” και σωστές ετικέτες, γι’ αυτό και ονομάζεται “clean label attack” όμως, στην πραγματικότητα τα δείγματα αυτά έχουν αλλοιωθεί με πολύ συγκεκριμένο τρόπο με αποτέλεσμα να προκληθούν σφάλματα στο μοντέλο εκμάθησης, ειδικά σε ορισμένες συγκεκριμένες εισόδους, χωρίς όμως να εντοπίζεται εύκολα η δηλητηρίαση, καθώς οι ετικέτες παραμένουν ορθές.[14],[15],[16]

Training time:



Test time:



Εικόνα 11

Το κοινό χαρακτηριστικό που έχουν όλες αυτές οι επιθέσεις είναι ότι εστιάζουν στο σύνολο των δεδομένων και όχι στο ίδιο το μοντέλο. Αυτό έχει ως αποτέλεσμα να μην υπάρχει μεγάλη επιρροή αν το μοντέλο είναι black-box ή white-box. Επίσης, πολλά από αυτά τα μοντέλα, όπως και στην έρευνα που θα αναφερθεί στην ενότητα 7, χρησιμοποιούν σύνολα δεδομένων που υπάρχουν διαθέσιμα δωρεάν στο διαδίκτυο και που ο καθένας μπορεί να κατεβάσει και να τα επεξεργαστεί. Όμως, υπάρχουν και αρκετοί επιτιθέμενοι που δημοσιεύουν “δηλητηριασμένα” σύνολα δεδομένων για να επιτύχουν τους κακόβουλους σκοπούς τους.

6. Ανασκόπηση των επιθέσεων τύπου backdoor και των αντίστοιχων μέτρων αντιμετώπισής τους στα συστήματα βαθιάς μάθησης

Το άρθρο [17] είναι μια συνολική ανασκόπηση των επιθέσεων τύπου backdoor και των αντίστοιχων μέτρων αντιμετώπισής τους στα συστήματα βαθιάς μάθησης (deep learning). Οι επιθέσεις αυτές είναι ιδιαίτερα επικίνδυνες, δεδομένου ότι μπορεί να στοχεύουν σημαντικές εφαρμογές, όπως αναγνώριση προσώπου ή αυτόνομη οδήγηση, όπου οι συνέπειες μπορεί να είναι καταστροφικές. Η εργασία επισημαίνει ότι, παρόλο που έχουν γίνει πολλές προσπάθειες για την ανάπτυξη μέτρων αντιμετώπισης αυτών των επιθέσεων, οι άμυνες εξακολουθούν να υστερούν σε σχέση με τις επιθέσεις. Επιπλέον, αναφέρεται ότι δεν υπάρχει ακόμα ένα ενιαίο μέτρο άμυνας που να μπορεί να προστατεύσει από όλες τις μορφές των επιθέσεων τύπου backdoor. Το συγκεκριμένο paper παρέχει μια ταξινόμηση των επιθέσεων backdoor με βάση τις δυνατότητες του επιτιθέμενου και το στάδιο του κύκλου ζωής της μηχανικής μάθησης που επηρεάζεται (π.χ., συλλογή δεδομένων, συνεργατική μάθηση, μεταπαραγωγή) ή όπως αναφέρονται στην έρευνα "επιφάνειες επίθεσης" (attack surfaces) (εικ. 11).

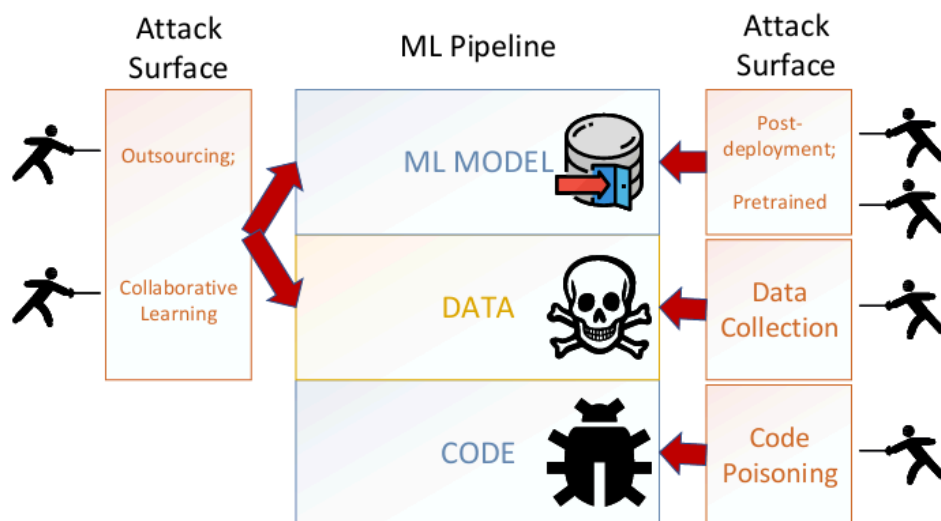


Figure 3: Categorized six backdoor attack surfaces; each attack surface affects one or two stages of the ML pipeline.

Εικόνα 12

Πρόκειται για έξι διαφορετικές κατηγορίες που είναι οι εξής:

- **Poisoning του κώδικα:** Ο εισβολέας τροποποιεί τον κώδικα του λογισμικού που χρησιμοποιείται για την εκπαίδευση του μοντέλου. Πολλές φορές, γίνεται χρήση δημόσιων frameworks για την ενίσχυση της ανάπτυξης ενός μοντέλου. Ένα δημόσιο framework βαθιάς μάθησης θα μπορούσε να οδηγήσει σε ευπάθειες για τα μοντέλα μηχανικής μάθησης που είναι χτισμένα πάνω σε αυτό. Ένας επιτιθέμενος μπορεί να εκμεταλλευτεί τις ευπάθειες για να εξαπολύσει διάφορες επιθέσεις που κυμαίνονται από επιθέσεις άρνησης παροχής υπηρεσιών (DoS attack) έως επιθέσεις hijacking ροής ελέγχου που είτε θέτουν σε κίνδυνο το σύστημα είτε αποφεύγουν την ανίχνευση.
- **Εκπαίδευση μέσω outsourcing:** Ο χρήστης αναθέτει την εκπαίδευση του μοντέλου σε τρίτους λόγω έλλειψης δεξιοτήτων ή υπολογιστικών πόρων. Σε αυτό το σενάριο, ο χρήστης ορίζει την αρχιτεκτονική του μοντέλου, παρέχει το σύνολο των δεδομένων και αναθέτει την εκπαίδευση σε υπηρεσία μηχανικής μάθησης ως υπηρεσία (MLaaS). Ως εκ τούτου, ένας κακόβουλος πάροχος MLaaS ελέγχει τη φάση της εκπαίδευσης και εισχωρεί στο μοντέλο ML κατά τη διάρκεια της διαδικασίας εκπαίδευσης.
- **Προεκπαιδευμένα μοντέλα (pretrained models):** Οι χρήστες μπορεί να κατεβάζουν έτοιμα μοντέλα για να τα χρησιμοποιήσουν ή να τα βελτιώσουν μέσω μεταφοράς μάθησης (transfer learning). Αυτά τα προεκπαιδευμένα μοντέλα μπορεί να περιέχουν κρυφά “triggers”. Για παράδειγμα, ο επιτιθέμενος μπορεί να δημοσιεύσει και να διαφημίσει στο διαδίκτυο έναν (μολυσμένο) εξαγωγέα χαρακτηριστικών (feature extractor) για ταξινόμηση εικόνας. Ή μπορεί να κατεβάζει πρώτα ένα δημοφιλές “καθαρό” προεκπαιδευμένο μοντέλο, το χειραγωγεί και το επανεκπαιδεύει με παραποιημένα δεδομένα (δηλαδή, εμφυτεύει ένα backdoor στο μοντέλο). Στη συνέχεια, ο επιτιθέμενος αναδημοσιεύει το μοντέλο.
- **Συλλογή δεδομένων (data collection):** Κατά τη συλλογή δεδομένων πολλές φορές γίνεται χρήση από διάφορες πηγές. Ο επιτιθέμενος μπορεί να έχει δημοσιεύσει δηλητηριασμένα (poisoned) δεδομένα, τα οποία οδηγούν σε

ανεπιθύμητες συμπεριφορές του μοντέλου. Συγκεκριμένα, οι επιθέσεις “δηλητηρίασης” clean-label και image-scaling attacks ανήκουν σε μια τέτοια “επιφάνεια επίθεσης”. Τέτοιες επιθέσεις διατηρούν τη συνέπεια μεταξύ ετικετών και τιμής δεδομένων, παρακάμπτοντας έτσι τις χειροκίνητες ή οπτικές επιθεωρήσεις.

- Συνεργατική μάθηση (collaborative learning): Συστήματα, όπως το federated learning, επιτρέπουν σε διάφορους πελάτες να συνεισφέρουν στην εκπαίδευση του ίδιου μοντέλου χωρίς να μοιράζονται τα δεδομένα τους. Ένας κακόβουλος πελάτης μπορεί να προσθέσει backdoor στο κοινό μοντέλο. Για παράδειγμα, η Google εκπαιδεύει μοντέλα πρόβλεψης λέξεων από δεδομένα που εντοπίζονται στα τηλέφωνα των χρηστών. Η συνεργατική μάθηση ή η κατανεμημένη μάθηση έχει σχεδιαστεί για να προστατεύει τη διαρροή της ιδιωτικότητας των δεδομένων που ανήκουν στους πελάτες ή τους συμμετέχοντες. Κατά τη διάρκεια της φάσης μάθησης, ο server δεν έχει πρόσβαση στα δεδομένα εκπαίδευσης των συμμετεχόντων. Αυτό καθιστά τη συνεργατική μάθηση ευάλωτη σε διάφορες επιθέσεις, συμπεριλαμβανομένης της επίθεσης backdoor. Ένα μοντέλο που μαθαίνεται από κοινού μπορεί εύκολα να παραβιαστεί με backdoor όταν ένας πολύ μικρός αριθμός συμμετεχόντων είναι εκτεθειμένος ή ελέγχεται από έναν επιτιθέμενο. Τόσο η δηλητηρίαση τοπικών δεδομένων όσο και η δηλητηρίαση του μοντέλου μπορούν να πραγματοποιηθούν από τον επιτιθέμενο για να εμφυτεύσει backdoor στο κοινό μοντέλο.
- Μετά την ανάπτυξη (post-deployment): Μετά την εγκατάσταση του μοντέλου, μπορούν να πραγματοποιηθούν επιθέσεις, όπως αλλοιώσεις στις παραμέτρους του μοντέλου μέσω φυσικών ή άλλων επιθέσεων. Σκεφτείτε ένα τυπικό σενάριο επίθεσης όπου ένας επιτιθέμενος και ένας χρήστης μοιράζονται τον ίδιο server. Ο χρήστης εκτελεί ένα μοντέλο μηχανικής μάθησης και φορτώνει τα βάρη του στη μνήμη. Ο επιτιθέμενος αντιστρέφει έμμεσα τα bits των βαρών ενεργοποιώντας το rowhammer bug, με αποτέλεσμα την υποβάθμιση της ακρίβειας.

Στην συνέχεια αναλύονται οι διάφορες μορφές επιθέσεων backdoor:

- Class-specific και Class-agnostic: Κάποιες επιθέσεις μπορεί να στοχεύουν μόνο σε συγκεκριμένες κλάσεις δεδομένων, ενώ άλλες μπορεί να

στοχεύουν ανεξάρτητα από την κατηγορία των δεδομένων. Ένα backdoored μοντέλο μπορεί να ταξινομήσει εσφαλμένα την είσοδο από οποιαδήποτε κλάση που φέρει το trigger στη κλάση-στόχο. Δηλαδή, η παρουσία του trigger καθορίζει την επίθεση. Στην άλλη περίπτωση, το μοντέλο μπορεί να ταξινομήσει εσφαλμένα την είσοδο από συγκεκριμένες κλάσεις που έχουν στιγματιστεί με το trigger στη στοχευόμενη κλάση. Δηλαδή, η παρουσία του trigger, μαζί με τη συγκεκριμένη κλάση καθορίζει την επίθεση.

- Πολλαπλά triggers: Μια επίθεση μπορεί να περιλαμβάνει πολλά διαφορετικά trigger που οδηγούν σε ίδιο ή διαφορετικό κακόβουλο αποτέλεσμα. Υπάρχουν δύο ειδών κατηγορίες εδώ. Πολλαπλά triggers στην ίδια ετικέτα όπου σε αυτήν την περίπτωση, η παρουσία οποιουδήποτε trigger μεταξύ πολλαπλών triggers μπορεί να οδηγήσει το backdoored μοντέλο στην ίδια στοχευόμενη ετικέτα. Πολλαπλά triggers σε πολλαπλές ετικέτες. Το backdoored μοντέλο μπορεί να καταληφθεί από πολλαπλά triggers ωστόσο, κάθε trigger στοχεύει σε διαφορετική ετικέτα.
- Δυναμικά triggers: Οι επιθέσεις μπορεί να περιλαμβάνουν δυναμικά triggers που αλλάζουν εμφάνιση και τοποθεσία, καθιστώντας δυσκολότερη την ανίχνευση. Όλες αυτές οι ρυθμίσεις του trigger μπορούν να είναι αυθαίρετες για τον επιτιθέμενο. Για παράδειγμα, το trigger μπορεί να είναι οποιουδήποτε μεγέθους και σε οποιαδήποτε θέση. Ειδικότερα, για τους τομείς ήχου και κειμένου, το σχήμα του trigger ενδέχεται να μην είναι εφαρμόσιμο. Στον τομέα της οπτικής αναγνώρισης, αυτό συμβαίνει όταν το trigger αναμειγνύεται με την εικόνα εισόδου. Αυτό μπορεί επίσης να αναφέρεται ως αορατότητα του trigger. Στον τομέα του ήχου, αυτό σχετίζεται με το πλάτος κύματος του trigger του ήχου. Στον τομέα του κειμένου, αυτό σημαίνει ότι η σημασιολογία έχει παραμείνει.

Έπειτα, ταξινομεί και τα μέτρα αντιμετώπισης σε τέσσερις κατηγορίες (εικ. 12):

- **Τυφλή αφαίρεση backdoor (blind backdoor removal):** Αφαίρεση του backdoor χωρίς να γνωρίζουμε αν το μοντέλο είναι όντως μολυσμένο ή όχι ή αν μία είσοδος έχει triggers η όχι. Ο κύριος σκοπός είναι η απομάκρυνση ή η καταστολή του backdoor, διατηρώντας παράλληλα το CDA (clean data accuracy) των καθαρών εισόδων, ιδίως αν ληφθεί υπόψη ότι πρόκειται τελικά για ένα “καθαρό” μοντέλο.
- **Offline επιθεώρηση:** Έλεγχος των δεδομένων ή του μοντέλου σε μια εκ των υστέρων βάση για τον εντοπισμό των backdoors. Ωστόσο, η επιθεώρηση “δηλητηριασμένων” δεδομένων είναι εφικτή μόνο στο πλαίσιο της συλλογής δεδομένων. Είναι απίθανο για άλλες “επιφάνειες επίθεσης” όπου ο επιτιθέμενος δεν θα εκθέσει τα δηλητηριασμένα δεδομένα στον αμυνόμενο. Η επιθεώρηση δεδομένων προϋποθέτει ότι τα “δηλητηριασμένα” δεδομένα είναι διαθέσιμα για τους αμυνόμενους, δεδομένου ότι ο επιτιθέμενος εκμεταλλεύεται την “δηλητηρίαση” δεδομένων για να εισάγει backdoors. Σε αντίθεση με την επιθεώρηση δεδομένων που προϋποθέτει πρόσβαση σε “δηλητηριασμένα” δεδομένα, η επιθεώρηση του μοντέλου χρησιμοποιεί μια πιο ρεαλιστική υπόθεση ότι τα “δηλητηριασμένα” δεδομένα δεν είναι διαθέσιμα. Ως εκ τούτου, είναι κατάλληλα για την αντιμετώπιση επιθέσεων backdoor που προκύπτουν από διάφορες επιφάνειες επίθεσης και δεν περιορίζονται μόνο στη συλλογή δεδομένων.
- **Online επιθεώρηση:** Συνεχής παρακολούθηση των εισερχόμενων δεδομένων ή του μοντέλου κατά τη διάρκεια της χρήσης. Η επιθεώρηση δεδομένων ελέγχει τις εισόδους για να αποφασίσει αν περιέχεται ή όχι ένα trigger μέσω της ανίχνευσης ανωμαλιών, αρνούμενη έτσι την αντίπαλη είσοδο και στη συνέχεια ενεργοποιώντας τον συναγερμό αν έχει οριστεί. Η επιθεώρηση μοντέλων βασίζεται σε μια τεχνική ανωμαλιών για να διακρίνει τις μη φυσιολογικές συμπεριφορές ενός μοντέλου που προέρχονται από ένα backdoor.
- **Αφαίρεση backdoor μετά τον εντοπισμό (post backdoor removal):** Αφαίρεση του backdoor αφού πρώτα εντοπιστεί το πρόβλημα. Μόλις

εντοπίζεται το backdoor, είτε μέσω της επιθεώρησης μοντέλου είτε μέσω της επιθεώρησης δεδομένων, μπορεί να ληφθεί υπόψη η αφαίρεση του. Ένας τρόπος είναι η αφαίρεση των αλλοιωμένων εισόδων και η εκ νέου εκπαίδευση του μοντέλου, η οποία φαίνεται να είναι πρακτική μόνο στο πλαίσιο της συλλογής δεδομένων, καθώς ο χρήστης δεν επιτρέπεται να έχει πρόσβαση στις εισόδους trigger στο πλαίσιο άλλων "επιφανειών επίθεσης". Ο άλλος τρόπος είναι η αφαίρεση των συμπεριφορών backdoor από ένα μοντέλο με επανεκπαίδευση ή λεπτομερή ρύθμιση του μοντέλου με backdoor χρησιμοποιώντας αλλοιωμένα δεδομένα που περιλαμβάνουν το trigger αλλά χαρακτηρίζονται σωστά, τα οποία μαθαίνουν εκ νέου τα διορθωμένα όρια απόφασης

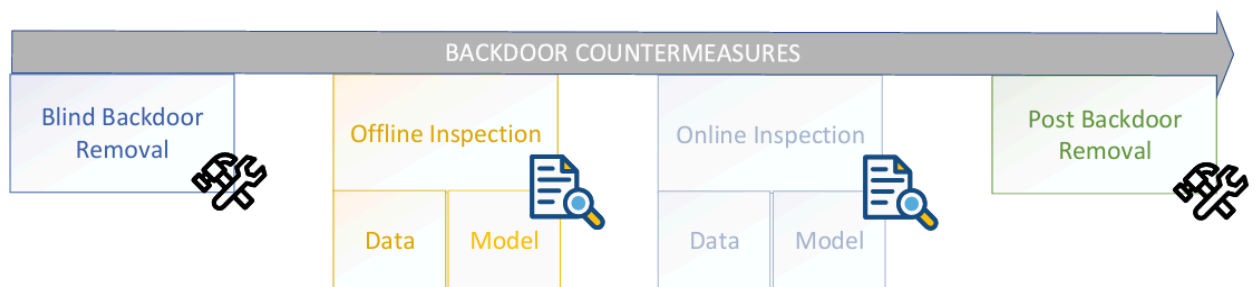


Figure 9: Categories of backdoor countermeasures: blind backdoor removal without first check whether the model is backdoored or not; offline inspection on the data or model; online inspection on the (upcoming) data or model; post backdoor removal or restoration after backdoor trigger is identified.

Εικόνα 13

Σαν συμπέρασμα, το άρθρο επισημαίνει την ανάγκη για περαιτέρω έρευνα στον τομέα της ασφάλειας των συστημάτων βαθιάς μάθησης και τονίζει τις προκλήσεις που υπάρχουν στην ανάπτυξη πιο αποτελεσματικών και πρακτικών μεθόδων άμυνας.

7. Σκοπός της έρευνας

Σε αυτό το κεφάλαιο, θα αναλυθεί η έρευνα που πραγματοποιήθηκε στα πλαίσια της εργασίας. Η μεθοδολογία της διεξαγωγής της, τα αποτελέσματα καθώς και οι δυσκολίες που υπήρξαν και οι τρόποι αντιμετώπισής τους.

Για την παρούσα έρευνα αναπτύχθηκε ένα Intrusion Detection System (IDS) με την χρήση της τεχνητής νοημοσύνης και πιο συγκεκριμένα, την χρήση νευρωνικού δικτύου, για την ανίχνευση επιθέσεων στην δικτυακή κίνηση του Internet of Things. Το μοντέλο παρουσίασε ποσοστό ακρίβειας (Accuracy) $\approx 99\%$. Σκοπός της έρευνας υπήρξε εύρεση ευπαθειών στο παρόν μοντέλο ώστε να κλονιστεί η απόδοση του και να κριθεί αναξιόπιστο. Δηλαδή, να μην αναγνωρίσει μία κακόβουλη ενέργεια ή να ανιχνεύσει λανθασμένα απλή κίνηση ως κακόβουλη. Υπάρχουν πολλές επιθέσεις κατά της τεχνητής νοημοσύνης. Μια συγκεκριμένη κατηγορία ασχολείται με την μεταβολή των δεδομένων εισόδου και ακούει στο όνομα data poisoning attack.

Για την συγκεκριμένη εργασία επιλέχθηκε η τεχνική του label-flipping (LF). Η τεχνική αυτή αποτελεί μια επίθεση “δηλητηρίασης” των δεδομένων (data poisoning), όπου ο επιτιθέμενος μεταβάλλει ένα μέρος του συνόλου των δεδομένων εκπαίδευσής (training dataset) του εκάστοτε μοντέλου, αλλάζοντας τις ετικέτες ορισμένων εγγραφών από μια κλάση σε μια άλλη. Δυστυχώς, αυτή η επίθεση είναι εύκολο να πραγματοποιηθεί και δύσκολο να εντοπιστεί και επηρεάζει αρνητικά την απόδοση του συνολικού μοντέλου.[13]

7.1. Το σύνολο των δεδομένων

Το dataset που επιλέχθηκε για αυτήν την εργασία προέρχεται από τον ιστότοπο του καναδικού ινστιτούτου κυβερνοασφάλειας του πανεπιστημίου του New Brunswick.[19] Για να παραχθεί το dataset πραγματοποιήθηκαν 33 επιθέσεις σε ένα δίκτυο συσκευών IoT αποτελούμενο από 105 συσκευές. Οι επιθέσεις αυτές ταξινομούνται σε επτά κατηγορίες. Αυτές είναι DDoS, DoS, Recon, Web-based, Brute Force, Spoofing και Mirai. Η κάθε μία από αυτές τις κατηγορίες έχει ένα σύνολο επιθέσεων που πραγματοποιήθηκαν. Πιο συγκεκριμένα:

- Για την κατηγορία DDoS οι επιθέσεις είναι: ACK fragmentation, UDP flood, SlowLoris, ICMP flood, RSTFIN flood, PSHACK flood, HTTP flood, UDP fragmentation, TCP flood, SYN flood, SynonymousIP flood
- Για την κατηγορία DoS: TCP flood, HTTP flood, SYN flood, UDP flood
- Για την κατηγορία Recon: Ping sweep, OS scan, Vulnerability scan, Port scan, Host discovery
- Για την κατηγορία Web-based: Sql injection, Command injection, Backdoor malware, Uploading attack, XSS, Browser hijacking
- Για την κατηγορία Brute Force: Dictionary brute force
- Για την κατηγορία Spoofing: Arp spoofing, DNS spoofing
- Για την κατηγορία Mirai: GREIP flood, Greeth flood, UDPPlain

Το dataset είναι χωρισμένο σε 169 CSV αρχεία το καθένα με διαφορετικό αριθμό εγγραφών που κυμαίνεται περίπου από 218.000 έως 450.000 εγγραφές, εκ των οποίων η απλή κίνηση (benign traffic) αποτελεί περίπου το 2%. Τα χαρακτηριστικά κάθε δείγματος αριθμούνται σε 46 και είναι τα εξής: flow_duration, Header_Length, Protocol Type, Duration, Rate, Srate, Drate, fin_flag_number, syn_flag_number, rst_flag_number, psh_flag_number, ack_flag_number, ece_flag_number, cwr_flag_number, ack_count, syn_count, fin_count, urg_count, rst_count, HTTP, HTTPS, DNS, Telnet, SMTP, SSH, IRC, TCP, UDP, DHCP, ARP, ICMP, IPv, LLC, Tot sum, Min, Max, AVG, Std, Tot size, IAT, Number, Magnitue, Radius, Covariance, Variance, Weight. Τέλος, υπάρχει και η ετικέτα (label) της κάθε εγγραφής όπου αναφέρεται αν έχουμε απλή κίνηση (benign traffic) ή το είδος της επίθεσης (πχ. DoS-SYN_Flood, MITM-Arp Spoofing, Mirai-greeth_flood κ.α.)

7.2. Υλοποίηση του Intrusion Detection System & επίθεση

Παρακάτω φαίνεται ο κώδικας σε python για το Intrusion Detection System (IDS) με την χρήση τεχνητής νοημοσύνης, καθώς και η επίθεση που πραγματοποιήθηκε. Πρόκειται για μία παραλλαγή ενός απλού νευρωνικού δικτύου, το Multi-Layer perceptron (MLP) χρησιμοποιώντας τη βιβλιοθήκη Keras.

```
# -*- coding: utf-8 -*-
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from keras.models import Sequential
from keras.layers import Dense, Input

dataset= pd.read_csv("C:\\CICIoT2023\\part-00001-363d1ba3-8ab5-4f96-bc25-4d5862db7cb9-c000.csv",sep=",", decimal=".")

x= dataset.drop(columns=["label"], axis=1).values.astype('float32')
y= dataset["label"].apply(lambda x: 1 if x == 'BenignTraffic' else 0).values.astype('float32')

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Εικόνα 14

```
# Define the model
model = Sequential([
    Input(shape=(X_train.shape[1],)),
    Dense(128, activation='relu'),
    Dense(64, activation='relu'),
    Dense(32, activation='relu'),
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy','precision','recall'])

# Train the model
model.fit(X_train, y_train, epochs=10, batch_size=32, validation_data=(X_test, y_test))

# Evaluate the model
test_loss, test_acc, test_prec, test_rec = model.evaluate(X_test, y_test)
print('Test accuracy:', test_acc, 'Test recall:', test_rec, 'Test precision:', test_prec, 'Test loss:', test_loss)
```

Εικόνα 15

Αρχικά, εισάγονται διάφορες βιβλιοθήκες, όπως pandas για χειρισμό δεδομένων, StandardScaler και train_test_split για preprocessing από την sklearn, καθώς και οι Sequential, Dense και Input από την keras για τον ορισμό του μοντέλου νευρωνικού δικτύου.

Εν συνεχεία, γίνεται διαχωρισμός των χαρακτηριστικών (X) και ετικετών (Y) ορίζοντας για τις τιμές του Y “1” για κανονική κίνηση (BenignTraffic) και “0” για οτιδήποτε άλλο (δηλαδή επιθέσεις). Το σύνολο δεδομένων διαχωρίζεται σε σύνολα εκπαίδευσης (train) και δοκιμής (test), χρησιμοποιώντας αναλογία 80-20. Έπειτα, εφαρμόζεται ο standardScaler για την κανονικοποίηση των χαρακτηριστικών (εικ. 13). Αυτό βοηθά στη βελτίωση της σύγκλισης του μοντέλου κατά την εκπαίδευση.

Ορίζεται ένα απλό νευρωνικό δίκτυο (εικ. 14) χρησιμοποιώντας Sequential με:

- Ένα επίπεδο εισόδου με το ίδιο σχήμα όπως ο αριθμός των χαρακτηριστικών.
- Τρία κρυφά επίπεδα με 128, 64, και 32 νευρώνες αντίστοιχα, καθένα με τη συνάρτηση ενεργοποίησης ReLU.
- Ένα επίπεδο εξόδου με έναν νευρώνα και τη συνάρτηση ενεργοποίησης sigmoid (για δυαδική ταξινόμηση).

Τέλος, το μοντέλο μεταγλωττίζεται χρησιμοποιώντας τον βελτιστοποιητή Adam, την απώλεια δυαδικής διασταυρούμενης εντροπίας (binary cross-entropy) που είναι κατάλληλη για δυαδική ταξινόμηση και δείκτες απόδοσης όπως η ακρίβεια (accuracy), ακρίβεια θετικής πρόβλεψης (precision) και η ευαισθησία (recall/sensitivity).

Στην εικόνα 15 φαίνεται η απόδοση του μοντέλου σε ένα μέρος του συνόλου δεδομένων με 275,258 εγγραφές.

```
Test accuracy: 0.9939672350883484 Test recall: 0.9273256063461304 Test
precision: 0.8350785374641418 Test loss: 0.0139619754627347
```

Εικόνα 16

Παρακάτω, παρατίθεται ο παραλλαγμένος κώδικας έχοντας εφαρμόσει την επίθεση label-flipping (εικ. 16). Το νέο αυτό κομμάτι κώδικα που προστέθηκε επιλέγει τυχαία και ανακατεύει μεταξύ τους τις ετικέτες ενός υποσυνόλου του συνόλου δεδομένων (n=60.000 σειρές). Αλλάζει μόνο τις ετικέτες, ενώ τα υπόλοιπα δεδομένα παραμένουν αμετάβλητα. Έπειτα το “δηλητηριασμένο” πλέον dataset δίνεται για εκπαίδευση του μοντέλου αφήνοντας τον υπόλοιπο κώδικα απaráλλακτο.

```
# -*- coding: utf-8 -*-
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from keras.models import Sequential
from keras.layers import Dense, Input

dataset= pd.read_csv("C:\CICIoT2023\part-00001-363d1ba3-8ab5-4f96-bc25-4d5862db7cb9-c000.csv",sep=",", decimal=".")

#label flipping attack

n= 60000
random_rows= dataset.sample(n).index

half_indices= dataset.index[random_rows]

shuffled_labels= dataset.loc[half_indices, 'label'].sample(frac=1, random_state=42).values

dataset.loc[half_indices,'label'] = shuffled_labels

x= dataset.drop(columns=["label"], axis=1).values.astype('float32')
y= dataset["label"].apply(lambda x: 1 if x == 'BenignTraffic' else 0).values.astype('float32')

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Define the model
model = Sequential([
    Input(shape=(X_train.shape[1],)),
```

Εικόνα 17

7.3. Αποτελέσματα

Στην εικόνα 17 υπάρχει η νέα απόδοση του μοντέλου μετά από την “δηλητηρίαση” των δεδομένων.

```
Test accuracy: 0.9823130369186401 Test recall: 0.48961424827575684 Test precision: 0.6573705077171326 Test loss: 0.05886618047952652
```

Εικόνα 18

Όπως φαίνεται και στην εικόνα η ακρίβεια του μοντέλου (accuracy) έπεσε περίπου 1% μόνο. Όμως, οι δείκτες των precision και recall έπεσαν περίπου 17,8 και 43,8 αντίστοιχα. Αυτό συμβαίνει γιατί στο συγκεκριμένο μέρος του συνόλου των δεδομένων υπάρχουν 275,258 εγγραφές όπου από αυτές μόλις το 2% είναι απλή δικτυακή κίνηση “Benign Traffic”. Στην εικόνα 18 χρησιμοποιούμε ένα μέρος του dataset που έχει 451,498 εγγραφές, όπου η κανονική κίνηση αποτελεί ξανά το 2% του συνόλου δεδομένων. Ανακατεύοντας πάλι τις ετικέτες από n=60.000 σειρές.

```
Test accuracy: 0.9878737330436707 Test recall: 0.7468839883804321 Test precision: 0.7331764698028564 Test loss: 0.042343780398368835
```

Εικόνα 19

Αυτή την φορά όλες οι μετρικές είναι καλύτερες καθώς το μέγεθος του δείγματος μεγάλωσε,. Όμως, παρόλη την βελτίωση του μοντέλου τα ποσοστά συνεχίζουν να καθιστούν το μοντέλο αναξιόπιστο. Παρακάτω, υπάρχουν πίνακες αναλυτικής σύγκρισης των δύο dataset πριν και μετά από την επίθεση (εικ. 20 & 21).

	Εγγραφές	Benign Traffic	Accuracy	Precision	Recall	Binary cross-entropy loss
Dataset A	275,258	6,387	99.39%	83.50%	92.73%	1.39%
Dataset B	451,498	10,507	99.35%	81.48%	93.34%	1.50%

Εικόνα 20

	Εγγραφές	Accuracy	Precision	Recall	Binary cross-entropy loss	Ποσοστό δηλητηρίασης
Dataset A	275,258	98.23%	65.73%	48.96%	5.88%	21.79%
Dataset B	451,498	98.78%	73.31%	74.68%	4.23%	13.28%

Εικόνα 21

7.4. Εμπόδια και αντιμετώπιση αυτών

Στην συγκεκριμένη έρευνα, υπήρξαν δυσκολίες οι οποίες έπαιξαν καίριο ρόλο στην εξέλιξη και την ολοκλήρωσή της. Τόσο στην συγγραφή του κώδικα όσο και στην κατανόηση του συνόλου δεδομένων και των αποτελεσμάτων για την εξαγωγή των συμπερασμάτων.

Αρχικά, η προσέγγιση για τον διαχωρισμό των επιθέσεων από την απλή κίνηση έγινε δημιουργώντας μία καινούρια στήλη στο αρχείο μέσω του excel. Όμως αυτό είχε ως αποτέλεσμα να τροποποιηθεί το αρχείο με τέτοιο τρόπο ώστε να χρειάζονται διαφορετικές παραμετροποιήσεις για την εισαγωγή του σε σχέση με τα υπόλοιπα που ήταν ακέραια. Αυτό συνέβαινε καθώς, ο διαχωρισμός των στηλών γίνεται με την χρήση του χαρακτήρα κόμμα “,” και τα δεκαδικα ψηφία χρησιμοποιούν τον χαρακτήρα της τελείας “.” όμως, μετά από την τροποποίηση του αρχείου έπρεπε αυτές οι παραμετροποιήσεις να αλλάξουν σε κόμμα “,” και για τις δύο περιπτώσεις. Επιπλέον, θα έπρεπε χειροκίνητα να τροποποιηθούν 169 CSV αρχεία οπότε επιλέχθηκε μία πιο δυναμική προσέγγιση μέσω της python.

Σε αυτό το σημείο πρέπει να ειπωθεί ότι δεν είχε εφαρμοστεί καμία τεχνική προεπεξεργασίας των δεδομένων (data preprocessing). Η ακρίβεια (accuracy) του μοντέλου βρισκόταν στο $\approx 97,5\%$ ενώ μετά από την επίθεση η ακρίβεια έπεφτε στο $\approx 96,5\%$ και η απώλεια δυαδικής διασταυρούμενης εντροπίας (binary crossentropy) ήταν υψηλή. Με την σκέψη ότι αυτό μπορεί να εξηγεί την μικρή πτώση της ακρίβειας, εφαρμόστηκε κανονικοποίηση των δεδομένων χρησιμοποιώντας τον standardscaler με σκοπό την μείωση της συνάρτησης απώλειας και αυτό είχε ως αποτέλεσμα το τελικό μοντέλο να έχει ακρίβεια $\approx 99\%$ όπου μετά από την επίθεση έφτανε στο $\approx 98\%$. Συνεπώς το μοντέλο που αρχικά έπρεπε να “χαλάσει” έγινε ακόμα καλύτερο και η απόκλιση μετά από την επίθεση να παραμείνει ίδια. Όμως, πλέον η απώλεια δυαδικής διασταυρούμενης εντροπίας μειώθηκε.

Έπειτα, υπήρξε σύγχυση με τις τιμές των χαρακτηριστικών, καθώς η εφαρμογή excel απεικόνιζε τους δεκαδικούς αριθμούς ως ακέραιους με αποτέλεσμα να είναι διαφορετικό το περιεχόμενο από αυτό της απεικονίσεις στην python όπως φαίνεται και στις εικόνες 22 & 23. Αυτό έθεσε αμφιβολίες για την σωστή εισαγωγή του αρχείου στην python και την ορθή χρήση την βιβλιοθήκης pandas.

1	flow_duration ▾	Header_Length ▾	Protocol Type ▾	Duration ▾	Rate ▾
2	0	5346	594	6336	1,1458E+16
3	0	540	60	640	1,02782E+16
4	2,20462E+15	9396	60	640	6,71213E+15

Εικόνα 22

	flow_duration	Header_Length	...	Weight	label
0	0.000000	53.46	...	141.55	DDoS-SYN_Flood
1	0.000000	54.00	...	141.55	DDoS-TCP_Flood
2	2.204616	93.96	...	141.55	DDoS-SynonymousIP_Flood
3	0.053618	12497.00	...	141.55	DDoS-UDP_Flood
4	0.000000	0.00	...	141.55	DDoS-ICMP_Flood

Εικόνα 23

Τελος, το πιο παραπλανητικό εμπόδιο ήταν η απόδοση του συστήματος. Στα πειράματα, η μόνη μετρική που χρησιμοποιήθηκε αρχικά ήταν η ακρίβεια (accuracy). Όμως, μετά από την επίθεση η ακρίβεια (accuracy) μειωνόταν μόλις 2% ακόμα και όταν η επίθεση εφαρμοζόταν σε ποσοστό $\geq 50\%$.

Εδώ αξίζει να αναφερθούμε ξανά στο γεγονός ότι η κανονική δικτυακή κίνηση (Benign Traffic) σε κάθε ένα από τα μέρη του dataset, αντιστοιχεί στο 2%. Αυτό σημαίνει ότι ο αλγόριθμος ακόμα και αν ταξινομούσε όλα τα Benign Traffic λανθασμένα ως επιθέσεις, πάλι η ακρίβεια (accuracy) θα ήταν υψηλή. Γι αυτό, έπρεπε να βασιστούμε σε άλλους δείκτες για την μέτρηση της απόδοσης του συστήματος όπως η ακρίβεια θετικής πρόβλεψης (precision) και η ευαισθησία (recall) που αποδίδουν καλύτερα όταν τα δεδομένα δεν είναι κατανεμημένα ομοιόμορφα.

Συμπεράσματα Έρευνας

Στην παρούσα εργασία αναπτύχθηκε ένα Intrusion Detection System με την χρήση τεχνητής νοημοσύνης το οποίο ταξινομεί την κίνηση στο διαδίκτυο των πραγμάτων (IoT) σε απλή και κακόβουλη. Πρόκειται για ένα νευρωνικό δίκτυο με ποσοστό ακρίβειας (accuracy) $\approx 99\%$. Σκοπός της εργασίας υπήρξε η μείωση της αποδοτικότητάς του καθιστώντας το αναξιόπιστο. Για την επίτευξη αυτού του στόχου αξιοποιήθηκε η επίθεση της “δηλητηρίασης” των δεδομένων (data poisoning) και πιο συγκεκριμένα η τεχνική label flipping, κατά την οποία ανακατεύονται οι ετικέτες από ένα υποσύνολο του dataset ώστε να προκληθεί σύγχυση στο μοντέλο. Το μοντέλο μετά από την “δηλητηρίαση” έδωσε ακρίβεια $\approx 98\%$ που πρόκειται για μία αμελητέα διαφορά. Με μία δεύτερη σκέψη όμως, αν συμπεριλάβουμε τους δείκτες precision και recall στις μετρήσεις, είναι εμφανές ότι η αξιοπιστία του μοντέλου έχει κλονιστεί. Αυτό συνέβει γιατί στο σύνολο των δεδομένων δεν υπάρχει ίση κατανομή μεταξύ απλής και κακόβουλης κίνησης. Για την ακρίβεια, παρατηρήθηκε πως σε κάθε μέρος του dataset η απλή κίνηση αποτελεί μόλις το 2% των δειγμάτων. Αυτό σημαίνει ότι η ακρίβεια μπορεί να είναι παραπλανητική. Γι αυτό χρησιμοποιούμε τις άλλες δύο μετρικές. Η ακρίβεια θετικής πρόβλεψης (precision) δίνει βαρύτητα στα False positives, δηλαδή στα δείγματα εκείνα που ταξινομήθηκαν λανθασμένα ως θετικά. Ειδικότερα η ευαισθησία (recall) είναι ίσως η πιο σημαντική, καθώς δίνει μεγαλύτερη βαρύτητα στα False Negatives. Δηλαδή σε αυτά τα δείγματα που ταξινομήθηκαν λανθασμένα ως αρνητικά.

Στην περίπτωση των κυβερνοεπιθέσεων, είναι προτιμότερο να ταξινομηθεί λανθασμένα μια απλή ενέργεια ως κακόβουλη καθώς το αντίστροφο, δηλαδή να ταξινομηθεί μία κακόβουλη ενέργεια ως κανονική, θα ήταν καταστροφικό.

Εν κατακλείδι, η επίθεση της αντιστροφής ετικετών (label flipping attack) πραγματοποιήθηκε με επιτυχία. Σε ένα dataset με κανονική κατανομή μεταξύ απλής και κακόβουλης κίνησης, η συνολική απόδοση του μοντέλου (accuracy) θα είχε μεγαλύτερη απόκλιση. Όμως, στην πραγματική ζωή η κατανομή αυτή είναι κάθε άλλο παρά κανονική. Το παρόν dataset ανταποκρίνεται καλύτερα στην πραγματικότητα στην περίπτωση που αντιστραφούν τα ποσοστά των δύο κλάσεων.

Βιβλιογραφία

1. Russell, S., & Norvig, P. (2021). *Τεχνητή νοημοσύνη* (4η έκδ., Π. Καναβός, Μετάφρ.). Εκδόσεις Κλειδάριθμος. (Πρωτότυπη έκδοση 2020).
2. Xiaojun, L., Xiande, Z., Kexi, Z., Zhenli, D., Kai, Z., & Xi, W. (2021). Study on the application fields and development prospects of artificial intelligence. In *Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)* (pp. 101-105). Dali, China. <https://doi.org/10.1109/ICAIE53562.2021.00029>
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://www.deeplearningbook.org/>
4. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. <https://doi.org/10.1016/j.patrec.2005.10.010>
5. Patil, B., & Gala, D. (2023). An analysis of enabling technologies for the Internet of Things. *International Journal of Technology*, 8, 214-219. <https://doi.org/10.55248/gengpi.2022.31222>
6. Wang, S., Hou, Y., Gao, F., & Ji, X. (2016). A novel IoT access architecture for vehicle monitoring system. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)* (pp. 639-642). IEEE. <https://doi.org/10.1109/WF-IoT.2016.7845396>
7. Abdmeziem, M., Tandjaoui, D., & Romdhani, I. (2015). Architecting the Internet of Things: State of the art. In *Studies in Systems, Decision and Control* (pp. 55-75). Springer. https://doi.org/10.1007/978-3-319-22168-7_3
8. Scarfone, K., & Mell, P. (2007). *Guide to intrusion detection and prevention systems (IDPS)* (NIST Special Publication 800-94). National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=50951
9. Γκρίτζαλης, Σ., Κάτσικας, Σ., & Γκρίτζαλης, Δ. (2003). Ασφάλεια δικτύων υπολογιστών: Τεχνολογίες και υπηρεσίες σε περιβάλλοντα ηλεκτρονικού επιχειρείν και ηλεκτρονικής διακυβέρνησης. Εκδόσεις Παπασωτηρίου.
10. Fan, J., Yan, Q., Li, M., Qu, G., & Xiao, Y. (2022). A survey on data poisoning attacks and defenses. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)* (pp. 48-55). IEEE. <https://doi.org/10.1109/DSC55868.2022.00014>

11. Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., & Goldstein, T. (2024). Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. *arXiv*. <https://arxiv.org/abs/2006.12557>
12. Li, Q., Wang, X., Wang, F., & Wang, C. (2023). A label flipping attack on machine learning model and its defense mechanism. In W. Meng, R. Lu, G. Min, & J. Vaidya (Eds.), *Algorithms and architectures for parallel processing* (pp. 490-506). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22677-9_26
13. Jebreel, N. M., Domingo-Ferrer, J., Sánchez, D., & Blanco-Justicia, A. (2022). Defending against the label-flipping attack in federated learning. *arXiv*. <https://arxiv.org/abs/2207.01982>
14. Tian, Y., Zhang, W., Simpson, A., Liu, Y., & Jiang, Z. L. (2021). Defending against data poisoning attacks: From distributed learning to federated learning. *The Computer Journal*, 66(3), 711-726. <https://doi.org/10.1093/comjnl/bxab192>
15. Paudice, A., Muñoz-González, L., & Lupu, E. C. (2019). Label sanitization against label flipping poisoning attacks. In C. Alzate, A. Monreale, H. Assem, A. Bifet, T. S. Buda, B. Caglayan, B. Drury, E. García-Martín, R. Gavalda, I. Koprinska, S. Kramer, N. Lavesson, M. Madden, I. Molloy, M.-I. Nicolae, & M. Sinn (Eds.), *ECML PKDD 2018 Workshops* (pp. 5-15). Springer International Publishing. https://doi.org/10.1007/978-3-030-13453-2_1
16. Oprea, A., Singhal, A., & Vassilev, A. (2022). Poisoning attacks against machine learning: Can machine learning be trustworthy? *Computer*, 55(11), 94-99. <https://doi.org/10.1109/MC.2022.3190787>
17. Gao, Y., Doan, B. G., Zhang, Z., Ma, S., Zhang, J., Fu, A., Nepal, S., & Kim, H. (2020). Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv*. <https://arxiv.org/abs/2007.10760>
18. Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). CIIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensor* (submitted to *Journal of Sensors*). <https://www.unb.ca/cic/datasets/iotdataset-2023.html>

Εικόνες

Εικόνα 1:

https://gadictos.com/wp-content/uploads/2019/05/1_1mpE6fsq5LNxH31xeTWi5w-scaled.jpeg

Εικόνα 2:

https://assets-global.website-files.com/5d7b77b063a9066d83e1209c/60d2424009416f21db643e21_Group%20807.jpg

Εικόνα 3:<https://www.mdpi.com/2076-3417/10/5/1897>

Εικόνα 4:

https://uploads-ssl.webflow.com/6266b596eef18c1931f938f9/644aea65cefe35380f198a5a_class_guide_cm08.png

Εικόνα 5:

https://www.alps.academy/wp-content/uploads/2022/10/confusion_matrix_evaluation_metrics-768x432.png

Εικόνα 6:<https://ieeexplore.ieee.org/document/7845396>

Εικόνα 7:<https://slideplayer.com/slide/12624146/>

Εικόνα 8:

<https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1>

Εικόνα 9:

https://github.com/michaelTJC96/Label_Flipping_Attack

Εικόνα 10:

https://www.researchgate.net/figure/An-Example-of-Targeted-Poisoning-Attacks_fig3_348214457

Εικόνα 11:https://www.youtube.com/watch?v=Eeois_a8Gfw

Εικόνες 12 & 13:<https://arxiv.org/abs/2007.10760>