ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής

Εργασία Μαθήματος ***Διοίκηση Πληροφοριακών Συστημάτων***

Αρ. Άσκησης – Τίτλος Ασκησης	***Διαχείριση Ασφάλειας Συστημάτων Τεχνητής Νοημοσύνης (ΤΝ)***
Όνομα φοιτητή – Αρ. Μητρώου (όλων σε περίπτωση ομαδικής εργασίας)	Γιαννίκος Παναγιώτης – ΜΠΚΕΔ24007
	Μπαλτζής Δημήτρης – ΜΠΚΕΔ24026
	Ραυτόπουλος Μάριος – ΜΠΚΕΔ24034
Ημερομηνία παράδοσης	25/02/2025

Εκφώνηση της άσκησης

Διαχείρηση Ασφάλειας Συστημάτων Τεχνητής Νοημοσύνης (ΤΝ)

Περιγραφή:

Επιλεξτε ένα συστημα ΤΝ που χρησιμοιπειται σε κάποιο κρίσιμο οργανισμό (πχ υγείας, χρηματοοικονομικό, παιδείας, κλπ) και

1) Αναλύστε τα στοιχεία του συστήματος

2) Προσδιορίστε το κύκλο ζωής τους και τους εμπλεκόμενους χρήστες

3) Σε κάθε φάση του κύκλοθ ζωής και καθε στοιχείο του συστήματος ΤΝ , βείτε τις απειλές (χρησιμοποιώντας την βάση του OWASP)

4) Προτείνετε τα αντίμετρα για κάθε απειλή

5) Ακολουθήστε τα βήματα ανάλυσης επικινδυνότητητας και υπολογίστε την επικνδυνότητα για κάθε απειλή, για κάθε στοιχείο του ΤΝ συστήματος για κάθε φάση του κύκλου ζωής του συστήματος

Περιεχόμενα

1. Εισαγωγή

Για την ανάλυση ενός συστήματος ΤΝ επιλέγεται ένα νοητό σύστημα ΤΝ (SteelInspectAI) που χρησιμοποιείται σε ένα εργοστάσιο χάλυβα για την αυτοματοποιημένη αξιολόγηση της ποιότητας των παραγόμενων προϊόντων. Το σύστημα αυτό αξιολογεί την ποιότητα του χάλυβα που παράγεται σε πραγματικό χρόνο, εντοπίζοντας ελαττώματα και αποκλίσεις από τα πρότυπα. Το σύστημα ενσωματώνεται στην παραγωγική διαδικασία και παρέχει αναφορές στους διαχειριστές του εργοστασίου.

Οι κύριες λειτουργίες του συστήματος ΤΝ είναι οι εξής:

- Ανάλυση οπτικών δεδομένων από αισθητήρες και κάμερες.
- Συγκριτική αξιολόγηση με πρότυπα ποιότητας.
- Παροχή συστάσεων για διορθωτικές ενέργειες.

2. Ανάλυση Στοιχείων του Συστήματος

Ένα σύστημα ΤΝ αποτελείται από επιμέρους συστατικά στοιχεία που συνεργάζονται για την επίτευξη ενός συγκεκριμένου σκοπού, συνθέτοντας τη συνολική του λειτουργία.

Τα δομικά συστατικά ενός συστήματος ΤΝ, και συγκεκριμένα του οργανισμού είναι τα εξής:

1. **Είσοδος (Δεδομένα εισόδου)**. Η είσοδος αποτελεί τη βάση της μάθησης και της λήψης αποφάσεων ενός συστήματος ΤΝ, καθώς αξιοποιείται για την προσαρμογή και λειτουργία του μέσω της ανάλυσης δεδομένων.

   Στον οργανισμό τα δεδομένα εισόδου αποτελούνται από εικόνες/βίντεο προϊόντων, δεδομένα αισθητήρων, ιστορικά δεδομένα ποιότητας και προδιαγραφές προϊόντων. Τα οπτικά δεδομένα προέρχονται από τη χρήση βιομηχανικών καμερών και σχετίζονται με την επιφάνεια και τη δομή των προϊόντων χάλυβα. Τα δεδομένα αισθητήρων αφορούν πληροφορίες σχετικά με τη διαδικασία παραγωγής, όπως θερμοκρασίες, πίεση ή πάχος.

   Τα ιστορικά δεδομένα ποιότητας είναι δεδομένα από προηγούμενες αξιολογήσεις για τη δημιουργία προτύπων ή τη σύγκριση. Οι προδιαγραφές προϊόντων αφορούν δεδομένα για τις απαιτούμενες προδιαγραφές ποιότητας, όπως ανοχές σε πάχος, σκληρότητα, ή ομαλότητα.
2. **Αλγόριθμοι και Μοντέλα.** Οι αλγόριθμοι και τα μοντέλα αποτελούν τον πυρήνα του συστήματος ΤΝ. Είναι υπεύθυνα για την επεξεργασία των δεδομένων και τη λήψη αποφάσεων.

   Στον οργανισμό οι αλγόριθμοι και τα μοντέλα περιλαμβάνουν ανίχνευση και ταξινόμηση ατελειών, πρόβλεψη ποιότητας και βελτιστοποίηση παραγωγής. Στην βελτιστοποίηση παραγωγής, αναλύονται οι αιτίες αποβλήτων και παρέχονται συστάσεις για την προσαρμογή της διαδικασίας παραγωγής.

3. **Διαδικασίες Λειτουργίας.** Στις διαδικασίες λειτουργίας περιλαμβάνεται η εκπαίδευση, η δοκιμή και αξιοπιστία, η λειτουργία και η επανεκπαίδευση. Ο σκοπός των διαδικασιών αυτών είναι η διασφάλιση της σωστής λειτουργίας και η συνεχής βελτίωση του συστήματος.
4. **Έξοδος (Δεδομένα εξόδου).** Οι έξοδοι είναι το τελικό αποτέλεσμα της λειτουργίας του συστήματος ΤΝ. Περιλαμβάνουν προβλέψεις, συστάσεις ή δράσεις που εκτελεί το σύστημα. Στον οργανισμό η έξοδος αφορά αναφορές ποιότητας, που περιγράφουν την ποιότητα κάθε προϊόντος, ειδοποιήσεις, για ανίχνευση σοβαρών προβλημάτων και συστάσεις βελτιστοποίησης, παρέχοντας συμβουλές για ρύθμιση μηχανημάτων ή διαδικασιών για τη μείωση ελαττωμάτων.
5. **Υποδομή.** Στην υποδομή περιλαμβάνονται τα επιμέρους στοιχεία που χρησιμοποιούνται για τον σχεδιασμό και τη λειτουργία ενός συστήματος ΤΝ, όπως το υλικό και το λογισμικό. Στον οργανισμό κάποια από τα μέρη του υλικού περιλαμβάνουν βιομηχανικές κάμερες υψηλής ανάλυσης, εξοπλισμό αισθητήρων για δεδομένα παραγωγής και εξυπηρετητές επεξεργασίας δεδομένων παραγωγής. Στο λογισμικό περιλαμβάνονται βιβλιοθήκες AI, συστήματα διαχείρισης δεδομένων για αποθήκευση εισερχόμενων και εξερχόμενων δεδομένων.

Το σύστημα ΤΝ του εργοστασίου αποτελείται από κάποια αγαθά (assets) τα οποία συνεργάζονται μεταξύ τους για την βελτιστοποίηση της παραγωγής και της ποιότητας των προϊόντων. Τα αγαθά μπορούν να κατηγοριοποιηθούν με βάση τη λειτουργία και τη σχέση τους με το σύστημα συνολικά. Τα αγαθά που είναι ενσωματωμένα στο σύστημα ΤΝ κατηγοριοποιούνται ως εξής:

1. **Δεδομένα και Βάσεις Δεδομένων**

   SteelData_v1: Βάση δεδομένων με ιστορικά δεδομένα παραγωγής και ποιότητας. Χρησιμοποιείται για την τροφοδότηση του μοντέλου ΤΝ με δεδομένα εκπαίδευσης και παρέχει πληροφορίες για την ανίχνευση ανωμαλιών.

   Backup_Storage_AWS: Αντίγραφα ασφαλείας δεδομένων για αποκατάσταση σε περίπτωση απώλειας. Διασφαλίζει την ακεραιότητα και τη διαθεσιμότητα των ιστορικών δεδομένων.
2. **Υποδομές Μηχανικής Μάθησης**

   QualityAI_v2: Μοντέλο ΤΝ για την αξιολόγηση ποιότητας. Επεξεργάζεται εικόνες από τις κάμερες και αναγνωρίζει ελαττωματικά προϊόντα.
3. **Εξοπλισμός Παρακολούθησης και Ανάλυσης**

   InspectionCameras&Sensors: Κάμερες υψηλής ανάλυσης και σένσορες τοποθετημένα στη γραμμή παραγωγής, τα οποία συλλέγουν δεδομένα, όπως εικόνες προϊόντων, για ανάλυση.
4. **Υποδομές επεξεργασίας και δικτύωσης**

   SteelInspectAIServer: Εξυπηρετητής εκτέλεσης της υπηρεσίας διασύνδεσης με το σύστημα AI.

FactoryServer: Εξυπηρετητής επεξεργασίας δεδομένων παραγωγής. Εκτελεί τις αναλύσεις του μοντέλου QualityAI_v2 και αποθηκεύει προσωρινά τα δεδομένα.

Firewall_FORTINET01: Προστασία του δικτύου εργοστασίου, διασφαλίζοντας ότι το δίκτυο είναι προστατευμένο από κυβερνοεπιθέσεις.

5. **Διεπαφές και Διαχείριση**

   FactoryControlPortal: Διαδικτυακή πλατφόρμα για την παρακολούθηση της παραγωγής και τη δημιουργία αναφορών.

   MaintenanceBot: Αυτοματοποιημένος μηχανισμός διαγνωστικών και συντήρησης. Ανιχνεύει προβλήματα στο σύστημα και προγραμματίζει διαδικασίες συντήρησης.

6. **Εξειδικευμένο Ανθρώπινο Δυναμικό**

   Alex Steel (Data Analyst): Αναλυτής υπεύθυνος για την ερμηνεία αποτελεσμάτων από το μοντέλο TN. Προσαρμόζει τη στρατηγική παραγωγής βάσει δεδομένων.

   John Greek (AI/M Engineer): Υπεύθυνος για την ανάπτυξη και τη συντήρηση του μοντέλου TN. Παρακολουθεί τη λειτουργία του συστήματος, διασφαλίζει τη βελτιστοποίηση και τη συμμόρφωση με κανονισμούς.

3. Κύκλος Ζωής και Εμπλεκόμενοι Χρήστες

Ο κύκλος ζωής ενός συστήματος TN διαφέρει από εκείνον ενός παραδοσιακού πληροφοριακού συστήματος, καθώς αλλάζει δυναμικά με το χρόνο.

Το σύστημα TN ακολουθεί έναν πλήρη κύκλο ζωής που περιλαμβάνει τον σχεδιασμό, την ανάπτυξη, τον έλεγχο, την εφαρμογή, τη συντήρηση, την αναβάθμιση και την απόσυρση. Οι εμπλεκόμενοι χρήστες περιλαμβάνουν Domain Experts, επιχειρηματικούς ενδιαφερόμενους (Business Stakeholders), Data Scientists, Data Engineers, Machine Learning Engineers, DevOps Engineers, IT Specialists.

Οι φάσεις του κύκλου ζωής του συστήματος Τεχνητής Νοημοσύνης περιγράφονται παρακάτω, ακολουθούμενες από τα κύρια χαρακτηριστικά στοιχεία και τους εμπλεκόμενους χρήστες που αντιστοιχούν σε κάθε φάση.

Φάσεις Κύκλου Ζωής:

1. **Σχεδίαση:**
   - Καθορισμός απαιτήσεων και βασικών στόχων του συστήματος TN, όπως το σκοπό επίτευξης του μοντέλου, το είδος των δεδομένων και του προβλήματος.
   - Σχεδίαση διαδικασιών συλλογής, οργάνωσης και επεξεργασίας των δεδομένων.
   - Επιλογή του τύπου των αλγορίθμων και της προσέγγισης που θα χρησιμοποιηθεί (πχ μηχανική μάθηση, νευρωνικά δίκτυα).

Εμπλεκόμενοι: Domain Experts, επιχειρηματικοί ενδιαφερόμενοι, Data Scientists, Data Engineers.

2. **Ανάπτυξη:**
    - Επιλογή κατάλληλων αλγορίθμων.
    - Εκπαίδευση του μοντέλου με τα δεδομένα εκπαίδευσης.
    - Επικύρωση του μοντέλου με χρήση δεδομένων επικύρωσης (validation data).

    Εμπλεκόμενοι: Data Scientists, Machine Learning Engineers, Domain Experts.
3. **Έλεγχος**:
    - Αξιολόγηση της ακρίβειας και της αξιοπιστίας του μοντέλου, με τη χρήση δεδομένων δοκιμής (test data).
    - Ανίχνευση και διόρθωση προβλημάτων.

    Εμπλεκόμενοι: Data Scientists, QA Engineers, Domain Experts.
4. **Εφαρμογή:**
    - Εγκατάσταση του συστήματος σε περιβάλλον παραγωγής.
    - Ενσωμάτωση στα υπάρχοντα συστήματα και διαδικασίες.
    - Ρύθμιση των παραμέτρων της παραγωγικής υποδομής.

    Εμπλεκόμενοι: Machine Learning Engineers, DevOps Engineers, IT Specialists.
5. **Συντήρηση:**
    - Παρακολούθηση της απόδοσης του συστήματος σε πραγματικό χρόνο.
    - Ενημέρωση του μοντέλου σε περίπτωση αλλαγών στα δεδομένα.
    - Διόρθωση σφαλμάτων και ενημερώσεις ασφαλείας.

    Εμπλεκόμενοι: Machine Learning Engineers, Data Engineers, IT Specialists.
6. **Αναβάθμιση και Απόσυρση:**
    - Αναβάθμιση συστήματος για βελτίωση απόδοσης ή προσθήκη νέων λειτουργιών.
    - Απόσυρση παρωχημένων συστημάτων που δεν είναι πλέον χρήσιμα.

    Εμπλεκόμενοι: Machine Learning Engineers, Data Engineers, επιχειρηματικοί ενδιαφερόμενοι.

4. Αναγνώριση Απειλών

Το σύστημα ΤΝ αντιμετωπίζει διάφορες απειλές. Οι απειλές αυτές μπορούν να κατηγοριοποιηθούν με βάση τον OWASP σε τρεις κατηγορίες:

1. κατά τη διάρκεια ανάπτυξης (κατά την απόκτηση και προετοιμασία των δεδομένων, και κατά την εκπαίδευση/παραγωγή του μοντέλου),
2. κατά τη χρήση του μοντέλου (παρέχοντας είσοδο και διαβάζοντας την έξοδο), και

3. κατά τη διάρκεια εκτέλεσης σε περιβάλλον παραγωγής.

Η ανάλυση των διάφορων απειλών γίνεται με βάση τις κατηγορίες που ορίζονται από τον OWASP. Κάθε φάση του κύκλους ζωής του συστήματος που περιγράφηκε, εμπεριέχεται σε μία από αυτές τις τρεις κατηγορίες. Οι απειλές του συστήματος AI φαίνονται στον παρακάτω πίνακα (Πίνακας 1) και ταξινομούνται με βάση το κύκλο ζωής, το πεδίο επίθεσης και το αντίκτυπο της απειλής.

Κύκλος Ζωής (lifecycle)	**Πεδίο Επίθεσης (attack surface)**	**Απειλή (Threat/Risk category)**	**Αγαθό (Asset)**	**Αντίκτυπο (Impact)**
Development	Engineering environment	Δηλητηρίαση Δεδομένων (Data Poisoning)	QualityAI_v2 (Model behavior)	Ακεραιότητα
		Δηλητηρίαση Μοντέλου κατά την ανάπτυξη (Model poisoning development time)		
Operation	Model Use (provide input/ read output)	Άμεση έγχυση προτροπής (Direct prompt injection)	FactoryServer	
		Έμμεση έγχυση προτροπής (Indirect prompt injection)		
		Παράκαμψη ασφαλείας (Evasion)	QualityAI_v2	
	Break into deployed model	Δηλητηρίαση Μοντέλου κατά την εκτέλεση (Model poisoning in runtime)	QualityAI_v2 FactoryServer	
Operation	Model use	Αποκάλυψη δεδομένων στην έξοδο του μοντέλου (Data disclosure in model output)	QualityAI_v2	Εμπιστευτικότητα
Development	Engineering environment	Παραβίαση Ελέγχου Πρόσβασης (Broken Access Control)	SteelData_v1 Backup_Storage_AWS	
Development	Engineering environment	Διαρροή δεδομένων (Training data leak)	SteelData_v1	
Operation	Model use	Άρνηση εξυπηρέτησης μοντέλου (Denial of model service)	SteelInspectAIServer	Διαθεσιμότητα
Operation	IT	Spoofing Attack	InspectionCameras&Sensors	Εμπιστευτικότητα,

		IP Spoofing Firewall Rule Bypass	Firewall_FORTINET01	Ακεραιότητα, Διαθεσιμότητα
		SQL Injection	FactoryControlPortal	
		Spoofing, Injection Attacks	MaintenanceBot	

Πίνακας 1. Καταγραφή στοιχείων του συστήματος με τις απειλές σε κάθε φάση του κύκλου ζωής

Το πεδίο επίθεσης (attack surface) αφορά στο σύνολο των σημείων εισόδου και αλληλεπίδρασης του συστήματος που μπορούν να αποτελέσουν στόχο για κακόβουλες επιθέσεις. Στον παραπάνω πίνακα ανάλογα με το κύκλο ζωής και το πεδίο επίθεσης γίνεται ο διαχωρισμός και η κατηγοριοποίηση των απειλών.

Η ανάπτυξη (development) στον κύκλο ζωής αναφέρεται στην φάση ανάπτυξης ενώ η λειτουργία (operation) στη φάση λειτουργίας του μοντέλου. Στην περίπτωση λειτουργίας εντάσσονται η χρήση του μοντέλου (Model Use), δηλαδή η εφαρμογή δεδομένων εισόδου και η παραγωγή δεδομένων εξόδου στο μοντέλο, καθώς και η εκτέλεση του σε περιβάλλον παραγωγής.
Με το engineering environment γίνεται αναφορά στα εργαλεία, στην υποδομή και στις διαδικασίες που χρησιμοποιούνται για τη σχεδίαση, την ανάπτυξη, την εκπαίδευση, τον έλεγχο και την εφαρμογή των AI μοντέλων.

4.1 Απειλές κατά της ακεραιότητας

Η ακεραιότητα του μοντέλου μπορεί να παραβιαστεί τόσο κατά την ανάπτυξη όσο και κατά την λειτουργία του. Κατά την ανάπτυξη του μοντέλου εντοπίζονται κάποιες απειλές οι οποίες αφορούν δηλητηρίαση δεδομένων και δηλητηρίαση μοντέλου. Στη δηλητηρίαση δεδομένων ο επιτιθέμενος παραποιεί ή αλλοιώνει τα δεδομένα εκπαίδευσης ενός μοντέλου, ώστε να επηρεάσει τη συμπεριφορά του. Για παράδειγμα στο μοντέλο QualityAI_v2 θα μπορούσαν να αλλοιωθούν οι ετικέτες δεδομένων εκπαίδευσης ώστε προϊόντα χαμηλής ποιότητας να μαρκάρονται ως άριστα. Έτσι, μπορεί να προκύψει εκπαίδευση του μοντέλου με εσφαλμένα δεδομένα.

Η δηλητηρίαση μοντέλου αποτελεί μία πιο στοχευμένη μορφή δηλητηρίασης δεδομένων, όπου ο επιτιθέμενος μπορεί να επηρεάσει το μοντέλο απευθείας κατά την εκπαίδευση του, εισάγοντας κακόβουλα μοτίβα. Με αυτό τον τρόπο, οι λανθασμένες καταχωρήσεις μπορεί να οδηγήσουν σε κακή ταξινόμηση, παραβιάσεις ασφαλείας, ή ενσωμάτωση backdoors.

Κατά τη χρήση του μοντέλου εντοπίζονται οι απειλές άμεση έγχυση προτροπής (indirect prompt injection), έμμεση έγχυση προτροπής (indirect prompt injection), παράκαμψη ασφαλείας (evasion). Στις δυο πρώτες επιθέσεις εισάγονται επιβλαβείς εντολές και πληροφορίες στο μοντέλο με σκοπό την παράκαμψη των κανόνων ασφαλείας του AI. Η τρίτη επίθεση έχει ως στόχο την παραπλάνηση του συστήματος AI με σκοπό να λειτουργεί με μη αναμενόμενο τρόπο. Στην περίπτωση του συστήματος AI στο εργοστασιακό περιβάλλον, οι επιθέσεις αυτές μπορεί να παρακάμψουν

σημαντικούς κανόνες ασφαλείας και να αποκτηθεί πρόσβαση σε ευαίσθητα δεδομένα του οργανισμού.

Μία άλλη απειλή είναι η δηλητηρίαση μοντέλου κατά την εκτέλεση, όπου μπορεί να επηρεαστεί η συμπεριφορά ή η απόδοση του μοντέλου σε πραγματικό χρόνο. Αυτό μπορεί κυρίως να συμβεί όταν το μοντέλο αλληλεπιδρά με κάποιο εξωτερικό ή δυναμικό περιβάλλον, το οποίο μπορεί να γίνει αντικείμενο εκμετάλλευσης. Στο εργοστασιακό περιβάλλον, μία πιθανή εκμετάλλευση είναι κατά τη διάρκεια της επανεκπαίδευσης να εισαχθούν κακόβουλα δεδομένα στην είσοδο, για παράδειγμα μέσω API ή άλλων διαύλων, τα οποία μπορεί να επηρεάσουν αρνητικά τη λειτουργία του μοντέλου.

4.2 Απειλές κατά της εμπιστευτικότητας

Οι απειλές που σχετίζονται με την παραβίαση της εμπιστευτικότητας, αφορούν κυρίως αποκάλυψη ευαίσθητων δεδομένων. Μία απειλή κατά την λειτουργία του μοντέλου είναι η αποκάλυψη δεδομένων στην έξοδο του μοντέλου, καθώς το QualityAI_v2 χρησιμοποιεί εικόνες από κάμερες για την ανίχνευση ελαττωματικών προϊόντων. Έτσι αν η έξοδός του δεν έχει σχεδιαστεί σωστά, μπορεί να αποκαλύψει ευαίσθητες πληροφορίες, είτε άμεσα είτε έμμεσα.

Ένας άλλος τύπος απειλής αφορά στην παραβίαση ελέγχου πρόσβασης, η οποία σχετίζεται με τη μη εξουσιοδοτημένη πρόσβαση σε ευαίσθητα δεδομένα κατά την ανάπτυξη του μοντέλου.
Τα SteelData_v1 και Backup_Storage_AWS είναι συστήματα που αποθηκεύουν ή διαχειρίζονται κρίσιμα δεδομένα παραγωγής. Αν δεν υπάρχει σωστός έλεγχος πρόσβασης, επιτιθέμενοι ή μη εξουσιοδοτημένοι χρήστες μπορούν να αποκτήσουν πρόσβαση σε αυτά και να διαρρεύσουν, τροποποιήσουν ή διαγράψουν σημαντικά δεδομένα.

Ακόμη μία απειλή είναι η διαρροή των δεδομένων εκπαίδευσης του μοντέλου. Το SteelData_v1 περιέχοντας τα δεδομένα εκπαίδευσης του μοντέλου αποκάλυψη των δεδομένων εκπαίδευσης του μοντέλου, που μπορεί να περιέχουν εμπιστευτικές ή εμπορικά ευαίσθητες πληροφορίες.
Το SteelData_v1 περιέχει και διαχειρίζεται τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου AI, σχετικού με την παραγωγή και την ανίχνευση ελαττωματικών προϊόντων. Αν αυτά τα δεδομένα διαρρεύσουν, μπορεί να εκτεθεί κρίσιμη επιχειρησιακή γνώση, ευαίσθητες πληροφορίες του οργανισμού, όπως αυτές των παραγωγικών διαδικασιών.

4.3 Απειλές κατά της διαθεσιμότητας

Σε αυτή τη κατηγορία περιέχονται οι απειλές που μπορούν να προκαλέσουν διακοπή της ορθής λειτουργίας του συστήματος AI. Μία συνηθισμένη επίθεση είναι η άρνηση εξυπηρέτησης μοντέλου κατά τη φάση λειτουργίας του, όπου μπορεί να επηρεαστεί η παραγωγική διαδικασία.

4.4 Απειλές κατά των τριών θεμέλιων αρχών ασφάλειας

Σε αυτή την κατηγορία περιέχονται κάποιες απειλές σε βασικά συστατικά μέρη του IT μέρους, τα οποία μπορούν να επηρεάσουν τα τρία συστατικά της ασφάλειας. Κάποιες από τις απειλές αυτές σχετίζονται με την παραποίηση ταυτότητας για την απόκτηση πρόσβασης και εκμετάλλευσης των ευαίσθητων πληροφοριών. Για παράδειγμα στην περίπτωση των καμερών και αισθητήρων μία απειλή σχετίζεται με την μη εξουσιοδοτημένη πρόσβαση και τη πιθανή τροποποίηση ή παραποίηση των μετρήσεων. Άλλες απειλές σχετίζονται με το τείχος προστασίας, τη διαδικτυακή πλατφόρμα για την παρακολούθηση της παραγωγής και τον αυτοματοποιημένο μηχανισμό διαγνωστικών και συντήρησης. Σε αυτές τις περιπτώσεις οι απειλές σε αυτά τα μέρη μπορούν επηρεάσουν είτε έμμεσα είτε άμεσα το σύστημα AI.

5. Αντίμετρα για την αντιμετώπιση των απειλών

Για κάθε απειλή η λήψη αντιμέτρων είναι απαραίτητη για να μετριαστούν ή να εξαλειφθούν οι απειλές. Στο παρακάτω πίνακα για κάθε απειλή που αντιμετωπίζει ένα αγαθό δίνεται μία περιγραφή της και τα αντίμετρα που μπορούν να εφαρμοστούν.

Απειλή (Threat/Risk category)	**Αγαθό (Asset)**	**Περιγραφή Απειλής**	**Αντίμετρα**
Δηλητηρίαση Δεδομένων (Data Poisoning)	QualityAI_v2 (Model behavior)	Εισαγωγή κακόβουλα τροποποιημένων δεδομένων στο σύστημα με σκοπό να επηρεάσουν τη συμπεριφορά του μοντέλου.	• Επικύρωση των δεδομένων με αξιόπιστες πηγές πριν χρησιμοποιηθούν στην εκπαίδευση. • Χρήση κατακερματισμού ή ψηφιακών υπογραφών για τη διασφάλιση της μη τροποποίησης των δεδομένων.
Δηλητηρίαση Μοντέλου κατά την ανάπτυξη (Model poisoning development time)		Παρέμβαση στη διαδικασία εκπαίδευσης ή τροποποίηση των βαρών του μοντέλου, με σκοπό τη λανθασμένη συμπεριφορά του μοντέλου.	• Εφαρμογή μηχανισμούς ελέγχου πρόσβασης για τον περιορισμό της μη εξουσιοδοτημένης πρόσβασης. • Χρήση κατακερματισμού ή ψηφιακών υπογραφών για τη διασφάλιση της μη τροποποίησης του μοντέλου.
Άμεση έγχυση προτροπής (Direct prompt injection)	FactoryServer	Εισαγωγή σκόπιμα διαμορφωμένων προτροπών με στόχο τον έλεγχο της συμπεριφοράς του μοντέλου	• Μηχανισμοί ανίχνευσης και απόρριψης κακόβουλων προτροπών

Έμμεση έγχυση προτροπής (Indirect prompt injection)	FactoryServer	Έμμεση κακόβουλη προτροπή, όπως κείμενα, ιστοσελίδες, με στόχο την επίδραση στη συμπεριφορά του μοντέλου	• Μηχανισμοί ανίχνευσης και απόρριψης κακόβουλων προτροπών
Παράκαμψη ασφαλείας (Evasion)	QualityAI_v2	Τεχνική αποφυγής εντοπισμού ή αποτροπής από μηχανισμούς ασφαλείας, όπως ανίχνευσης εισβολών ή φίλτρα ασφαλείας	• Συστηματική παρακολούθηση και ανάλυση της συμπεριφοράς του συστήματος σε πραγματικό χρόνο, με σκοπό τον εντοπισμό μη αναμενόμενης συμπεριφοράς σχετικά με την παράκαμψη ασφαλείας
Δηλητηρίαση Μοντέλου κατά την εκτέλεση (Model poisoning in runtime)	QualityAI_v2 FactoryServer	Τροποποίηση του μοντέλου κατά τη φάση λειτουργίας του με σκοπό την αλλοίωση της συμπεριφοράς του	• Περιορισμός πρόσβασης σε μη εξουσιοδοτημένους χρήστες • Εφαρμογή μηχανισμών για την εξασφάλιση της ακεραιότητας των παραμέτρων του μοντέλου
Αποκάλυψη δεδομένων στην έξοδο του μοντέλου (Data disclosure in model output)	QualityAI_v2	Αποκάλυψη ευαίσθητων ή εμπιστευτικών πληροφοριών με τις οποίες έχει εκπαιδευτεί το μοντέλο	• Εφαρμογή φίλτρων στην έξοδο για το έλεγχο πιθανής διαρροής ευαίσθητων πληροφοριών πριν αυτές κοινοποιηθούν στους τελικούς χρήστες. • Ενσωμάτωση μηχανισμών στο μοντέλο για την αναγνώριση και αποφυγή παραγωγής ευαίσθητων πληροφοριών στην έξοδο.
Παραβίαση Ελέγχου Πρόσβασης (Broken Access Control)	SteelData_v1 Backup_Storage_AWS	Μη εξουσιοδοτημένοι χρήστες αποκτούν πρόσβαση σε πόρους ή δεδομένα λόγω ανεπαρκών μηχανισμών ελέγχου πρόσβασης	• Καταγραφή και παρακολούθηση των προσβάσεων στα δεδομένα, με σκοπό τον εντοπισμό μη εξουσιοδοτημένης δραστηριότητας
Διαρροή δεδομένων (Training data leak)	SteelData_v1	Κακόβουλη αποκάλυψη των δεδομένων που χρησιμοποιούνται για την εκπαίδευση του μοντέλου	• Εφαρμογή κρυπτογράφησης για την ασφαλή αποθήκευση των δεδομένων εκπαίδευσης • Απόκρυψη των ευαίσθητων πληροφοριών, μειώνοντας τον κίνδυνο διαρροής σε περίπτωση αποκάλυψης
Άρνηση εξυπηρέτησης μοντέλου	SteelInspectAIServer	Αποτροπή της ορθής λειτουργίας του συστήματος	• Περιορισμός ρυθμού αιτημάτων

(Denial of model service)		μέσω της μη διαθεσιμότητας του	• Ανίχνευση και αποκλεισμός κακόβουλων διευθύνσεων IP • Εξισορρόπηση φορτίου για τη καλύτερη διαχείριση του φόρτου των αιτημάτων
Spoofing Attack	InspectionCameras&Sensors	Παραπλάνηση των συστημάτων παρακολούθησης με σκοπό τις λανθασμένες αναλύσεις, αποφάσεις ή ενέργειες του συστήματος	• Εφαρμογή μηχανισμών για τον έλεγχο της ακεραιότητας των δεδομένων που λαμβάνονται από τα συστήματα αυτά • Χρήση πολλαπλών αισθητήρων και καμερών για το συνδυασμό των δεδομένων από διαφορετικές πηγές
Spoofing Attack Firewall Rule Bypass	Firewall_FORTINET01	Παράκαμψη κανόνων τείχους προστασίας για την απόκτηση πρόσβασης στο δίκτυο που βρίσκεται το σύστημα AI	• Εφαρμογή Φιλτραρίσματος MAC • Χρήση στατικών διευθύνσεων IP
SQL Injection	FactoryControlPortal	Εισαγωγή κακόβουλου κώδικα σε ερωτήματα SQL μέσω εισόδων χρήστη, με στόχο την πρόσβαση, τροποποίηση ή καταστροφή δεδομένων στη βάση δεδομένων	• Χρήση παραμετροποιημένων ερωτημάτων αντί για δυναμική δημιουργία ερωτημάτων SQL • Έλεγχος και επικύρωση όλων των εισόδων των χρηστών, απορρίπτοντας οποιαδήποτε μη αναμενόμενη ή κακόβουλη είσοδο
Spoofing, Injection Attacks	MaintenanceBot	Παραποίηση ταυτότητας με σκοπό την μη εξουσιοδοτημένη πρόσβαση στον μηχανισμό συντήρησης. Εισαγωγή κακόβουλου κώδικα με σκοπό την εκτέλεση μη εξουσιοδοτημένων εντολών	• Εφαρμογή μηχανισμών επαλήθευσης ταυτότητας των χρηστών που επικοινωνούν με το σύστημα αυτό • Περιορισμός δικαιωμάτων ελαχιστοποιώντας τα δικαιώματα, ώστε ο κακόβουλος κώδικας να μην μπορεί να εκτελεί επιβλαβείς ενέργειες

Πίνακας 2. Περιγραφή απειλών και αντιμέτρων

6. Ανάλυση Επικινδυνότητας

Η ανάλυση επικινδυνότητας (Risk Assessment) του συστήματος πραγματοποιείται για κάθε αγαθό και κάθε απειλή και βασίζεται στο επίπεδο επικινδυνότητας της απειλής στο αγαθό (Threat), στο επίπεδο της ευπάθειας του αγαθού στην απειλή (Vulnerability), και στη

συνέπεια στο αγαθό (Likelihood of Impact), που είναι η πιθανότητα να συμβεί ένας γεγονός και ποσό σοβαρές θα είναι οι επιπτώσεις του. Η ανάλυση επικινδυνότητας υπολογίζεται ως το γινόμενο των παραπάνω χαρακτηριστικών και δίνεται από τον παρακάτω τύπο

*Risk=Threat * Vulnerability * Likelihood of Impact*

Για κάθε ένα από τα χαρακτηριστικά της ανάλυσης επιλέγεται κλίμακα 1-5
(1 Πολύ Χαμηλή, 2 Χαμηλή, 3 Μεσαία, 4 Υψηλή, 5 Πολύ Υψηλή)

Κύκλος Ζωής (lifecycle)	**Πεδίο Επίθεσης (attack surface)**	**Απειλή (Threat/Risk category)**	**Αγαθό (Asset)**	**Threat** [1-5]	**Vulnerability** [1-5]	**Likelihood of Impact** [1-5]	**Risk** [1-5]
Development	Engineering environment	Δηλητηρίαση Δεδομένων (Data Poisoning)	QualityAI_v2 (Model behavior)	4	4	5	80
		Δηλητηρίαση Μοντέλου κατά την ανάπτυξη (Model poisoning development time)	QualityAI_v2 (Model behavior)	3	4	4	48
Operation	Model Use (provide input/ read output)	Άμεση έγχυση προτροπής (Direct prompt injection)	FactoryServer	3	3	4	48
		Έμμεση έγχυση προτροπής (Indirect prompt injection)	FactoryServer	2	2	4	16
		Παράκαμψη ασφαλείας (Evasion)	QualityAI_v2	4	4	3	48
	Break into deployed model	Δηλητηρίαση Μοντέλου κατά την εκτέλεση (Model poisoning in runtime)	QualityAI_v2, FactoryServer	4	4	4	64
Operation	Model use	Αποκάλυψη δεδομένων στην έξοδο του μοντέλου (Data disclosure in model output)	QualityAI_v2	4	4	5	80

Development	Engineering environment	Παραβίαση Ελέγχου Πρόσβασης (Broken Access Control)	SteelData_v1, Backup_Storage_AWS	4	5	4	80
Development	Engineering environment	Διαρροή δεδομένων (Training data leak)	SteelData_v1	3	4	4	48
Operation	Model use	Άρνηση εξυπηρέτησης μοντέλου (Denial of model service)	SteelInspectAIServer	3	3	3	27
Operation	IT	Spoofing Attack	InspectionCameras&Sensors	4	3	4	48
		IP Spoofing Firewall Rule Bypass	Firewall_FORTINET01	3	4	4	48
		SQL Injection	FactoryControlPortal	3	3	3	27
		Spoofing, Injection Attacks	MaintenanceBot	3	3	4	36

Πίνακας 3. Πίνακας ποσοτικής ανάλυσης επικινδυνότητας

Οι τιμές στον πίνακα προκύπτουν από την αξιολόγηση των κινδύνων με βάση τρεις βασικούς παράγοντες: το μέγεθος της απειλής, την ευπάθεια του συστήματος και την πιθανότητα να επηρεάσει τη λειτουργία. Ο στόχος είναι η απόκτηση μίας συνολικής εικόνας από τη συγκρίσιμη εκτίμηση των κινδύνων, ώστε να τεθούν σε προτεραιότητα τα πιο κρίσιμα προβλήματα ασφαλείας.

Από την ανάλυση της επικινδυνότητας πρόκυψε ότι οι απειλές Δηλητηρίαση Δεδομένων (Data Poisoning) – QualityAI_v2, αποκάλυψη δεδομένων στην έξοδο του μοντέλου (Data disclosure in model output) – QualityAI_v2 και παραβίαση ελέγχου πρόσβασης (Broken Access Control) – SteelData_v1, Backup_Storage_AWS έχουν τον υψηλότερο κίνδυνο (Risk = 80). Έτσι, οι απειλές αυτές είναι κρίσιμες για το σύστημα AI και πρέπει να έχουν υψηλότερη προτεραιότητα.