ΠΑΝΕΠΙΣΤΗΜΙΟ

ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ
ΣΧΟΛΗ

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Πως μπορεί η Τεχνητή Νοημοσύνη να υποστηρίξει την ανίχνευση και αντιμετώπιση επιθέσεων κοινωνικής μηχανικής;

(How can Artificial Intelligence support the detection and remediation of social engineering attacks?)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πιστοφίδου Νικολέτα

Επιβλέπων: Κρητικός Κυριάκος (Αναπληρωτής Καθηγητής)
Μέλη εξεταστικής επιτροπής: Καρύδα Μαρία (Καθηγήτρια), Κοκολάκης Σπύρος (Καθηγητής)

Σάμος, 2024

Περίληψη

Η παρούσα διπλωματική εργασία διερευνά πώς η Τεχνητή Νοημοσύνη (ΤΝ) μπορεί να αξιοποιηθεί για την ανίχνευση, αποτροπή και αντιμετώπιση επιθέσεων κοινωνικής μηχανικής. Ειδικότερα, σκοπός της έρευνας είναι η αναγνώριση και η αξιολόγηση τεχνικών, μεθόδων και αλγορίθμων βασισμένων στην Τεχνητή Νοημοσύνη, οι οποίοι μπορούν να εφαρμοστούν για την αποτελεσματική προστασία έναντι τέτοιου είδους επιθέσεων. Η συνεισφορά της εργασίας έγκειται στη συστηματική ανάλυση και συγκριτική αποτίμηση αυτών των μεθόδων, αναδεικνύοντας τις πλέον αποδοτικές και αποτελεσματικές προσεγγίσεις για την ανίχνευση και μετρίαση ποικίλων επιθέσεων κοινωνικής μηχανικής. Τα ευρήματα προσφέρουν κρίσιμες πληροφορίες για την ενίσχυση της ασφάλειας στον κυβερνοχώρο μέσω της Τεχνητής Νοημοσύνης, καθοδηγώντας τόσο την ακαδημαϊκή έρευνα όσο και την πρακτική εφαρμογή στο πεδίο της κυβερνοασφάλειας.

Λέξεις κλειδιά: Επιθέσεις κοινωνικής μηχανικής, τεχνητή νοημοσύνη, αλγόριθμοι, ανίχνευση, αποτροπή, αντιμετώπιση, κυβερνοασφάλεια.

Abstract

This thesis explores how Artificial Intelligence (AI) can be utilized for the detection, prevention, and mitigation of social engineering attacks. The aim of the research is to identify and evaluate AI-based techniques, methods, and algorithms that can be applied to effectively protect against such attacks. The contribution of this work lies in the systematic analysis and comparative evaluation of these methods, highlighting the most efficient and effective approaches for detecting and mitigating various types of social engineering attacks. The produced findings provide critical insights for enhancing cybersecurity through AI, guiding both academic research and practical application in the field of cybersecurity.

Keywords: Social engineering attacks, artificial intelligence, algorithms, detection, deterrence, response, cybersecurity.

ΠΕΡΙΕΧΟΜΕΝΑ

Λίστα Εικόνων

Λίστα Πινάκων

Ακρωνύμια:

Ακρωνύμιο	Σημασία Ακρωνυμίου
ΓΚΠΔ	Γενικός Κανονισμός για την Προστασία Δεδομένων
ΤΛ	Τεχνολογία Λογισμικού
ΤΝ	Τεχνητή Νοημοσύνη

Acronyms:

Acronym	Acronym Definition
ABAC	Attribute-Based Access Control
APT	Advanced Persistent Threat
BA & DL	Biometric Authentication and Deep Learning
C2/C&C	Command and Control
CNNS	Convolutional Neural Networks
DAISY	Dense Descriptor
DDoS	Distributed Denial of Service
DNS	Domain Name System
DQN	Deep Q-Networks
GANS	Generative Adversarial Networks
GBM	Gradient Boosting Machine
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Models
HOG	Histogram of Oriented Gradients

Acronym	Acronym Definition
IDS	Intrusion Detection Systems
KNN	K-Nearest Neighbors
MDP	Markov Decision Process
MFA	Multi-Factor Authentication
MITM	Man-in-the-Middle
ML	Machine Learning
NLP	Natural Language Processing
RNNs	Recurrent Neural Networks
SDLC	Secure Software Development Lifecycle
SIEM	Security Information and Event Management
SIFT	Scale-Invariant Feature Transform
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
UBA	User Behavior Analytics

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή

Σήμερα, είναι δύσκολο να εντοπιστεί μια εταιρεία, ένα ίδρυμα, μια οικογένεια ή ένα άτομο που να μην χρησιμοποιεί το διαδίκτυο και την τεχνολογία. Εμείς, ως άνθρωποι, βρισκόμαστε καταβεβλημένοι από τον αριθμό των ψηφιακών συσκευών και εφαρμογών που χρησιμοποιούμε καθημερινά. Πολλοί από εμάς δεν μπορούμε καν να ελέγξουμε τη χρήση της τεχνολογίας με τον τρόπο που θα επιθυμούσαμε. Κάποιοι μπορεί να είναι εθισμένοι στο διαδίκτυο και να μην μπορούν να σταματήσουν να το χρησιμοποιούν, ενώ άλλοι μπορεί να μην το χρησιμοποιούν αρκετά για να συμβαδίζουν με τις ραγδαίες αλλαγές στην τεχνολογία. Δεν υπάρχει αμφιβολία ότι η τεχνολογία έχει αυξήσει την παραγωγικότητα και την αποδοτικότητά μας με διάφορους τρόπους. Ωστόσο, πρέπει να λάβουμε υπόψη τις επιπτώσεις που έχει στην προσωπική και κοινωνική μας ζωή και στην ψυχική και σωματική μας ευεξία.

Παρά τη θετική πλευρά της διάδοσης των συσκευών που συνδέονται στο Διαδίκτυο, μια σκοτεινή πλευρά των εγκλημάτων στον κυβερνοχώρο έχει απλώσει τη σκιά της πάνω από τη ζωή μας. Αυτή η ξαφνική αύξηση της διασύνδεσης, ενώ μας έχει κάνει πιο αποτελεσματικούς, μας έχει καταστήσει ταυτόχρονα ευάλωτους. Το διαδίκτυο μας έχει κάνει ταυτόχρονα κατακτητές και φυλακισμένους (Chakraborty et al., 2023). Ωστόσο, αυτό έχει επίσης δημιουργήσει τρωτά σημεία που μπορούν να αξιοποιηθούν από τους εγκληματίες. Η κλοπή ταυτότητας και οι παραβιάσεις της ιδιωτικής ζωής γίνονται όλο και πιο συχνές, ειδικότερα καθώς οι χάκερ εκμεταλλεύονται τις πλατφόρμες ψηφιακών μέσων. Το έγκλημα στον κυβερνοχώρο είναι ένας όρος ομπρέλα που περιλαμβάνει όλες τις μορφές εγκλημάτων που σχετίζονται με τον κυβερνοχώρο - στην ουσία, παράνομες δραστηριότητες που ξεκινούν με τη χρήση υπολογιστών. Καθώς αποτελεί μια από τα πιο προσοδοφόρες δραστηριότητες της σύγχρονης εποχής, κάθε χρόνο, οι εγκληματίες του κυβερνοχώρου αποκομίζουν κέρδη δισεκατομμυρίων δολαρίων από την κλοπή δεδομένων, την αλλοίωση τους και την παραβίαση κρίσιμων υποδομών.

Η ενίσχυση της ασφάλειας στον κυβερνοχώρο έχει γίνει μια κρίσιμη πτυχή για κάθε υποδομή ώστε να μην εκτεθεί από κάθε είδους επίθεση που προέρχεται από εξωτερικές πηγές, όπως ιούς, κακόβουλο λογισμικό ή απόπειρες παραβίασης. Ωστόσο, οι περισσότερες παραβιάσεις της ασφάλειας ή εγκλήματα στον κυβερνοχώρο οφείλονται πλέον σε ανθρώπινο λάθος και όχι σε εξωτερική απειλή. Αυτό ισχύει με την λογική πως ορισμένα συστήματα είναι καλά προστατευμένα ή θα είναι αρκετά κοστοβόρο και χρονοβόρο να διεισδύσει κάποιος σε αυτά, ενώ η εκμετάλλευση του ανθρώπου είναι αρκετά πιο εύκολη και μπορεί να δημιουργήσει

εύκολα εκμεταλλεύσιμες τρωτότητες σε οποιοδήποτε σύστημα. Η εκμετάλλευση αυτή γίνεται στα πλαίσια ειδικών επιθέσεων που ονομάζονται επιθέσεις κοινωνικής μηχανικής.

Οργανωτικά & διοικητικά αντίμετρα, όπως η εκπαίδευση του προσωπικού και η εφαρμογή πολιτικών ασφαλείας, συμβάλλουν σημαντικά στην αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής. Η ευαισθητοποίηση και η καθιέρωση αυστηρών διαδικασιών για την προστασία των πληροφοριών μειώνουν την ανθρώπινη ευπάθεια μέσω προγραμμάτων εκπαίδευσης και προσομοιώσεων επιθέσεων. Ωστόσο, παρά την αξία αυτών των γενικών μέτρων, δεν είναι από μόνα τους αρκετά για την πλήρη αντιμετώπιση των απειλών της κοινωνικής μηχανικής. Αναλυτικότερα, η εκπαίδευση του προσωπικού δεν εξαλείφει τα ανθρώπινα σφάλματα ή την απροσεξία. Για παράδειγμα, οι εργαζόμενοι μπορεί να πέσουν θύματα επιθέσεων λόγω έλλειψης συγκέντρωσης. Παράλληλα, τα τεχνικά αντίμετρα που έχουν προταθεί για την άμεση αντιμετώπιση ή την ανίχνευση των επιθέσεων κοινωνικής μηχανικής, όπως τα φίλτρα προστασίας από phishing και οι ανιχνευτές ανωμαλιών, συχνά δεν διαθέτουν την απαραίτητη ακρίβεια, ενώ επικεντρώνονται κυρίως σε συγκεκριμένα είδη επιθέσεων, αφήνοντας πιθανή την έκθεση σε άλλα δυνατά σενάρια **[5].** Αυτό, λοιπόν, το κενό προστασίας δημιουργεί την ανάγκη για πιο εξελιγμένες και ολοκληρωμένες λύσεις.

Η Τεχνητή Νοημοσύνη (ΤΝ) έρχεται να καλύψει αυτό το κενό, προσφέροντας πλεονεκτήματα, όπως η δυνατότητα για δυναμική ανίχνευση και προσαρμογή σε νέες μορφές επιθέσεων. Η ΤΝ μπορεί να αναλύσει μεγάλες ποσότητες δεδομένων σε πραγματικό χρόνο, να αναγνωρίσει σύνθετα μοτίβα και να προβλέψει ενδεχόμενες επιθέσεις με μεγαλύτερη ακρίβεια. Η προσαρμοστικότητά της και η ικανότητά της να μαθαίνει από νέα δεδομένα την καθιστούν μια πολύτιμη προσθήκη για την αντιμετώπιση των εξελισσόμενων απειλών της κοινωνικής μηχανικής. Παρότι όμως η χρήση της ΤΝ στην κοινωνική μηχανική έχει διερευνηθεί σε σημαντικό βαθμό σήμερα και αποτελεί ισχυρό σύμμαχο έναντι τέτοιου είδους επιθέσεων, υπάρχουν ακόμα περιθώρια για βελτίωση και διερεύνηση νέων προσεγγίσεων.

Η παριστάμενη διπλωματική εργασία εξετάζει πώς η Τεχνητή Νοημοσύνη (ΤΝ) μπορεί να συμβάλλει στην αποτροπή, ανίχνευση και μετριασμό των επιθέσεων κοινωνικής μηχανικής, αξιολογώντας την αποτελεσματικότητα των αλγορίθμων ΤΝ κατά των επιθέσεων κοινωνικής μηχανικής. Μέσα από μια εμπεριστατωμένη επισκόπηση της υπάρχουσας βιβλιογραφίας, η εργασία θα προσδιορίσει ποιοι παράγοντες συμβάλλουν στην επιτυχία ή την αποτυχία των προσεγγίσεων ΤΝ που εφαρμόζονται σήμερα. Συγκρίνει διάφορες τεχνικές και αλγορίθμους βασισμένους στην ΤΝ, αξιολογώντας την αποτελεσματικότητά τους. Μέσα από την ανάλυση αυτών των περιπτώσεων, αναδεικνύονται οι παράγοντες επιτυχίας και αποτυχίας, παρέχοντας κατευθύνσεις για τις βέλτιστες πρακτικές και εντοπίζοντας προκλήσεις. Η εργασία τονίζει τις ανεκμετάλλευτες δυνατότητες της ΤΝ και την ανάγκη για περαιτέρω έρευνα στον τομέα αυτό. Τα παραγόμενα ευρήματα της εργασίας αυτής προσφέρουν κρίσιμες πληροφορίες για την ενίσχυση της ασφάλειας στον κυβερνοχώρο

μέσω της ΤΝ, καθοδηγώντας τόσο την ακαδημαϊκή έρευνα όσο και την πρακτική εφαρμογή στο πεδίο της κυβερνοασφάλειας.

1.2 Δομή Εργασίας

Η παρούσα εργασία χωρίζεται στα επόμενα 5 κεφάλαια, με την εξής δομή:

- *Κεφάλαιο 2, Θεωρητικό Υπόβαθρο*: Παρέχεται ένα πλήρες Θεωρητικό υπόβαθρο για την κυβερνοασφάλεια, τις επιθέσεις κοινωνικής μηχανικής και τον ρόλο της ΤΝ, με σκοπό την καλύτερη κατανόηση συνολικά της αναφοράς της διπλωματικής.

- *Κεφάλαιο 3, Μεθοδολογία Βιβλιογραφικής Επισκόπησης*: Εξετάζονται τα κύρια ερευνητικά ερωτήματα και τα σημεία στα οποία αυτά απαντώνται στο υπόλοιπο μέρος της αναφοράς καθώς και ο τρόπος στρατηγικής αναζήτησης και φιλτραρίσματος των εντοπισμένων άρθρων. Επιπλέον, παρέχεται μια πρώτη ποσοτική ανάλυση με τη μορφή στατιστικών βάσει των άρθρων που τελικά επιλέχθηκαν.

- *Κεφάλαιο 4, Ανάλυση*: Μελετώνται τα σχετικά άρθρα που τελικώς επιλέχθηκαν, παρέχεται μια ιεραρχική κατηγοριοποίησή τους και πραγματοποιείται μια σύγκρισή τους με βάση ένα σύνολο από σχετικά κριτήρια.

- *Κεφάλαιο 5, Προκλήσεις και Μελλοντική Έρευνα*: Εξετάζονται προκλήσεις και ερευνητικά κενά που αφορούν την αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής από τη ΤΝ καθώς και ερευνητικές κατευθύνσεις που μπορούν να τα αντιμετωπίσουν.

- *Κεφάλαιο 6, Συμπεράσματα*: Παρουσιάζονται τα κύρια συμπεράσματα που παρήχθησαν από την διπλωματική εργασία καθώς και αναφέρεται η προσωπική εμπειρία που αποκομίσθηκε από την εκπόνησή της.

Κεφάλαιο 2: Θεωρητικό Υπόβαθρο

Η κατανόηση των λεπτομερειών των επιθέσεων κοινωνικής μηχανικής και του ρόλου των αναδυόμενων τεχνολογιών είναι υψίστης σημασίας στο διαρκώς εξελισσόμενο τοπίο της κυβερνοασφάλειας.

Το δεύτερο αυτό κεφάλαιο εμβαθύνει στο θεωρητικό υπόβαθρο που είναι απαραίτητο για την κατανόηση των αποχρώσεων των απειλών κυβερνοασφάλειας, εστιάζοντας ιδιαίτερα στην κοινωνική μηχανική και την ιστορική της εξέλιξη. Επιπλέον, το κεφάλαιο αυτό εξετάζει τα κίνητρα, τους κινδύνους, τις μεθόδους και τις επιπτώσεις των επιθέσεων κοινωνικής μηχανικής, καθώς και παραδείγματα από τον πραγματικό κόσμο, ενώ παράλληλα παρέχει μια ολοκληρωμένη βάση για την κατανόηση του ανθρώπινου στοιχείου αλλά και ευρύτερων απειλών στον κυβερνοχώρο, καθώς και τα (κλασικά) αντίμετρα που μπορούν να αντιμετωπίσουν αυτές τις απειλές.

Τέλος, ορίζει και αναλύει τι είναι η ΤΝ και οι τεχνολογίες της με γνώμονα την πλήρη κατανόησή της αλλά και των προοπτικών εφαρμογής της στην ασφάλεια των συστημάτων και ειδικότερα στην ενίσχυση υπαρχόντων ή την ανάπτυξη νέων αντιμέτρων για επιθέσεις κοινωνικής μηχανικής.

2.1 Κυβερνοασφάλεια

2.1.1 Ορισμοί

Σε μια εποχή όπου η ψηφιακή συνδεσιμότητα διαπερνά κάθε πτυχή της κοινωνίας, η κυβερνοασφάλεια αποτελεί προπύργιο ενάντια σε μια σειρά απειλών που επιδιώκουν να εκμεταλλευτούν τα τρωτά σημεία των διασυνδεδεμένων συστημάτων. Στον πυρήνα της, η κυβερνοασφάλεια περιλαμβάνει την πρακτική της προστασίας συστημάτων υπολογιστών, δικτύων και δεδομένων από ψηφιακές επιθέσεις, κλοπή, βλάβη ή μη εξουσιοδοτημένη πρόσβαση, προκειμένου να διασφαλιστεί η ακεραιότητα, η εμπιστευτικότητα και η διαθεσιμότητα των πληροφοριών (Jansen & Grance, 2011). Ενώ το τοπίο της κυβερνοασφάλειας είναι τεράστιο και πολύπλευρο, η θεμελιώδης κατανόηση των βασικών εννοιών, απειλών, τρωτών σημείων και αντιμέτρων είναι απαραίτητη για την ενίσχυση της άμυνάς μας έναντι των κακόβουλων φορέων. Προς αυτό τον σκοπό, παρουσιάζουμε τους ορισμούς για τις τρεις πιο βασικές έννοιες της κυβερνοασφάλειας:

1. *Απειλές*: Κάθε πιθανή αιτία που μπορεί να διαταράξει την κανονική λειτουργία ενός πληροφοριακού συστήματος ή να προκαλέσει ζημιά σε ένα αγαθό ή οργανισμό (Stallings, W., & Brown, L. (2012). Οι απειλές οφείλονται σε εξωτερικούς ή εσωτερικούς παράγοντες. Οι

πρώτοι αφορούν χάκερς ή κακόβουλα λογισμικά (πχ. πράκτορες λογισμικού) που προχωρούν σε εσκεμμένες επιθέσεις, ενώ οι δεύτεροι αφορούν άτομα, λογισμικό ή υλικά εντός ενός οργανισμού που μπορεί συνήθως κατά λάθος να προκαλέσουν κάποια ζημία (αν και τα άτομα εντός ενός οργανισμού μπορεί να προκαλέσουν εσκεμμένες ζημίες σε ορισμένες περιπτώσεις). Επομένως, οι απειλές μπορεί να είναι σκόπιμες ή ακούσιες. Επιπλέον, διαχωρίζονται σε φυσικές (πχ. φωτιές), τεχνικές (βλάβες υλικού ή σφάλματα λογισμικού) & ανθρώπινες (πλαστοπροσωπία, μη εξουσιοδοτημένη πρόσβαση).

2. *Ευπάθειες*: Αδυναμίες σε αγαθά (πχ. υλικό, λογισμικό, άτομα, δεδομένα), ομάδες αγαθών ή μέτρων ασφαλείας σε ένα σύστημα που μπορούν να εκμεταλλευτούν οι κακόβουλοι παράγοντες. Αυτές μπορεί να περιλαμβάνουν κενά ασφαλείας σε λογισμικά, αδύναμους κωδικούς πρόσβασης ή μη σωστά ενημερωμένα συστήματα. Η εκμετάλλευση ευπάθειας γίνεται από επιθέσεις που πραγματοποιούν απειλές και συχνά οδηγεί σε παραβίαση ασφάλειας (Anderson, R. (2020).

3. *Αντίμετρα*: Αποτελούν τεχνικά, διοικητικά, οργανωτικά ή και νομικά μέσα διαχείρισης ενός κινδύνου. Μπορεί να περιλαμβάνουν μηχανισμούς ή διαδικασίες που αποσκοπούν στον περιορισμό ή την εξάλειψη των επιπτώσεων μιας ή περισσοτέρων απειλών. Εμπεριέχουν τις προληπτικές στρατηγικές, τις τεχνολογίες και τις βέλτιστες πρακτικές που χρησιμοποιούνται για τον μετριασμό των απειλών στον κυβερνοχώρο, την μείωση των τρωτοτήτων καθώς και την άμυνα κατά των επιθέσεων στον κυβερνοχώρο. Στα αντίμετρα περιλαμβάνονται πρωτόκολλα ασφάλειας δικτύου, αλγόριθμοι κρυπτογράφησης, έλεγχοι πρόσβασης, μέθοδοι αυθεντικοποίησης, συστήματα ανίχνευσης εισβολών, σχέδια αντιμετώπισης περιστατικών, τείχη προστασίας, συστήματα προστασίας τελικών σημείων, συστήματα προστασίας δεδομένων και προγράμματα εκπαίδευσης των εργαζομένων (Schwartau, 2018).

4. *Επίθεση:* Μια επίθεση είναι μια σκόπιμη προσπάθεια από κάποιον να θέσει σε κίνδυνο ή να εκμεταλλευτεί μια αδυναμία ενός (πληροφοριακού) συστήματος ή των αγαθών που αυτό εμπεριέχει. Κύρια σκοπιμότητα των επιθέσεων που προσπαθούν να αποκτήσουν μη εξουσιοδοτημένη πρόσβαση σε ένα σύστημα αποτελεί η πρόκληση βλαβών ή η κλοπή δεδομένων (Stallings, 2012).

5. *Κίνδυνος / Ρίσκο:* Το ρίσκο αναφέρεται στο γινόμενο της πιθανότητας να συμβεί μια επίθεση και των συνεπειών που θα ακολουθήσουν εφόσον αυτή όντως συμβεί. Ο κίνδυνος αξιολογείται με βάση το επίπεδο τρωτότητας ενός συστήματος και τις πιθανές επιπτώσεις από την εκμετάλλευσή του (Stallings, 2012).

2.1.2 Ανθρώπινος Παράγοντας

Ο ανθρώπινος παράγοντας αποτελεί μια κρίσιμη αλλά συχνά παραγνωρισμένη πτυχή της ασφάλειας στον κυβερνοχώρο. Παρά τις εξελίξεις στην τεχνολογία και τις αυτοματοποιημένες λύσεις ασφάλειας, η ανθρώπινη συμπεριφορά παραμένει ο πρωταρχικός

καθοριστικός παράγοντας της αποτελεσματικότητας της ασφάλειας στον κυβερνοχώρο (Schwartau, 2018).

Μέτρα για την "διόρθωση" της ανθρώπινης συμπεριφοράς προς ενίσχυση της κυβερνοασφάλειας περιλαμβάνουν:

- *Υγιεινή στον κυβερνοχώρο*: Η υιοθέτηση ορθών πρακτικών κυβερνοϋγιεινής (Cyber hygiene), όπως η τακτική ενημέρωση του λογισμικού, η χρήση ισχυρών κωδικών πρόσβασης και η προσοχή όταν γίνεται πρόσβαση σε συνδέσμους, είναι απαραίτητη για τον μετριασμό των κινδύνων στον κυβερνοχώρο (Mitnick, 2017).
- *Ευαισθητοποίηση των εργαζομένων*: Οι εργαζόμενοι αποτελούν συχνά την πρώτη γραμμή άμυνας έναντι των απειλών στον κυβερνοχώρο. Η εκπαίδευση των εργαζομένων σχετικά με τους κοινούς κινδύνους στον κυβερνοχώρο, τις τακτικές κοινωνικής μηχανικής και τη σημασία των βέλτιστων πρακτικών κυβερνοασφάλειας μπορεί να συμβάλει στην προώθηση μιας κουλτούρας ασφάλειας εντός των οργανισμών.
- *Ψυχολογικοι παράγοντες*: Το άγχος, η βιασύνη, και η κόπωση μπορούν να αυξήσουν την πιθανότητα λάθους και να εκθέσουν τον οργανισμό σε κίνδυνο. Η καλλιέργεια περιβάλλοντος που μειώνει αυτούς τους παράγοντες μπορεί να βελτιώσει την κυβερνοασφάλεια.

2.1.3 Τύποι Κυβερνοεπιθέσεων

Οι κυβερνοεπιθέσεις περιλαμβάνουν κακόβουλους παράγοντες που εκμεταλλεύονται τρωτά σημεία και προκαλούν ζημιά στο σύστημα. Καθοριστικός παράγοντας για την επιτυχία των κυβερνοεπιθέσεων είναι, όπως αναλύθηκε προηγουμένως, η ανθρώπινη συμπεριφορά.

Παρακάτω παραθέτουμε μια κατηγοριοποίηση διαφόρων τύπων επιθέσεων κυβερνοασφάλειας, η οποία απεικονίζεται στην Εικόνα 1:

1. *Επιθέσεις κακόβουλου λογισμικού (Malware attacks)*: Το κακόβουλο λογισμικό, συμπεριλαμβανομένων των ιών, των σκουληκιών, των Trojans (Δούρειων Ίππων) και του ransomware (λυτρισμικού), αποτελεί σημαντική απειλή για την ασφάλεια στον κυβερνοχώρο, η οποία έχει σχεδιαστεί για να προκαλέσει βλάβη ή να υποκλέψει, τροποποιήσει ή διαγράψει δεδομένα από μια συσκευή. Το κακόβουλο λογισμικό συνήθως εκχύεται σε ένα σύστημα με διάφορους εναλλακτικούς τρόπους (*πχ*. κακόβουλα συνημμένα σε μηνύματα ηλεκτρονικής αλληλογραφίας, επιθέσεις drive-by download κα.) ώστε να επιτελέσει τους κακόβουλους σκοπούς του.

2. *Επιθέσεις κοινωνικής μηχανικής (Social engineering attacks)*: Θα αναλυθούν διεξοδικά στην ενότητα 2.2.6 αυτού του κεφαλαίου. Επιγραμματικά, οι επιθέσεις αυτές

περιλαμβάνουν τεχνικές εξαπάτησης με στόχο την απόσπαση εμπιστευτικών πληροφοριών ή την παραβίαση συστημάτων ασφαλείας, μέσω της χειραγώγησης ατόμων.

3. *Επιθέσεις (κατανεμημένης) άρνησης παροχής υπηρεσιών ((D)DoS attacks)*: Οι επιτιθέμενοι κατακλύζουν έναν διακομιστή ή δίκτυο με υπερβολική κίνηση αιτήσεων ή δεδομένων, καθιστώντας τον μη διαθέσιμο σε νόμιμους χρήστες, διαταράσσοντας έτσι τις προσφερόμενες υπηρεσίες. Η απλή επίθεση DoS πραγματοποιείται από μια μόνο πηγή ενώ η κατανεμημένη επίθεση DoS πραγματοποιείται από ένα σύνολο από πηγές που είναι ελεγχόμενο από τον επιτιθέμενο (όπως είναι ένα botnet).

4. *Επιθέσεις προηγμένων επίμονων απειλών (APT attacks)*: Αποτελούν ιδιαίτερα εξελιγμένες επιθέσεις στον κυβερνοχώρο που συνήθως περιλαμβάνουν τη μακροχρόνια, αθόρυβη εκμετάλλευση ευπαθειών, με κύριους στόχους κυβερνήσεις, εταιρείες και κρίσιμες υποδομές. Η ανίχνευση και αντιμετώπισή τους απαιτεί τη χρήση προηγμένων τεχνικών, όπως τεχνικές μηχανικής μάθησης.

5. *Επιθέσεις εκμετάλλευσης ευπαθειών μηδενικής ημέρας (Zero-Day Exploits attacks)*: Οι επιτιθέμενοι εντοπίζουν και εκμεταλλεύονται ευπάθειες λογισμικού ή συστημάτων που είναι άγνωστες στους κατασκευαστές τους ή στο κοινό κατά τη στιγμή της επίθεσης. Ο όρος "Zero-Day" αναφέρεται στο γεγονός ότι ο οργανισμός ή η εταιρεία έχει μηδενικές ημέρες για να διορθώσει την ευπάθεια, αφού αυτή γίνεται εκμεταλλεύσιμη πριν καν υπάρξει διαθέσιμη ενημέρωση ασφαλείας ή επίγνωση του προβλήματος.

6. *Επιθέσεις ενδιάμεσου (Man-in-the-Middle (MiTM) attacks)*: Αυτές οι επιθέσεις εμπίπτουν στην κατηγορία των δικτυακών επιθέσεων, καθώς οι επιτιθέμενοι παρεμβάλλονται στη μετάδοση δεδομένων μεταξύ δύο μερών, παρακολουθώντας ή τροποποιώντας τα δεδομένα που ανταλλάσσονται χωρίς να το γνωρίζουν τα εμπλεκόμενα μέρη. Σε ορισμένες περιπτώσεις, οι επιτιθέμενοι μπορεί να αναπαριστούν ένα από τα δύο μέρη της επικοινωνίας (οπότε έχουμε ένα είδος πλαστοπροσωπίας) για να εξαπατήσουν το άλλο ώστε να επικοινωνήσει μαζί τους (και όχι με το πραγματικό).

7. *Επιθέσεις Έκχυσης (Injection Attacks)*: Οι επιθέσεις αυτές περιλαμβάνουν διάφορες μεθόδους εκμετάλλευσης ευπαθειών σε εφαρμογές ιστού (web applications) μέσω της εισαγωγής κακόβουλου κώδικα ή δεδομένων σε πεδία εισόδου (πχ. φορμών) που δεν έχουν κατάλληλη προστασία ή δεν ελέγχονται καθόλου (από τις εφαρμογές αυτές). Οι πιο κοινές μορφές επιθέσεων έκχυσης είναι:

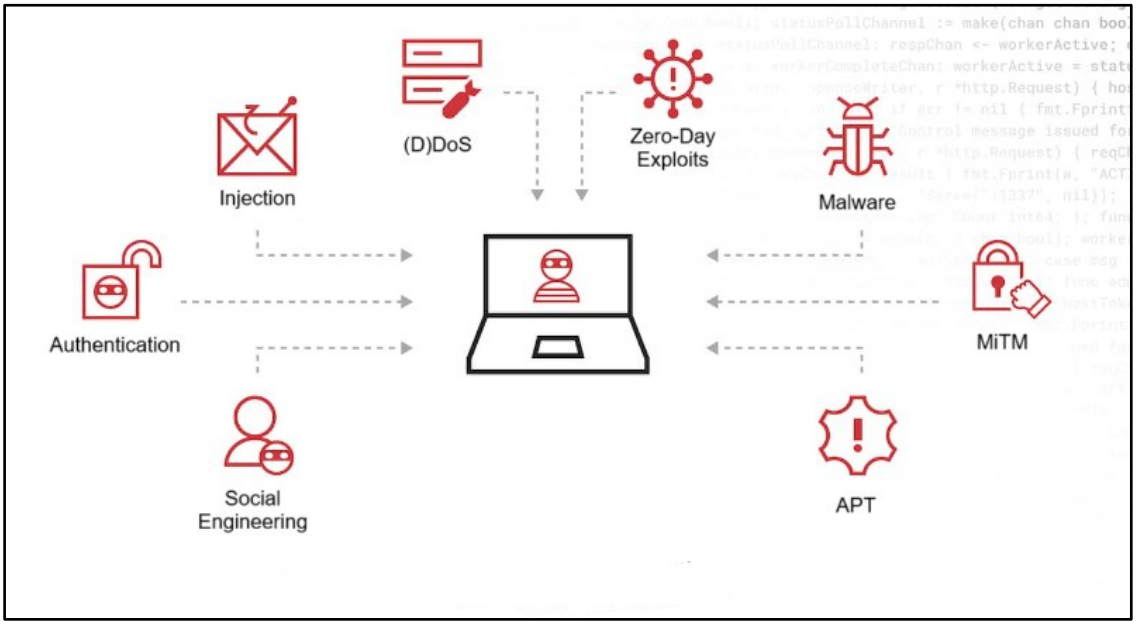
- *SQL Injection*: Οι αναφερόμενες επιθέσεις επιτρέπουν στους επιτιθέμενους να εισάγουν κακόβουλο κώδικα SQL σε πεδία εισόδου της εφαρμογής. Στόχος τους είναι να παρακάμψουν ή να συμπληρώσουν τα κανονικά ερωτήματα βάσης δεδομένων που υποβάλλει η εφαρμογή σε μια υποκείμενη βάση δεδομένων ώστε να αποκτήσουν πρόσβαση σε ευαίσθητα δεδομένα (πχ. διαπιστευτήρια χρηστών) ή να αλλοιώσουν το

περιεχόμενο της βάσης δεδομένων (πχ. διαγραφή χρηστών ή δεδομένων χρηστών της εφαρμογής).

- *Cross-Site Scripting (XSS)*: Οι συγκεκριμένες επιθέσεις αξιοποιούν τρωτά σημεία σε εφαρμογές που δεν καθαρίζουν σωστά τα εισαγόμενα δεδομένα, επιτρέποντας την εισαγωγή κακόβουλου κώδικα (συνήθως JavaScript) σε ιστοσελίδες που επισκέπτονται άλλοι χρήστες. Κατά την επίσκεψη του θύματος στις μολυσμένες ιστοσελίδες, εκτελείται ο κακόβουλος κώδικας και οι επιτιθέμενοι αποκτούν με αυτό τον τρόπο μη εξουσιοδοτημένη πρόσβαση σε δεδομένα συνεδριών (sessions) ή cookies ή μπορούν να εκτελέσουν ενέργειες εξ ονόματος του θύματος αυτού.
- *Command Injection (Έκχυσης Εντολών)*: Η επικαλούμενη επίθεση επιτρέπει στους δράστες να εισάγουν και να εκτελέσουν αυθαίρετες εντολές συστήματος σε εφαρμογές που δεν φιλτράρουν σωστά τα δεδομένα εισόδου. Αυτές οι επιθέσεις στοχεύουν κυρίως εφαρμογές που περνούν απευθείας δεδομένα εισόδου σε λειτουργικά συστήματα ή διακομιστές, επιτρέποντας στους επιτιθέμενους να εκτελέσουν οποιαδήποτε εντολή στο υποκείμενο σύστημα ή σε έναν διασυνδεδεμένο διακομιστή.

8. *Επιθέσεις αυθεντικοποίησης (Authentication Attacks)*: Αυτού του είδους οι επιθέσεις στοχεύουν στην παραβίαση των μηχανισμών αυθεντικοποίησης των χρηστών, επιτρέποντας στους επιτιθέμενους να αποκτήσουν μη εξουσιοδοτημένη πρόσβαση σε λογαριασμούς (χρηστών) και συστήματα. Οι πιο κοινές μορφές επιθέσεων αυθεντικοποίησης είναι:

- *Brute Force*: Κατά την επίθεση ωμής βίας, οι επιτιθέμενοι δοκιμάζουν συστηματικά όλες τις πιθανές παραλλαγές διαπιστευτηρίων και ιδιαίτερα κωδικών μέχρι να βρεθεί η σωστή (παραλλαγή), εκμεταλλευόμενοι συχνά αδύναμους ή κοινούς κωδικούς πρόσβασης. Με αυτό τον τρόπο, κατορθώνουν να έχουν μη εξουσιοδοτημένη πρόσβαση στο λογαριασμό του χρήστη-θύματος σε ένα σύστημα ή μια υπηρεσία.
- *Credential Stuffing*: Η επίθεση αυτή εκμεταλλεύεται την επαναχρησιμοποίηση των ίδιων ονομάτων χρήστη και κωδικών πρόσβασης από τους χρήστες σε πολλές πλατφόρμες. Οι επιτιθέμενοι χρησιμοποιούν λίστες κλεμμένων διαπιστευτηρίων από παραβιάσεις δεδομένων για να προσπαθήσουν να αποκτήσουν πρόσβαση σε άλλες υπηρεσίες και λογαριασμούς του ίδιου χρήστη (δηλ. του χρήστη που έχουν διαρρεύσει / κλαπεί τα διαπιστευτήριά του).



Εικόνα 1:Οπτική αναπαράσταση των 8 κύριων τύπων επιθέσεων κυβερνοασφάλειας Εργαλείο: DALL·E (AI-generated)

2.1.4 Αντίμετρα Κυβερνοασφάλειας

Μετά την εξέταση των κύριων τύπων επιθέσεων που απειλούν την ασφάλεια στον κυβερνοχώρο, είναι απαραίτητο να διερευνηθούν τα κατάλληλα αντίμετρα που μπορούν να εφαρμοστούν για την προστασία συστημάτων και δεδομένων. Κάθε τύπος επίθεσης απαιτεί εξειδικευμένες στρατηγικές και εργαλεία για τον έγκαιρο εντοπισμό των κινδύνων και την πρόληψη ή την αντιμετώπιση/μετριασμών των ζημιών. Ακολούθως θα παρουσιαστούν τα κατάλληλα αμυντικά μέτρα και οι τεχνικές λύσεις που μπορούν να ληφθούν για την αποτελεσματική αντιμετώπιση των επιθέσεων που αναλύθηκαν προηγουμένως. Αυτή η ολιστική προσέγγιση μεταξύ επιθέσεων και αντιμέτρων είναι κρίσιμη για την ενίσχυση της ασφάλειας στον κυβερνοχώρο.

Τα βασικά αντίμετρα και οι βέλτιστες πρακτικές κυβερνοασφάλειας περιλαμβάνουν:

- *Έλεγχος πρόσβασης με βάση ρόλους (RBAC) ή χαρακτηριστικά (ABAC)*: Ο RBAC (Role-Based Access Control) περιορίζει την πρόσβαση σε συστήματα και δεδομένα με βάση τους ρόλους των χρηστών εντός του οργανισμού. Από την άλλη μεριά, ο ABAC (Attribute-Based Access Control) χρησιμοποιεί πιο ευέλικτες παραμέτρους, όπως χαρακτηριστικά του χρήστη, του περιβάλλοντος ή των δεδομένων για τον καθορισμό των δικαιωμάτων πρόσβασης. Αυτά τα σχήματα/μοντέλα ελέγχου πρόσβασης διασφαλίζουν ότι μόνο εξουσιοδοτημένοι χρήστες μπορούν να προσπελάσουν κρίσιμες πληροφορίες, μειώνοντας έτσι τον κίνδυνο μη εξουσιοδοτημένης πρόσβασης.

- *Τμηματοποίηση δικτύου (Network Segmentation)*: Η τμηματοποίηση δικτύου περιλαμβάνει το διαχωρισμό ενός δικτύου σε μικρότερα ανεξάρτητα τμήματα ή ζώνες, ώστε να περιοριστεί η εξάπλωση μιας επίθεσης. Με αυτόν τον τρόπο, ακόμα και αν ένας επιτιθέμενος αποκτήσει πρόσβαση σε ένα τμήμα του δικτύου, δεν θα μπορεί να μεταφερθεί εύκολα σε άλλα κρίσιμα τμήματά του.

- *Αντίγραφα ασφαλείας και ανάκαμψη (Backup & Recovery)*: Τα αντίγραφα ασφαλείας (backups) διασφαλίζουν ότι τα δεδομένα μπορούν να ανακτηθούν σε περίπτωση απώλειας, καταστροφής ή επίθεσης (πχ., ransomware). Το πλάνο αποκατάστασης (recovery plan) περιλαμβάνει διαδικασίες για την επαναφορά των δεδομένων και των συστημάτων σε πλήρη λειτουργία. Είναι κρίσιμο να πραγματοποιούνται τακτικά αντίγραφα ασφαλείας και να δοκιμάζονται τα πλάνα αποκατάστασης, για να εξασφαλίζεται η ταχεία και αξιόπιστη ανάκαμψη από καταστροφικά συμβάντα.

- *Διαχείριση πληροφοριών και συμβάντων ασφαλείας (Security Information and Event Management (SIEM))*: Τα συστήματα SIEM συγκεντρώνουν και αναλύουν δεδομένα ασφαλείας από διάφορα μέρη του δικτύου ενός οργανισμού σε πραγματικό χρόνο, προκειμένου να εντοπίζουν και να αποτρέπουν επιθέσεις. Παρέχουν κεντρική ορατότητα των συμβάντων ασφαλείας και επιτρέπουν την ταχύτερη αντίδραση σε παραβιάσεις ή ανωμαλίες.

- *Ασφαλής κύκλος ζωής ανάπτυξης λογισμικού (Secure Software Development Lifecycle (SDLC))*: Ο SDLC περιλαμβάνει την ενσωμάτωση διαδικασιών & ελέγχων ασφαλείας σε όλα τα στάδια του κύκλου ανάπτυξης λογισμικού, από το σχεδιασμό έως την υλοποίηση και την διάθεση/διάταξη του λογισμικού στα διάφορα περιβάλλοντα (πχ. παραγωγής). Η ασφαλής ανάπτυξη λογισμικού μειώνει τον κίνδυνο ύπαρξης ευπαθειών στο λογισμικό που αναπτύσσεται που θα μπορούσαν να εκμεταλλευτούν οι επιτιθέμενοι.

- *Ευφυΐα απειλών (Threat Intelligence)*: Η ανάλυση πληροφοριών για απειλές (threat intelligence) περιλαμβάνει τη συλλογή και αξιολόγηση δεδομένων σχετικά με υπάρχουσες ή νέες απειλές στον κυβερνοχώρο, όπως κακόβουλο λογισμικό ή τακτικές επιθέσεων. Αυτές οι πληροφορίες βοηθούν τους οργανισμούς να προετοιμάζονται καλύτερα και να προσαρμόζουν τα αντίμετρα τους σύμφωνα με τη τρέχουσα κατάσταση (ειδικά για τρέχουσες απειλές που υλοποιούνται μέσω επιθέσεων) και τις τελευταίες εξελίξεις (ειδικά για τις νέες απειλές).

- *Παγίδες (Honeypots)*: Οι παγίδες είναι σκόπιμα τρωτά, απομονωμένα μέρη ενός συστήματος που χρησιμοποιούνται για να προσελκύσουν επιτιθέμενους και να συλλέξουν πληροφορίες για τις τεχνικές τους. Παρέχουν χρήσιμα δεδομένα για τη βελτίωση της άμυνας ενός συστήματος και βοηθούν στον εντοπισμό των επιτιθέμενων.

- *Φυσική ασφάλεια (Physical Security)*: Η φυσική ασφάλεια αφορά την προστασία των υποδομών από φυσικές απειλές, όπως η πρόσβαση μη εξουσιοδοτημένων ατόμων σε κρίσιμα συστήματα, κτίρια ή εξοπλισμό καθώς την προστασία από φυσικές καταστροφές. Περιλαμβάνει συστήματα ελέγχου πρόσβασης σε κτίρια, κάμερες παρακολούθησης, καθώς και συστήματα και διαδικασίες για την αποτροπή φυσικών καταστροφών (πχ. συστήματα πυρόσβεσης).

- *Διαχείριση ενημερώσεων και διορθώσεων (Patch Management)*: Η τακτική ενημέρωση του λογισμικού και των συστημάτων με τις τελευταίες διορθώσεις ασφαλείας ή με νέες εκδόσεις όπου έχουν επιδιορθωθεί γνωστές ευπάθειες συμβάλλει στον μετριασμό των τρωτών σημείων και στη μείωση του κινδύνου εκμετάλλευσης τους από επιτιθέμενους στον κυβερνοχώρο (Mitnick et al., 2017).

- *Πολυπαραγοντική αυθεντικοποίηση (MFA)*: Ενισχύει την ασφάλεια πρόσβασης προσθέτοντας επιπλέον επίπεδα αυθεντικοποίησης, όπως βιομετρική επαλήθευση ή κωδικούς μιας χρήσης. Μειώνει σημαντικά την πιθανότητα μη εξουσιοδοτημένης πρόσβασης, ακόμη και αν κλασικά διαπιστευτήρια, όπως κωδικοί, παραβιαστούν.

- *Κρυπτογράφηση δεδομένων (Data Encryption)*: Η εφαρμογή κρυπτογράφησης διασφαλίζει ότι, ακόμη και αν τα δεδομένα υποκλαπούν κατά τη μεταφορά ή την αποθήκευσή τους, παραμένουν ακατάληπτα χωρίς τα κατάλληλα κλειδιά αποκρυπτογράφησης.

- *Ανίχνευση και αποτροπή κακόβουλου λογισμικού (Anti-virus/Anti-malware)*: Η χρήση εξειδικευμένου λογισμικού που ανιχνεύει, αποτρέπει και αφαιρεί κακόβουλο λογισμικό διασφαλίζει ότι οι συσκευές και τα μηχανήματα είναι προστατευμένες/α από ιούς, δούρειους ίππους (trojans) και λυτρισμικό (ransomware).

- *Έτοιμες δηλώσεις και έλεγχος εγκυρότητας εισόδου (Prepared statements and input validation)*: Η χρήση προετοιμασμένων δηλώσεων (prepared statements) στις βάσεις δεδομένων αποτρέπει SQL Injection επιθέσεις, ενώ η σωστή επικύρωση των δεδομένων εισόδου προστατεύει τις εφαρμογές από επιθέσεις, όπως έκχυσης XSS και εντολών (Command Injection).

- *Συστήματα ανίχνευσης και αποτροπής εισβολών (IDS/IPS)*: Αυτά τα συστήματα ανιχνεύουν ύποπτες ενέργειες και προσπάθειες παραβίασης σε πραγματικό χρόνο, επιτρέποντας την έγκαιρη αντιμετώπιση πιθανών εισβολών στο δίκτυο. Τα συστήματα ανίχνευσης (Intrusion Detection Systems - IDS) εστιάζουν κυρίως στην ανίχνευση και ειδοποιούν τους διαχειριστές του συστήματος για μια ενδεχόμενη εισβολή. Από την άλλη μεριά, τα συστήματα αποτροπής εισβολών (Intrusion Prevention Systems - IPS)

όχι μόνο ανιχνεύουν αλλά και αποτρέπουν τις απειλές, *πχ*. παραπέμποντας πακέτα που εντοπίζονται στην πηγή της απειλής.

- *Καταγραφή και παρακολούθηση δραστηριότητας (Logging and Monitoring)*: Τα συστήματα παρακολούθησης καταγράφουν και αναλύουν τη δραστηριότητα δικτύου και συστημάτων για την ανίχνευση ύποπτων κινήσεων και ανωμαλιών καθώς και τον εντοπισμό προβλημάτων σε πολιτικές ασφάλειας ή κανόνες πρόσβασης.

- *Προστασία από επιθέσεις DDoS*: Εξειδικευμένα συστήματα προστασίας εντοπίζουν και περιορίζουν τον αντίκτυπο επιθέσεων κατανεμημένης άρνησης παροχής υπηρεσιών (DDoS), διασφαλίζοντας τη συνεχή λειτουργία των υπηρεσιών (ενός συστήματος).

- *Εκπαίδευση ευαισθητοποίησης σε θέματα ασφάλειας (Security awareness training)*: Η εκπαίδευση των εργαζομένων σχετικά με τις απειλές στον κυβερνοχώρο, τις τακτικές κοινωνικής μηχανικής και τις βέλτιστες πρακτικές κυβερνοασφάλειας είναι υψίστης σημασίας για τον μετριασμό των κινδύνων που σχετίζονται με την ανθρώπινη συμπεριφορά - αυτό το θέμα θα καλυφθεί διεξοδικά στην ενότητα 2.2.8 του κεφαλαίου.

Εν κατακλείδι, η κυβερνοασφάλεια αντιπροσωπεύει έναν πολύπλευρο κλάδο που αποσκοπεί στην προστασία των ψηφιακών περιουσιακών στοιχείων, τη διαφύλαξη της ιδιωτικής ζωής και τον μετριασμό των κινδύνων στον κυβερνοχώρο. Κεντρικό ρόλο στην αποτελεσματικότητα της ασφάλειας στον κυβερνοχώρο διαδραματίζουν η αναγνώριση των ανθρώπινων παραγόντων, η κατανόηση των κοινών απειλών και τρωτών σημείων στον κυβερνοχώρο και η εφαρμογή προληπτικών αντιμέτρων και βέλτιστων πρακτικών. Καλλιεργώντας μια κουλτούρα ασφάλειας, αγκαλιάζοντας τις τεχνολογικές καινοτομίες και δίνοντας προτεραιότητα στην ευαισθητοποίηση και την εκπαίδευση στον τομέα της κυβερνοασφάλειας, οι οργανισμοί μπορούν να ενισχύσουν την ανθεκτικότητά τους έναντι των εξελισσόμενων απειλών στον κυβερνοχώρο και να διαφυλάξουν την ακεραιότητα των ψηφιακών τους περιβαλλόντων (Mitnick et al., 2017).

2.2 Κοινωνική Μηχανική

2.2.1 Ορισμός Κοινωνικής Μηχανικής

Σύμφωνα με έναν ορισμό, κοινωνική μηχανική σημαίνει "κάθε πράξη που επηρεάζει ένα άτομο να προβεί σε μια ενέργεια που μπορεί να είναι ή να μην είναι προς το συμφέρον

του" (Mitnick, 2002). Αυτό λίγο πολύ περιλαμβάνει ολόκληρο τον όρο σε μια πρόταση, αλλά για να κατανοήσουμε πραγματικά το νόημά του πρέπει να προχωρήσουμε σε μια περισσότερη εξερεύνηση.

Έρευνες που πραγματοποιήθηκαν το 2020 έδειξαν ότι η κοινωνική μηχανική θα είναι μία από τις σημαντικότερες προκλήσεις της επερχόμενης δεκαετίας. Σύμφωνα με το αγγλικό λεξικό της Οξφόρδης, ο όρος "κοινωνική μηχανική" έχει δύο διαφορετικές έννοιες (social engineer, social engineering, n. 2017). Πρώτον, είναι "η χρήση κεντρικού σχεδιασμού σε μια προσπάθεια διαχείρισης της κοινωνικής αλλαγής και ρύθμισης της μελλοντικής ανάπτυξης και συμπεριφοράς μιας κοινωνίας". Δεύτερον, είναι "η χρήση εξαπάτησης προκειμένου να παρακινηθεί ένα άτομο να αποκαλύψει ιδιωτικές πληροφορίες ή ιδίως να παράσχει εν αγνοία του μη εξουσιοδοτημένη πρόσβαση σε ένα σύστημα ή δίκτυο υπολογιστών". Ενώ και οι δύο ορισμοί αφορούν ένα ή περισσότερα άτομα που παρακινούν τη συμπεριφορά άλλων, ο πρώτος βρίσκει ρητά την εφαρμογή του στον τομέα της πολιτικής και οικονομικής διαχείρισης, ενώ ο δεύτερος βρίσκει το σπίτι του μοναδικά στον τομέα του κυβερνοχώρου και ειδικότερα της κυβερνοασφάλειας.

Η κοινωνική μηχανική, σύμφωνα με τους Hutchins, Cloppert και Amin (2011), αναφέρεται σε επιθέσεις που στοχεύουν την ανθρώπινη αλληλεπίδραση, προκειμένου να χειραγωγήσουν τα θύματα ώστε να παραδώσουν κρίσιμες πληροφορίες ή να ανοίξουν την πόρτα σε πιο πολύπλοκες επιθέσεις στον κυβερνοχώρο. Οι επιθέσεις κοινωνικής μηχανικής θεωρούνται ένα σημαντικό προκαταρκτικό βήμα για πιο εξελιγμένες και καταστροφικές κυβερνοεπιθέσεις. Μάλιστα, οι επιθέσεις αυτές αυξάνονται σε ένταση και αριθμό, εδραιώνοντας την ανάγκη για νέες τεχνικές ανίχνευσης και εκπαιδευτικά προγράμματα για την ασφάλεια στον κυβερνοχώρο (Salahdine, 2019).

2.2.2 Ιστορική Αναδρομή Κοινωνικής Μηχανικής

Η ιστορία της κοινωνικής μηχανικής (social engineering) είναι στενά συνδεδεμένη με την ψυχολογική χειραγώγηση και την εξέλιξη της τεχνολογίας. Ιστορικά, μπορεί να εντοπιστεί σε διάφορους αιώνες. Συνεπώς, δεν αποτελεί ένα καινούργιο φαινόμενο. Αν και σήμερα συνδέεται κυρίως με την κυβερνοασφάλεια, η βασική αρχή της εξαπάτησης μέσω της εκμετάλλευσης της ανθρώπινης συμπεριφοράς έχει χρησιμοποιηθεί για χιλιάδες χρόνια. Μερικά χαρακτηριστικά ιστορικά παραδείγματα είναι τα εξής:

1. ***Δούρειος Ίππος (1184 π.Χ.)***: Ο θρύλος του Δούρειου Ίππου, όπου οι Έλληνες εξαπάτησαν τους Τρώες με ένα ξύλινο άλογο, αποτελεί μία από τις πρώτες καταγεγραμμένες περιπτώσεις κοινωνικής μηχανικής. Οι Τρώες πίστεψαν ότι είχαν

κερδίσει τον πόλεμο, αλλά οι Έλληνες τους εξαπάτησαν κρυμμένοι μέσα στο άλογο, το οποίο οι Τρώες έφεραν εντός της πόλης τους (Mitnick, 2011).

2. ***Αρχαίοι χειρισμοί και τακτικές εξαπάτησης***: Με την πάροδο των αιώνων, η χειραγώγηση της εμπιστοσύνης και των συναισθημάτων των ανθρώπων συνέχισε να είναι σημαντικό μέρος πολεμικών και πολιτικών τακτικών. Τέτοια παραδείγματα περιλαμβάνουν διπλωματικές εξαπατήσεις και τη χρήση πληροφοριών για τη χειραγώγηση αντιπάλων κατά τους μεσαιωνικούς χρόνους (Mitnick, 2011).
3. **Ανάπτυξη της κοινωνικής μηχανικής στον 20ο αιώνα**: Στις αρχές του 20ού αιώνα, η κοινωνική μηχανική άρχισε να αναγνωρίζεται ως ένας σημαντικός κίνδυνος στην ασφάλεια των επικοινωνιών και των πληροφοριών. Το πρώτο γνωστό παράδειγμα ψηφιακής κοινωνικής μηχανικής εμφανίστηκε τη δεκαετία του 1970, όταν οι «phreakers», όπως ο John Draper, εκμεταλλεύτηκαν το τηλεφωνικό σύστημα για δωρεάν κλήσεις. Αυτοί οι πρωτοπόροι των τηλεπικοινωνιακών απατών χρησιμοποιούσαν συσκευές που ονομάζονταν "blue boxes" για να παρακάμπτουν τα τηλεφωνικά δίκτυα (Salahdine & Kaabouch, 2019).
4. ***Η άνοδος της κοινωνικής μηχανικής κατά την εποχή του Διαδικτύου***: Καθώς οι υπολογιστές και το διαδίκτυο άρχισαν να γίνονται κυρίαρχα εργαλεία στην καθημερινή ζωή, η κοινωνική μηχανική αναπτύχθηκε παράλληλα. Η δεκαετία του 1990 σηματοδοτεί την έναρξη της μαζικής χρήσης των τεχνικών αυτών, με περιπτώσεις όπως του Kevin Mitnick, που χρησιμοποιούσε τεχνικές κοινωνικής μηχανικής για να αποκτήσει πρόσβαση σε ευαίσθητα δίκτυα χωρίς να χρειάζεται να χακάρει άμεσα υπολογιστές. Αυτή η περίοδος σηματοδοτεί επίσης την αρχή της συνειδητοποίησης του κινδύνου της κοινωνικής μηχανικής στην κυβερνοασφάλεια (Mitnick, 2011).
5. ***Σύγχρονες εξελίξεις***: Στις μέρες μας, οι επιθέσεις κοινωνικής μηχανικής έχουν γίνει όλο και πιο εξελιγμένες, με τεχνικές όπως το phishing, το pretexting και το baiting να είναι ευρέως διαδεδομένες. Οι επιτιθέμενοι χρησιμοποιούν πιο εξελιγμένα σενάρια και εργαλεία (ακόμη και την Τεχνητή Νοημοσύνη) για να ξεγελάσουν τα θύματά τους, ενώ η εξάπλωση των κοινωνικών δικτύων και των ψηφιακών πλατφορμών έχει δώσει νέα εργαλεία στους επιτιθέμενους (Salahdine & Kaabouch, 2019; Crowdstrike, 2023).

2.2.3 Κίνητρα των επιτιθέμενων κοινωνικής μηχανικής

Οι επιτιθέμενοι στον κυβερνοχώρο καθοδηγούνται από ένα ευρύ φάσμα στόχων, οι οποίοι συχνά καθορίζουν τον χαρακτήρα και την ένταση των επιθέσεών τους. Τα συνήθη κίνητρα περιλαμβάνουν:

1. *Οικονομικό κέρδος:* Το οικονομικό κέρδος αποτελεί κύριο κίνητρο πίσω από τις επιθέσεις κοινωνικής μηχανικής, όπου οι επιτιθέμενοι προσπαθούν να αποσπάσουν εμπιστευτικά οικονομικά δεδομένα, όπως πληροφορίες πιστωτικών καρτών ή κωδικούς

πρόσβασης σε τραπεζικούς λογαριασμούς. Μέσω αυτών των επιθέσεων, κακόβουλα άτομα προχωρούν σε δόλιες οικονομικές δραστηριότητες, όπως κλοπή ταυτότητας (και πχ. χρήση της για την μεταφορά ή εξαργύρωση οικονομικών ποσών) ή εκβιασμό. Οι επιθέσεις ransomware, στις οποίες οι δράστες εισάγουν λογισμικό κρυπτογράφησης μέσω κοινωνικής μηχανικής, απαιτώντας λύτρα για την αποκρυπτογράφηση των αρχείων, αποτελούν χαρακτηριστικό παράδειγμα αυτής της προσέγγισης (Timon Lopez et al., 2021).

2. *Κρυφές επιχειρήσεις και απόκτηση πληροφοριών:* Οργανισμοί, κυβερνητικές υπηρεσίες ή άτομα μπορεί να αποτελέσουν στόχο κρατικών φορέων και ομάδων κυβερνοκατασκοπείας με σκοπό την κλοπή ευαίσθητων πληροφοριών για πολιτική, στρατιωτική ή εμπορική κατασκοπεία. Οι επιτιθέμενοι μπορεί να επιχειρήσουν να αποκτήσουν διάφορα είδη δεδομένων, όπως εμπορικά μυστικά ή διαβαθμισμένα δεδομένα, προκειμένου να αποκτήσουν ανταγωνιστικό πλεονέκτημα ή να αποδυναμώσουν τους γεωπολιτικούς τους αντιπάλους (Alabdan, 2020).

3. *Χακτιβισμός*: Ο χακτιβισμός αναφέρεται στις δραστηριότητες ομάδων που πραγματοποιούν επιθέσεις σε υπολογιστές για την προώθηση κοινωνικών ή πολιτικών στόχων. Οι επιθέσεις αυτές μπορεί να περιλαμβάνουν αλλοιώσεις ιστοτόπων, κατανεμημένες επιθέσεις άρνησης παροχής υπηρεσιών (DDoS) ή παραβιάσεις δεδομένων με σκοπό τη διατάραξη λειτουργιών. Συχνά ξεκινούν με επιθέσεις κοινωνικής μηχανικής για την απόκτηση πρόσβασης σε κρίσιμα δεδομένα ή συστήματα, διευκολύνοντας τη διείσδυση ή τη διενέργεια στοχευμένων επιθέσεων.

4. *Κυβερνοπόλεμος*: Στο πλαίσιο παγκόσμιας σύγκρουσης, οι επιθέσεις στον κυβερνοχώρο μπορούν να χρησιμοποιηθούν ως μέσο πολέμου για να καταστρέψουν ζωτικές υποδομές, να βλάψουν στρατιωτικά συστήματα ή να αποσταθεροποιήσουν τα αντίπαλα έθνη. Οι δραστηριότητες κυβερνοπολέμου υπό κρατική αιγίδα περιλαμβάνουν περίπλοκες επιθέσεις με στόχο κυβερνητικά δίκτυα, αμυντικά συστήματα ή βασικά περιουσιακά στοιχεία, οι οποίες ενδεχομένως να ξεκινούν με μια επίθεση κοινωνικής μηχανικής (Chakraborty et al., 2023).

5. *Οργανωμένες ομάδες κυβερνοεγκλήματος*: Τα οργανωμένα συνδικάτα κυβερνοεγκλήματος λειτουργούν με κύριο στόχο την παραγωγή οικονομικού κέρδους μέσω παράνομων πράξεων, συμπεριλαμβανομένων της διαδικτυακής απάτης, της κλοπής ταυτότητας και της εμπορίας κλεμμένων δεδομένων στον σκοτεινό ιστό. Οι ομάδες αυτές μπορεί να επικεντρώνονται σε διαφορετικούς τύπους εγκληματικών δραστηριοτήτων στον κυβερνοχώρο, χρησιμοποιώντας προηγμένες μεθόδους και υποδομές για την πραγματοποίηση επιθέσεων μεγάλης κλίμακας που έχουν σημαντικές συνέπειες παγκοσμίως. Τέτοιες σύνθετες επιθέσεις, όταν στοχεύουν ασφαλή συστήματα, μπορεί να εκκινούν ομοίως με επιθέσεις κοινωνικής μηχανικής ώστε να παρακαμφθούν οι σχετικές υπηρεσίες ασφαλείας των συστημάτων.

Συνοψίζοντας, οι επιτιθέμενοι στον κυβερνοχώρο παρακινούνται από διάφορους παράγοντες, όπως το χρηματικό κέρδος, η κατασκοπεία, ο χακτιβισμός, ο κυβερνοπόλεμος

και οι οργανωμένες εγκληματικές δραστηριότητες. Η κατανόηση των κινήτρων πίσω από τις επιθέσεις στον κυβερνοχώρο είναι απαραίτητη για την ανάπτυξη ισχυρών μέτρων κυβερνοασφάλειας και τη μείωση των απειλών που προέρχονται από εχθρικά άτομα (Gibson, 2017). Σημειώνουμε εδώ πως τέτοιες σύνθετες επιθέσεις μπορεί να έχουν ως αφετηρία την κοινωνική μηχανική, ιδιαίτερα διότι υπάρχει πλέον στις μέρες μας η τάση να φτιάχνονται πιο ασφαλή συστήματα ή να ενισχυονται τα υπάρχοντα, οδηγούμενοι από τα μαθήματα της πρόσφατης πανδημίας. Επομένως, θα πρέπει να μελετηθούν οι τρόποι αποτροπής του πρώτου βήματος αφετηρίας (δηλ. της κοινωνικής μηχανικής) ώστε οι σύνθετες αυτές επιθέσεις να μην επιτύχουν ποτέ τον σκοπό τους.

2.2.4 Ψυχολογικές τεχνικές και παράγοντες που εκμεταλλεύονται οι επιτιθέμενοι

Η πιο σημαντική δεξιότητα για τους θύτες των επιθέσεων κοινωνικής μηχανικής είναι η άριστη κατανόηση του ψυχισμού των θυμάτων τους και η μετέπειτα εκμετάλλευσή του. Οι επιτιθέμενοι χρησιμοποιούν διάφορους ψυχολογικούς παράγοντες για την χειραγώγηση των θυμάτων ώστε να αυξήσουν τις πιθανότητες επιτυχίας των επιθέσεών τους.

Παρακάτω αναλύονται οι κύριες τεχνικές και παράγοντες που χρησιμοποιούν και εκμεταλλεύονται οι επιτιθέμενοι.

1. Εμπιστοσύνη και Αξιοπιστία

Οι επιτιθέμενοι προσπαθούν να κερδίσουν την εμπιστοσύνη του θύματος παρουσιάζοντας τον εαυτό τους ως αξιόπιστη πηγή. Αυτή η τεχνική είναι ιδιαίτερα αποτελεσματική σε περιπτώσεις επιθέσεων pretexting, όπου οι επιτιθέμενοι προσποιούνται ότι είναι άτομα με εξουσία ή συνεργάτες του θύματος με βάση ένα πειστικό αλλά ψευδές σενάριο.

2. Επείγουσα Ανάγκη

Η δημιουργία ενός αισθήματος επείγουσας ανάγκης αποτελεί μία από τις πιο συνηθισμένες τεχνικές χειραγώγησης. Οι επιτιθέμενοι συχνά προσπαθούν να πιέσουν τα θύματα να λάβουν άμεσες αποφάσεις, χωρίς να σκεφτούν επαρκώς τις συνέπειες. Για παράδειγμα, μπορεί να πείσουν τα θύματα πως ο υπολογιστής τους έχει κολλήσει κάποιο νέο ιό, έτσι ώστε αυτά να εκφορτώσουν και να εγκαταστήσουν ένα νεωτεριστικό αντι-ιικό στον υπολογιστή αυτό που δεν είναι τίποτε άλλο από ένα κακόβουλο λογισμικό, όπως είναι το λυτρισμικό (ransomware).

3. Εκμετάλλευση της Περιέργειας

Η περιέργεια είναι ένας σημαντικός παράγοντας που οδηγεί πολλούς χρήστες να αλληλεπιδράσουν με κακόβουλα αρχεία ή συνδέσμους. Οι επιτιθέμενοι χρησιμοποιούν δολώματα, όπως ψευδείς ενημερώσεις ή πληροφορίες που τραβούν την προσοχή του θύματος, για να τους πείσουν να κάνουν κλικ σε κακόβουλους συνδέσμους (*πχ.* επιθέσεις baiting**)**.

4. Εκμετάλλευση της Αρχής

Η αρχή και η εξουσία είναι έννοιες που εκμεταλλεύονται οι επιτιθέμενοι για να προκαλέσουν συμμόρφωση από το θύμα. Οι άνθρωποι έχουν την τάση να υπακούουν τυφλά σε πρόσωπα που φαίνεται να έχουν εξουσία, όπως ανώτεροι υπάλληλοι ή αξιωματούχοι (*πχ.* επιθέσεις pretexting ή επιθέσεις spear-phishing).

5. Φόβος και Απειλές

Οι επιθέσεις κοινωνικής μηχανικής συχνά βασίζονται στον φόβο για να παρακινήσουν τα θύματα να δράσουν. Αυτός ο φόβος μπορεί να περιλαμβάνει απειλές για την απώλεια προσωπικών πληροφοριών, χρημάτων ή ακόμη και τη φήμη του θύματος (*πχ.* επιθέσεις scareware).

6. Εξοικείωση

Η τεχνική αυτή βασίζεται στη χρήση γνωστών ονομάτων, σημάτων ή εταιρειών, ώστε να δημιουργήσει μια αίσθηση ασφάλειας στο θύμα. Οι επιτιθέμενοι χρησιμοποιούν επωνυμίες και λογότυπα για να παραπλανήσουν το θύμα και να το κάνουν να πιστέψει ότι αλληλεπιδρά με μια νόμιμη εταιρεία *(πχ.* επιθέσεις phishing).

7. Ευγένεια / Ντροπή

Η τεχνική αυτή βασίζεται στην εκμετάλλευση της ευγένειας ή της αμηχανίας του θύματος. Οι επιτιθέμενοι μπορεί να προσποιηθούν ότι χρειάζονται βοήθεια, όπως στην περίπτωση που φέρουν αντικείμενα και ζητούν από έναν εργαζόμενο να ανοίξει μια ελεγχόμενη πόρτα. Το θύμα, από ευγένεια ή ντροπή να ζητήσει διαπιστευτήρια, μπορεί να τους επιτρέψει την είσοδο. Έτσι, ο επιτιθέμενος καταφέρνει να παρακάμψει τα φυσικά μέτρα ασφαλείας χωρίς να εγείρει υποψίες (*πχ.* επιθέσεις tailgating/piggybacking).

2.2.5 Μεθοδολογία βημάτων που χρησιμοποιείται για την εκτέλεση επιθέσεων κοινωνικής μηχανικής

Οι επιτιθέμενοι κοινωνικής μηχανικής χρησιμοποιούν μια σειρά από βήματα μιας μελετημένης μεθοδολογίας, για να εξαπατήσουν τα θύματα και να αποκτήσουν πρόσβαση σε εμπιστευτικές πληροφορίες ή συστήματα, εκμεταλλευόμενοι αδυναμίες στα δίκτυα και την ανθρώπινη συμπεριφορά. Παρακάτω αναλύονται τα βασικά βήματα που χρησιμοποιούνται στην εκτέλεση τέτοιων επιθέσεων:

1. *Προετοιμασία της επίθεσης*: Σε αυτό το στάδιο (γνωστό ως reconnaissance), οι επιτιθέμενοι συλλέγουν πληροφορίες για το θύμα ή τον οργανισμό, όπως email, τηλεφωνικούς αριθμούς ή ακόμα και προφίλ στα κοινωνικά μέσα (social media). Αυτό βοηθά στο σχεδιασμό και την προετοιμασία για τη διενέργεια της επίθεσης (*πχ.* phishing ή pretexting, όπου απαιτούνται λεπτομερείς πληροφορίες για να εξαπατηθεί το θύμα). Μία γνωστή στρατηγική που επικεντρώνεται σε αυτήν την προετοιμασία ονομάζεται "Dumpster Diving" και αφορά τη συλλογή πληροφοριών μέσω της αναζήτησης ευαίσθητων δεδομένων σε φυσικά σκουπίδια οργανισμών. Οι επιτιθέμενοι μπορούν να βρουν στοιχεία που θα τους βοηθήσουν να σχεδιάσουν την επίθεση κοινωνικής μηχανικής, όπως σημειώσεις με ευαίσθητα δεδομένα (πχ. κωδικοί πρόσβασης) ή άλλα σημαντικά έγγραφα που δεν καταστράφηκαν σωστά. Αυτή η τεχνική υποδεικνύει ότι η κοινωνική μηχανική μπορεί να αφορά τόσο φυσικές όσο και ψηφιακές πηγές πληροφοριών (Perc et al., 2020).

2. *Δημιουργία σεναρίου*: Ο επιτιθέμενος δημιουργεί ένα ψεύτικο σενάριο, το οποίο προσαρμόζει στο εκάστοτε προφίλ θύματος. Για παράδειγμα, σε επιθέσεις vishing, ο επιτιθέμενος μπορεί να παριστάνει ότι είναι υπάλληλος μιας τράπεζας και να ζητήσει από το θύμα ευαίσθητες πληροφορίες. Ένα άλλο σενάριο, σε επιθέσεις smishing, όπου χρησιμοποιούνται SMS, θα μπορούσε να είναι η μίμηση μιας νόμιμης εταιρείας για την παραπλάνηση του θύματος.

3. *Αλληλεπίδραση με το θύμα*: Η ουσιαστική εκτέλεση της επίθεσης ξεκινά κυρίως με την άμεση επικοινωνία θύτη-θύματος. Για παράδειγμα, στην περίπτωση του phishing, αποστέλλεται ένα παραπλανητικό email με στόχο να ξεγελάσει το θύμα, ώστε να αποκαλύψει ευαίσθητες πληροφορίες (Perc et al., 2020).

4. *Εισαγωγή κακόβουλου λογισμικού*: Σε αυτό το βήμα, ο επιτιθέμενος χρησιμοποιεί την ανθρώπινη αλληλεπίδραση για να πείσει το θύμα να κατεβάσει κακόβουλο λογισμικό. Μέσω του baiting, οι επιτιθέμενοι προσφέρουν κάτι δελεαστικό (πχ., δωρεάν λογισμικό) που ενθαρρύνει το θύμα να κατεβάσει το κακόβουλο λογισμικό.

5. *Εκμετάλλευση δεδομένων*: Αφού ο επιτιθέμενος αποκτήσει πρόσβαση στα δεδομένα, τα χρησιμοποιεί για κακόβουλους σκοπούς. Στο phishing, τα δεδομένα μπορεί να χρησιμοποιηθούν για κλοπή ταυτότητας ή πρόσβαση σε χρηματοοικονομικά δεδομένα. Σε περιπτώσεις λυτρισμικού (ransomware), τα δεδομένα κρυπτογραφούνται και οι επιτιθέμενοι απαιτούν λύτρα για την αποκατάσταση/αποκρυπτογράφησή τους. Στην brute force επίθεση (επίθεση ωμής βίας), εφόσον οι δράστες μαντέψουν κωδικούς πρόσβασης (μέσω αδιάκοπων προσπαθειών δοκιμής και σφάλματος), χρησιμοποιούν τα αποκτηθέντα διαπιστευτήρια για να

αποκτήσουν πρόσβαση σε συστήματα ή λογαριασμούς και να εκμεταλλευτούν τα δεδομένα που έχουν παραβιαστεί.

Συνοψίζοντας, η εκτέλεση επιθέσεων κοινωνικής μηχανικής βασίζεται σε μια σειρά από στοχευμένα βήματα, όπως η συλλογή πληροφοριών, η δημιουργία ψευδών σεναρίων και η αλληλεπίδραση με το θύμα, με σκοπό την εκμαίευση δεδομένων και την εκμετάλλευσή τους. Ωστόσο, οι μέθοδοι κοινωνικής μηχανικής αποτελούν μόνο ένα μέρος της φαρέτρας των επιτιθέμενων. Συχνά χρησιμοποιούνται ως πρώτο στάδιο, πριν ακολουθήσουν πιο προηγμένες επιθέσεις κυβερνοασφάλειας, δημιουργώντας έτσι σύνθετες αλλά αποτελεσματικές επιθέσεις. Η κατανόηση αυτής της μεθοδολογίας είναι κρίσιμη για την αποτροπή και ανίχνευση τέτοιων επιθέσεων (Oltramari & Kott, 2018).

2.2.6 Τύποι επιθέσεων κοινωνικής μηχανικής και αντίκτυπος

Η κοινωνική μηχανική θεωρείται μία από τις πιο αποτελεσματικές μεθόδους παραβίασης της ασφάλειας, αφού βασίζεται στην εκμετάλλευση του ανθρώπινου παράγοντα που αποτελεί την αχίλλειο πτέρνα μεγάλων οργανισμών και επιχειρήσεων. Μέχρι και σήμερα πολλά πρόσωπα έχουν καταφέρει να γίνουν διάσημα εξαιτίας των επιθέσεών τους που στόχευσαν σε κορυφαίους οργανισμούς και προκάλεσαν μεγάλες επιπτώσεις στην παγκόσμια κυβερνοασφάλεια.

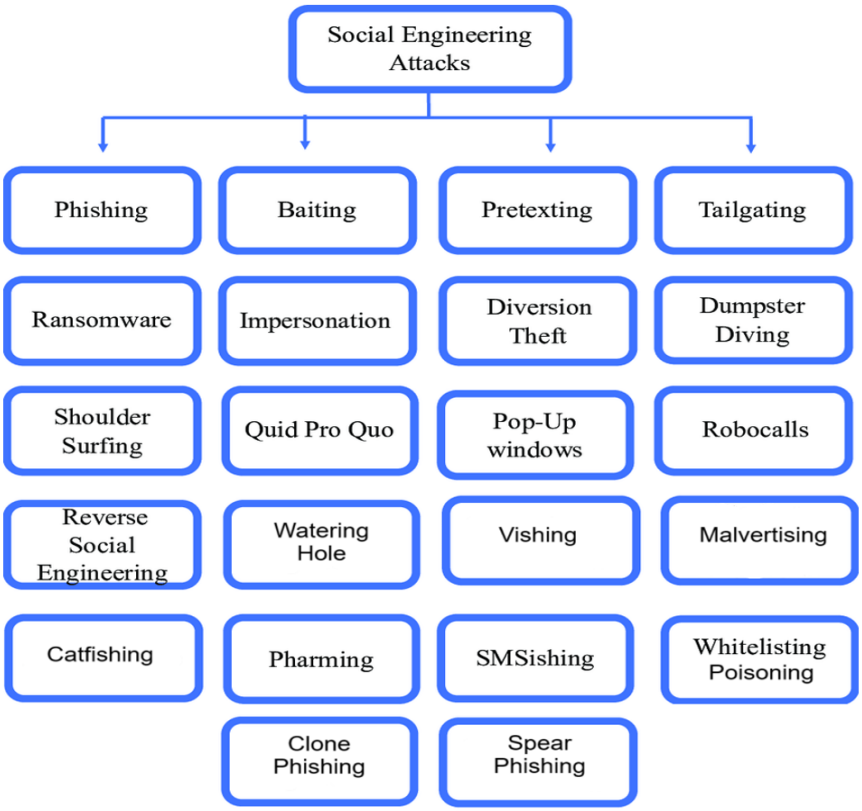
Ένα εξαιρετικά γνωστό πρόσωπο του τομέα είναι ο **Adrian Lamo,** ο οποίος έγινε γνωστός ως ένας από τους πιο διαβόητους χάκερ. Παραβίασε συστήματα μεγάλων οργανισμών, όπως Microsoft, Yahoo!, και New York Times, χρησιμοποιώντας κυρίως ευπάθειες στις υποδομές τους. Το στυλ του Lamo ήταν πιο «ηθικό» σε σύγκριση με άλλους χάκερς, καθώς συχνά ενημέρωνε τους οργανισμούς για τις παραβιάσεις τους και τις αδυναμίες ασφαλείας τους που αυτός εκμεταλλευόταν. Ωστόσο, έγινε παγκοσμίως γνωστός όταν ενημέρωσε τις αρχές για τις δραστηριότητες του Chelsea Manning, του στρατιώτη που διέρρευσε απόρρητα έγγραφα στο WikiLeaks (Lamo et al., 2022).

Επιπλέον, δεν θα μπορούσαμε να μην αναφερθούμε στον **Gary McKinnon**, γνωστός και ως "Solo", δηλαδή ενός από τους πιο διάσημους χάκερς στην ιστορία. Το 2001, ο McKinnon παραβίασε συστήματα της NASA και του Πενταγώνου, καταφέρνοντας να έχει πρόσβαση σε εξαιρετικά ευαίσθητες πληροφορίες των ΗΠΑ. Η παραβίαση ήταν τόσο εκτεταμένη που οι αμερικανικές αρχές τον κατηγόρησαν ότι προκάλεσε ζημιές ύψους εκατομμυρίων δολαρίων στα στρατιωτικά τους δίκτυα. Ο McKinnon υποστήριξε ότι ο στόχος του ήταν να αποκαλύψει πληροφορίες σχετικά με τη συγκάλυψη εξωγήινων τεχνολογιών. Ένα από τα μεγαλύτερά του χτυπήματα στην ιστορία αποτελεί αυτό της παραβίασης του δικτύου

της New York Times το 2002, όπου εισχώρησε στη βάση δεδομένων συνδρομητών και απέκτησε πρόσβαση σε εμπιστευτικές πληροφορίες (Erbschloe, M., 2024).

Τέλος, ένα από τα πιο γνωστά πρόσωπα στον τομέα της κοινωνικής μηχανικής είναι ο **Kevin Mitnick**, ένας από τους πιο διάσημους ειδικούς στην κυβερνοασφάλεια και πρώην χάκερ, ο οποίος μέσα από τις επιθέσεις και τις αναλύσεις του, ανέδειξε την ευκολία με την οποία οι άνθρωποι μπορούν να εξαπατηθούν (Mitnick et al., 2019).

Οι επιθέσεις κοινωνικής μηχανικής (Εικόνα 2) έχουν διάφορες μορφές, κάθε μία από τις οποίες έχει σχεδιαστεί για να χειραγωγήσει τα άτομα ώστε να αποκαλύψουν ευαίσθητες πληροφορίες, να εκτελέσουν ενέργειες (που μπορεί να είναι εις βάρος τους ή της εταιρείας τους) ή να λάβουν αποφάσεις που ωφελούν τον επιτιθέμενο. Οι επιθέσεις κοινωνικής μηχανικής περιλαμβάνουν ένα ευρύ φάσμα τακτικών που αποσκοπούν στη χειραγώγηση της ανθρώπινης ψυχολογίας και συμπεριφοράς για να εξαπατήσουν άτομα ή οργανισμούς ώστε να αποκαλύψουν ευαίσθητες πληροφορίες, να εκτελέσουν ενέργειες που θέτουν σε κίνδυνο την ασφάλεια ή να διευκολύνουν τη μη εξουσιοδοτημένη πρόσβαση σε συστήματα ή πόρους. Οι επιθέσεις αυτές εκμεταλλεύονται εγγενείς ανθρώπινες αδυναμίες, όπως η εμπιστοσύνη, η περιέργεια και η εξουσία, για να επιτύχουν τους στόχους τους.



Εικόνα 2: Επιθέσεις ΚΜ.

Ακολουθεί μια ταξινόμηση των κοινών επιθέσεων κοινωνικής μηχανικής με βάση τη δουλειά των Basit κ.α. (2021). Η ταξινόμηση αυτή εμφανίζεται στην Εικόνα 2.

1. *Phishing*: Σε μια επίθεση phishing, οι επιτιθέμενοι στέλνουν παραπλανητικά μηνύματα ηλεκτρονικού ταχυδρομείου που περιλαμβάνουν συνδέσμους σε κακόβουλους ιστότοπους. Τα μηνύματα αυτά φαίνονται να είναι νόμιμα (*πχ*. έχουν το λογότυπο μιας υπαρκτής εταιρείας που πλαστοπροσωπείται) για να εξαπατήσουν τους παραλήπτες. Η απάτη έχει ως αποτέλεσμα την αποκάλυψη σημαντικών, πιθανώς ιδιωτικών/προσωπικών δεδομένων του θύματος, όπως είναι διαπιστευτήρια σε υπηρεσίες. Οι επιθέσεις phishing μπορεί να οδηγήσουν σε παραβιάσεις δεδομένων, κλοπή ταυτότητας, οικονομικές απώλειες και μη εξουσιοδοτημένη πρόσβαση σε συστήματα ή λογαριασμούς (Hatfield, 2019). Για παράδειγμα, ένας επιτιθέμενος μπορεί να στείλει ένα μήνυμα ηλεκτρονικού ταχυδρομείου που παριστάνει ένα αξιόπιστο χρηματοπιστωτικό ίδρυμα, ζητώντας από τον παραλήπτη να ενημερώσει τις πληροφορίες του λογαριασμού του κάνοντας κλικ σε έναν σύνδεσμο. Κάνοντας κλικ στο σύνδεσμο, το θύμα οδηγείται σε έναν ψεύτικο ιστότοπο που έχει σχεδιαστεί να μοιάζει με τον νομότυπο του ιδρύματος ώστε να μπορεί να ξεγελάσει το θύμα για να παραδώσει τα στοιχεία σύνδεσής του (Perc et al., 2020).

Επιπτώσεις: Οι επιτιθέμενοι μπορούν να αποκτήσουν πρόσβαση σε προσωπικά δεδομένα, όπως κωδικούς πρόσβασης και στοιχεία πιστωτικών καρτών. Αυτό μπορεί να οδηγήσει σε οικονομικές απώλειες, κλοπή ταυτότητας, και παραβιάσεις λογαριασμών. Οι επιχειρήσεις μπορεί να υποστούν απώλεια εμπιστοσύνης των πελατών, ζημιά στη φήμη τους και νομικές συνέπειες αν παραβιαστούν δεδομένα πελατών.

2. *Vishing*: Το Vishing (φωνητικό ψάρεμα ή Voice Phishing) χρησιμοποιεί τηλεφωνικές κλήσεις ή το πρωτόκολλο VoIP (Voice over Internet Protocol) για να εξαπατήσει τους στόχους ώστε να αποκαλύψουν ευαίσθητες πληροφορίες ή να εκτελέσουν ανεπιθύμητες ενέργειες. Αυτό μπορεί να περιλαμβάνει την παραποίηση αναγνωριστικών κλήσης και τη χρήση τακτικών κοινωνικής μηχανικής για τη χειραγώγηση των θυμάτων **[29]**. Ουσιαστικά, πρόκειται για μία μορφή phishing που πραγματοποιείται μέσω τηλεφωνικών κλήσεων αντί ηλεκτρονικών μηνυμάτων (email). Για παράδειγμα, ένας επιτιθέμενος μπορεί να καλέσει έναν πελάτη τράπεζας, προσποιούμενος τον αντιπρόσωπο, και να ζητήσει τις πληροφορίες του λογαριασμού του με το πρόσχημα της επαλήθευσης της ταυτότητάς του για λόγους ασφαλείας.

Επιπτώσεις: Οι επιθέσεις Vishing μπορούν να οδηγήσουν σε κλοπή ταυτότητας, οικονομική απάτη, μη εξουσιοδοτημένη πρόσβαση και παραβίαση ευαίσθητων δεδομένων (Hatfield, 2019).

3. *Smishing*: Αυτή η μορφή phishing περιλαμβάνει την αποστολή παραπλανητικών μηνυμάτων κειμένου (SMS) για να εξαπατήσει τους παραλήπτες ώστε να κάνουν κλικ σε κακόβουλους συνδέσμους, να κατεβάσουν κακόβουλο λογισμικό ή να παράσχουν ευαίσθητες

πληροφορίες **[29]**. Για παράδειγμα, ένας επιτιθέμενος μπορεί να στείλει ένα μήνυμα κειμένου που προσποιείται ότι προέρχεται από μια υπηρεσία παράδοσης/διανομής, δίνοντας οδηγίες στον παραλήπτη να κάνει κλικ σε έναν σύνδεσμο για να παρακολουθήσει την πορεία παράδοσης ενός πακέτου, αλλά αντί αυτού τον οδηγεί σε έναν ιστότοπο phishing που έχει σχεδιαστεί για να κλέψει τα διαπιστευτήριά του (Rajasekharaiah et al., 2020).

Επιπτώσεις: Οι επιθέσεις Smishing μπορεί να οδηγήσουν σε μολύνσεις από κακόβουλο λογισμικό, παραβιάσεις δεδομένων, οικονομικές απώλειες και παραβίαση κινητών συσκευών.

4. *Pretexting*: Το Pretexting περιλαμβάνει τη δημιουργία ενός κατασκευασμένου σεναρίου ή προσχήματος (τα οποία βασίζονται σε πραγματικές πληροφορίες του θύματος, τις οποίες ο θύτης έπειτα από έρευνα έχει αντλήσει, με σκοπό την δημιουργία μιας πιο πειστικής ιστορίας) για τη χειραγώγηση των θυμάτων-στόχων ώστε να μοιραστούν εμπιστευτικές πληροφορίες με τον επιτιθέμενο ή να εκτελέσουν ανεπιθύμητες ενέργειες. Οι προηγμένες επιθέσεις αυτού του είδους επιχειρούν να εκμεταλλευτούν μια ανθρώπινη αδυναμία σε έναν οργανισμό ή μια εταιρεία για να προκαλέσουν ζημία σε αυτόν/αυτήν. Αυτή η μέθοδος απαιτεί από τον επιτιθέμενο να δημιουργήσει μια αξιόπιστη ιστορία που δεν αφήνει πολλά περιθώρια αμφισβήτησης από τον στόχο. Η στρατηγική είναι να χρησιμοποιηθούν συναισθήματα, όπως η περιέργεια, ο φόβος, το αίσθημα καθήκοντος, η ντροπή ή αμηχανία, ο επείγων χαρακτήρας, ενώ παράλληλα να οικοδομηθεί μια αίσθηση εμπιστοσύνης με το θύμα δημιουργώντας τις κατάλληλες συνθήκες για μια επιτυχημένη επίθεση (Timon-Lopez et al., 2021). Για παράδειγμα, ένας επιτιθέμενος μπορεί να υποδυθεί έναν τεχνικό γραφείου βοήθειας και να καλέσει έναν υπάλληλο, ισχυριζόμενος ότι χρειάζεται τα στοιχεία σύνδεσής του για μια αναβάθμιση του συστήματος. Ο υπάλληλος, πιστεύοντας ότι ο καλούντας είναι νόμιμος, παρέχει τις ζητούμενες πληροφορίες. Με αυτόν τον τρόπο οι εταιρείες μπορεί να χάσουν εμπιστευτικές πληροφορίες, ενώ οι εργαζόμενοι που πέφτουν θύματα αυτής της επίθεσης μπορεί να βρεθούν αντιμέτωποι με πειθαρχικές συνέπειες.

Επιπτώσεις: Η επίθεση αυτή μπορεί να οδηγήσει σε αποκάλυψη ευαίσθητων δεδομένων (πχ. εταιρικών μυστικών, διαπιστευτηρίων), μη εξουσιοδοτημένη πρόσβαση σε συστήματα και οικονομική απάτη (Hatfield, 2019). Επιπλέον, όπως αναφέρθηκε, οι εργαζόμενοι-θύματα μπορεί να βρεθούν αντιμέτωποι με πειθαρχικές συνέπειες ή ακόμη και να χάσουν τη δουλειά τους.

5. *Baiting*: Οι επιθέσεις Baiting δελεάζουν τους στόχους να προβούν σε μια συγκεκριμένη ενέργεια (πχ. να κάνουν κλικ σε έναν κακόβουλο σύνδεσμο, να κατεβάσουν ένα μολυσμένο αρχείο) προσφέροντας κάτι δελεαστικό, όπως δωρεάν λογισμικό, λήψεις πολυμέσων ή δωροκάρτες. Για παράδειγμα, ένας επιτιθέμενος μπορεί να διανείμει (απαρατήρητος) μονάδες USB που περιέχουν μολυσμένα από κακόβουλο λογισμικό αρχεία με την ένδειξη "Μπόνους υπαλλήλων" ή "Πολιτικές της εταιρείας". Όταν ανυποψίαστα άτομα συνδέσουν τις μονάδες USB, τα συστήματά τους θα μολυνθούν με κακόβουλο λογισμικό. Σε

κάποιες περιπτώσεις, το θύμα μπορεί να αναγκαστεί να εγκαταστήσει ένα κακόβουλο λογισμικό για να μπορέσει να αναπαράγει σωστά το πολυμέσο, επιτρέποντας στον επιτιθέμενο να εγκαταστήσει κακόβουλο λογισμικό **[30]**. Το δόλωμα είναι παρόμοιο με την επίθεση phishing, αλλά προσελκύει το θύμα μέσω στρατηγικών δελεασμού. Για παράδειγμα, οι χάκερ χρησιμοποιούν το δελεασμό των υποσχόμενων αγαθών εάν ένας χρήστης παραδώσει τα στοιχεία σύνδεσης σε έναν συγκεκριμένο ιστότοπο. Τα συστήματα δολώματος δεν περιορίζονται σε ψηφιακά on-line συστήματα αλλά μπορούν επίσης να ξεκινήσουν μέσω της χρήσης φυσικών μέσων (Timon-Lopez et al., 2021).

Επιπτώσεις: Μόλις ο επιτιθέμενος εγκαταστήσει κακόβουλο λογισμικό μέσω του δολώματος, μπορεί να αποκτήσει πρόσβαση σε προσωπικά δεδομένα, να κλέψει αρχεία ή ακόμα και να καταστρέψει δεδομένα. Το δόλωμα μπορεί επίσης να ανοίξει το δρόμο για επιθέσεις λυτρισμικού ή κλοπής πνευματικής ιδιοκτησίας.

6. *Quid Pro Quo*: Σε μια επίθεση quid pro quo, ο επιτιθέμενος προσφέρει κάτι πολύτιμο σε αντάλλαγμα για πληροφορίες ή πρόσβαση **[31]**. Για παράδειγμα, ο επιτιθέμενος μπορεί να επικοινωνήσει με έναν εργαζόμενο, προσποιούμενος ότι ανήκει στο τμήμα ανθρώπινου δυναμικού της εταιρείας, προσφέροντας μια υποτιθέμενη οικονομική ανταμοιβή ή μπόνους εφόσον ο εργαζόμενος αποκαλύψει τις προσωπικές του πληροφορίες, όπως τον αριθμό κοινωνικής ασφάλισης ή τα τραπεζικά στοιχεία.

Επιπτώσεις: Οι επιθέσεις με αντάλλαγμα μπορεί να οδηγήσουν σε μη εξουσιοδοτημένη πρόσβαση, παραβίαση δεδομένων και παραβίαση ευαίσθητων πληροφοριών, ενώ το αντάλλαγμα τελικώς δεν παρέχεται (Hatfield, 2019). Για παράδειγμα, ένας επιτιθέμενος μπορεί να παρουσιαστεί ως τεχνικός πληροφορικής και να προσφέρει δωρεάν αναβαθμίσεις λογισμικού εξ αποστάσεως με αντάλλαγμα τα στοιχεία σύνδεσης του θύματος. Παρόλο που η υποσχόμενη υπηρεσία δεν παρέχεται τελικά, ο επιτιθέμενος αποκτά μη εξουσιοδοτημένη πρόσβαση στο λογαριασμό του θύματος.

7. *Tailgating/Piggybacking*: Αυτή η φυσική επίθεση περιλαμβάνει την ακολουθία κάποιου με εξουσιοδοτημένη πρόσβαση σε έναν ασφαλή χώρο. Για παράδειγμα, ένας εισβολέας μπορεί να περιμένει κοντά σε ένα ασφαλές σημείο εισόδου και να ακολουθήσει στενά έναν εξουσιοδοτημένο υπάλληλο καθώς αυτός τραβάει την κάρτα πρόσβασής του, αποκτώντας μη εξουσιοδοτημένη είσοδο στην εγκατάσταση. Ο επιτιθέμενος αποκτά μη εξουσιοδοτημένη φυσική πρόσβαση σε εγκαταστάσεις ή συσκευές, εκμεταλλευόμενος πολλές φορές την εμπιστοσύνη, την καλοσύνη ή την ευγένεια εξουσιοδοτημένων ατόμων **[32]**. Για παράδειγμα, ο επιτιθέμενος μπορεί να κουβαλά ένα βάρος (πχ. πολλούς φακέλους) και να έχει και τα δυό του χέρια απασχολημένα. Οπότε, ένα εξουσιοδοτημένο άτομο από ευγένεια θα του ανοίξει την είσοδο.

Επιπτώσεις: Αυτός ο τύπος επίθεσης μπορεί να οδηγήσει σε μη εξουσιοδοτημένη πρόσβαση σε φυσικές εγκαταστάσεις, κλοπή εξοπλισμού ή δεδομένων και παραβίαση των πρωτοκόλλων ασφαλείας.

8. *Επιθέσεις πλαστοπροσωπίας (impersonation attacks)*: Στις επιθέσεις πλαστοπροσωπίας, ο επιτιθέμενος υιοθετεί μια ψεύτικη ταυτότητα για να πείσει το θύμα ότι αλληλεπιδρά με έναν αξιόπιστο ή εξουσιοδοτημένο πρόσωπο. Αυτό μπορεί να περιλαμβάνει τη μίμηση ανώτερων στελεχών, υπαλλήλων ή έμπιστων επαφών με σκοπό να αποσπαστούν ευαίσθητες πληροφορίες ή να εκτελέσει το θύμα ενέργειες που εξυπηρετούν τον επιτιθέμενο. Σε αντίθεση με άλλες επιθέσεις κοινωνικής μηχανικής, όπως το phishing ή το pretexting, η πλαστοπροσωπία έχει άμεση και συχνά προσωπική αλληλεπίδραση, όπως συμβαίνει στην περίπτωση impersonation in helpdesk attacks (ειδικότερος τύπος των *impersonation attacks*), όπου ο επιτιθέμενος θα μπορούσε να καλέσει έναν υπάλληλο της εταιρείας και να παρουσιαστεί ως τεχνικός υποστήριξης (helpdesk agent), ζητώντας τον κωδικό πρόσβασης ή άλλα διαπιστευτήρια με το πρόσχημα της επίλυσης ενός επείγοντος προβλήματος του συστήματος **[33]**. Αυτή η στρατηγική επιτρέπει την άμεση και στενή επαφή με το θύμα, αυξάνοντας την πειθώ του επιτιθέμενου και την πιθανότητα απόκτησης ευαίσθητων πληροφοριών.

Επιπτώσεις: Οι επιθέσεις πλαστοπροσωπίας μπορούν να οδηγήσουν σε απάτη, κλοπή ταυτότητας, μη εξουσιοδοτημένη πρόσβαση και παραβίαση ευαίσθητων δεδομένων.

9. *Watering Hole*: Σε μια επίθεση watering hole, οι επιτιθέμενοι μολύνουν ιστότοπους στους οποίους συχνάζει μια ομάδα ή ο οργανισμός-στόχος με σκοπό τη διανομή κακόβουλου λογισμικού ή τη συλλογή ευαίσθητων πληροφοριών. Στην περίπτωση της διανομής κακόβουλου λογισμικού, η επίθεση πραγματοποιείται μέσω μιας τεχνικής γνωστής ως drive-by-download, όπου το σύστημα των χρηστών μολύνεται (λόγω ευπαθειών του φυλλομετρητή) μόλις αυτοί επισκεφθούν τον μολυσμένο ιστότοπο, χωρίς να απαιτείται από αυτούς να εκτελέσουν κάποια ενέργεια, όπως η λήψη, η εκτέλεση ή το άνοιγμα ενός αρχείου. Για παράδειγμα, ένας εγκληματίας στον κυβερνοχώρο μπορεί να παραβιάσει ένα δημοφιλές βιομηχανικό φόρουμ ή έναν ειδησεογραφικό ιστότοπο που επισκέπτονται συχνά υπάλληλοι μιας εταιρείας-στόχου, ενσωματώνοντας κακόβουλο κώδικα που μολύνει τις συσκευές των επισκεπτών-υπαλλήλων με κακόβουλο λογισμικό (Rajasekharaiah et al., 2020). Αντίστοιχα, μέσω της επίθεσης XSS (Cross-Site Scripting), οι επιτιθέμενοι μπορούν να συλλέξουν ευαίσθητες πληροφορίες από τους χρήστες που αλληλεπιδρούν με τον ιστότοπο.

Επιπτώσεις: Οι επιθέσεις watering hole μπορεί να οδηγήσουν σε μολύνσεις από κακόβουλο λογισμικό, κλοπή δεδομένων και παραβίαση συστημάτων και δικτύων.

10. *Ransomware*: Ένας επιτιθέμενος που χρησιμοποιεί λυτρισμικό (ransomware) εισάγει το κακόβουλο αυτό λογισμικό στο σύστημα του θύματος για να κρυπτογραφήσει τα δεδομένα του και έπειτα του ζητά ένα ποσό για να τα αποκρυπτογραφήσει. Τα μηνύματα ηλεκτρονικού ταχυδρομείου ηλεκτρονικού «ψαρέματος», τα συνημμένα αρχεία κακόβουλου λογισμικού και οι μολυσμένοι ιστότοποι μπορούν να προκαλέσουν αυτού του είδους της μόλυνσης. Το θύμα λαμβάνει ειδοποίηση ότι τα αρχεία του έχουν κρυπτογραφηθεί και ότι πρέπει να πληρώσει ένα συγκεκριμένο ποσό, συνήθως σε κρυπτονόμισμα, για να λάβει το

κλειδί αποκρυπτογράφησης μετά την εγκατάσταση του λυτρισμικού στη συσκευή του (Greenberg, 2018).

Επιπτώσεις: Οι επιθέσεις λυτρισμικού μπορεί να έχουν σοβαρές συνέπειες, όπως την απώλεια πρόσβασης σε κρίσιμα δεδομένα, οικονομικές απώλειες, διακοπή της λειτουργίας οργανισμών και την πιθανή καταστροφή δεδομένων αν δεν καταβληθούν τα λύτρα. Για παράδειγμα, σε μια επίθεση ransomware, οι επιτιθέμενοι μπορούν να παραβιάσουν το δίκτυο μιας νοσοκομειακής μονάδας, κρυπτογραφώντας αρχεία ασθενών και κρίσιμα ιατρικά δεδομένα (Greenberg, 2018).

11. *Diversion Theft*: Μια τέτοια επίθεση συμβαίνει όταν προϊόντα ή δεδομένα που προορίζονται για ένα συγκεκριμένο άτομο ή ομάδα παραδίδονται τελικώς στη διεύθυνση του επιτιθέμενου. Συνήθως, οι επιτιθέμενοι χρησιμοποιούν την ταυτότητα αξιόπιστων ανθρώπων ή εξουσιοδοτημένων υπαλλήλων μιας εταιρείας για να παρεμβληθούν στη διαδικασία αποστολής ή διανομής και να τροποποιήσουν τη διεύθυνση παράδοσης (πχ. δίνοντας ψεύτικες οδηγίες παράδοσης στον σχετικό οδηγό). Στην επίθεση αυτη, οι δράστες μπορούν να στοχεύουν είτε σε φυσικά αντικείμενα είτε σε δεδομένα που μεταφέρονται ψηφιακά **[34]**. Για παράδειγμα, ένας επιτιθέμενος μπορεί να παραβιάσει τα συστήματα μιας εταιρείας και να εκτρέψει αποστολές πολύτιμων προϊόντων σε άλλη τοποθεσία, παρουσιάζοντας ψευδή έγγραφα αποστολής και αλλαγής παραλήπτη. Ένα πραγματικό παράδειγμα θα μπορούσε να είναι η περίπτωση όπου ένας επιτιθέμενος παρεμβαίνει στην επικοινωνία μεταξύ μιας εταιρείας και ενός προμηθευτή, αλλάζοντας τις λεπτομέρειες αποστολής και καταφέρνοντας να ανακατευθύνει πολύτιμα αγαθά σε μια διαφορετική τοποθεσία χωρίς να το αντιληφθούν οι εμπλεκόμενοι φορείς (Smith et al., 2021).

Επιπτώσεις: Οι επιθέσεις Diversion Theft μπορούν να οδηγήσουν σε απώλειες αγαθών ή κλοπή ευαίσθητων δεδομένων.

12. *Dumpster Diving*: Αυτό το είδος επίθεσης συμβαίνει όταν κάποιος ψάχνει στα σκουπίδια ενός ατόμου/οργανισμού για κρίσιμες πληροφορίες. Οι επιτιθέμενοι επωφελούνται από την πρακτική των επιχειρήσεων να απορρίπτουν συσκευές ή έγγραφα που δεν έχουν καταστραφεί σωστά. Προσωπικές πληροφορίες, σημειώσεις που προστατεύονται με κωδικό πρόσβασης, λογιστικά αρχεία, ακόμη και συσκευές αποθήκευσης, όπως USB και σκληροί δίσκοι, μπορούν να οδηγήσουν στην διαρροή των δεδομένων τους **[35]**.

Επιπτώσεις: Οι επιθέσεις Dumpster Diving μπορούν να οδηγήσουν σε παραβίαση ευαίσθητων πληροφοριών και κλοπή δεδομένων που θα μπορούσαν να χρησιμοποιηθούν για περαιτέρω κακόβουλες ενέργειες, όπως κλοπή ταυτότητας ή πρόσβαση σε εσωτερικά συστήματα του οργανισμού. Για παράδειγμα, ένας επιτιθέμενος μπορεί να βρει έγγραφα με πληροφορίες πελατών, καταστάσεις μισθοδοσίας ή οικονομικές αναφορές και να τα χρησιμοποιήσει για να αποκτήσει πρόσβαση σε συστήματα ή να εκβιάσει τον οργανισμό.

13. *Shoulder Surfing*: Μια τέτοια επίθεση συμβαίνει όταν ένας επιτιθέμενος παρακολουθεί κάποιον ενώ εισάγει σημαντικά δεδομένα, όπως κωδικούς πρόσβασης ή αριθμούς καρτών, σε δημόσιους ή ημι-δημόσιους χώρους, προσπαθώντας έτσι να αποκτήσει πρόσβαση σε αυτές τις πληροφορίες. Οι επιτιθέμενοι χρησιμοποιούν συχνά δημοφιλείς περιοχές, όπως ATM, καφετέριες και αεροδρόμια, για να παρακολουθούν τα θύματα κατά την εισαγωγή ευαίσθητων πληροφοριών σε κινητές συσκευές ή φορητούς υπολογιστές **[36]**. Ένα πραγματικό παράδειγμα είναι όταν ο επιτιθέμενος καταγράφει τον αριθμό PIN του θύματος σε ένα ATM και, στη συνέχεια, χρησιμοποιεί μια ειδική συσκευή για να αντιγράψει την κάρτα του θύματος από το ίδιο το ATM. Με την πρόσβαση τόσο στο PIN όσο και στην αντιγραμμένη κάρτα, ο επιτιθέμενος μπορεί να αποκτήσει πλήρη πρόσβαση στον τραπεζικό λογαριασμό του θύματος, προκαλώντας οικονομικές απώλειες και υποκλοπή δεδομένων.

Επιπτώσεις: Οι επιθέσεις Shoulder Surfing μπορούν να οδηγήσουν σε κλοπή ταυτότητας, πρόσβαση σε λογαριασμούς ή οικονομικές απώλειες, καθώς οι επιτιθέμενοι αποκτούν πρόσβαση σε κρίσιμες πληροφορίες χωρίς να χρειάζονται εξειδικευμένα εργαλεία.

14. *Reverse Social Engineering*: Σε αυτή την επίθεση, ο επιτιθέμενος αποφεύγει να πλησιάσει απευθείας το θύμα και αντ' αυτού δημιουργεί μια κατάσταση στην οποία το θύμα αναγκάζεται να επικοινωνήσει με αυτόν για βοήθεια. Ο επιτιθέμενος προκαλεί ένα σφάλμα στο σύστημα ή στις διεργασίες του θύματος και δημοσιεύει ότι μπορεί να το επιδιορθώσει παριστάνοντας τον γνήσιο και αξιόπιστο ειδικό. Το θύμα θα ζητήσει στη συνέχεια βοήθεια και κατά τη διάρκεια της «επίλυσης» του προβλήματος, ο επιτιθέμενος θα αποκτήσει ευαίσθητες πληροφορίες ή πρόσβαση σε ζωτικά συστήματα **[37]**. Για παράδειγμα, ένας επιτιθέμενος μπορεί να προκαλέσει προβλήματα σε λογισμικό που χρησιμοποιείται ευρέως από υπαλλήλους μιας εταιρείας και να εμφανιστεί σε φόρουμ υποστήριξης ως ειδικός στην επίλυση του προβλήματος. Έτσι, ένα θύμα-υπάλληλος, αναζητώντας βοήθεια, προσεγγίζει τον επιτιθέμενο και, κατά τη διάρκεια της διαδικασίας επίλυσης, δίνει στον επιτιθέμενο πρόσβαση σε ευαίσθητες πληροφορίες της εταιρείας.

Επιπτώσεις: Οι επιθέσεις Reverse Social Engineering μπορούν να οδηγήσουν σε κλοπή δεδομένων, παραβίαση συστημάτων ή ακόμα και εκβιασμό, καθώς το θύμα παραδίδει πρόθυμα πληροφορίες στον επιτιθέμενο.

15. *Pop-Up Windows*: Αυτή η επίθεση κοινωνικής μηχανικής χρησιμοποιεί αναδυόμενα παράθυρα (pop-ups) που εμφανίζονται στις οθόνες των θυμάτων κατά την περιήγησή τους στο διαδίκτυο. Τα αναδυόμενα παράθυρα μπορεί να εμφανίζονται ως έγκυρες ειδοποιήσεις από αξιόπιστους οργανισμούς ή υπηρεσίες, όπως τράπεζες ή υπηρεσίες ασφαλείας, ζητώντας προσωπικές πληροφορίες από το θύμα, όπως κωδικούς πρόσβασης ή στοιχεία πληρωμής. Οι επιτιθέμενοι συχνά προσπαθούν να πείσουν τα θύματα ότι υπάρχει ένα επείγον ζήτημα, όπως μια παραβίαση ασφαλείας, προκειμένου να τα παρακινήσουν να εμπλακούν στο κακόβουλο παράθυρο. Για παράδειγμα, ένα θύμα μπορεί να δει ένα pop-up που φαίνεται να προέρχεται από την τράπεζά του και του προτρέπει να εισάγει τα στοιχεία του για να "διορθώσει" ένα ψεύτικο πρόβλημα ασφαλείας. Εφόσον το θύμα προχωρήσει στην πράξη

αυτή, θα δώσει στην πραγματικότητα στον επιτιθέμενο πρόσβαση στον τραπεζικό του λογαριασμό. Ένα πραγματικό παράδειγμα είναι η χρήση pop-ups που προσποιούνται ενημερώσεις λογισμικού, όπου ο χρήστης κάνει κλικ και εγκαθιστά χωρίς να το γνωρίζει κακόβουλο λογισμικό (Jones et al., 2021).

Επιπτώσεις: Οι επιθέσεις μέσω αναδυόμενων παραθύρων μπορούν να οδηγήσουν σε κλοπή προσωπικών δεδομένων, εγκατάσταση κακόβουλου λογισμικού, ή πρόσβαση σε ευαίσθητους λογαριασμούς.

16. *Robocalls*: Οι επιθέσεις Robocalls συμβαίνουν όταν χρησιμοποιούνται αυτοματοποιημένες τηλεφωνικές κλήσεις ή ρομποκλήσεις (robocalls) για να εξαπατήσουν ανθρώπους. Σε αυτές τις κλήσεις, ένα ηχογραφημένο μήνυμα υποδύεται μια αξιόπιστη πηγή, όπως μια τράπεζα, μια κυβερνητική υπηρεσία ή μια τεχνολογική εταιρεία, και ενημερώνει το θύμα για ένα «επείγον ζήτημα» που πρέπει να διορθωθεί γρήγορα. Οι επιτιθέμενοι συχνά ζητούν από το θύμα να παράσχει προσωπικές πληροφορίες ή να προβεί σε μια συγκεκριμένη ενέργεια, όπως η πληρωμή ενός ψεύτικου λογαριασμού ή η επαναφορά των στοιχείων του λογαριασμού. Για παράδειγμα, ένας επιτιθέμενος μπορεί να προσποιηθεί ότι είναι από την εφορία και να ζητήσει την ταχεία πληρωμή ενός πλασματικού χρέους, τρομάζοντας το θύμα να παραδοθεί. Ένα πραγματικό παράδειγμα θα ήταν όταν οι επιτιθέμενοι χρησιμοποιούν ρομποτικές κλήσεις για να αποκτήσουν δεδομένα πιστωτικών καρτών από ανυποψίαστα θύματα (Brown et al., 2020).

Οι επιθέσεις vishing διαφέρουν από τις επιθέσεις robocall με βάση το γεγονός πως αφορούν ζωντανές επιθέσεις όπου ο επιτιθέμενος επικοινωνεί άμεσα με το θύμα. Αντιθέτως, μια επίθεση robocall περιλαμβάνει μια αυτοματοποιημένη ρομποκλήση. Αλλά μια τέτοια κλήση μπορεί έπειτα να οδηγήσει σε μια επίθεση vishing εφόσον το θύμα ανταποκριθεί (πχ. πάρει τηλέφωνο σε συγκεκριμένο αριθμό που ανήκει στον επιτιθέμενο).

Επιπτώσεις: Οι επιθέσεις με ρομποτικές κλήσεις μπορεί να οδηγήσουν σε παραβιάσεις απορρήτου και έκθεση σε απάτες.

17. *Pharming*: Αποτελεί μια επίθεση με την οποία οι επιτιθέμενοι εκτρέπουν κρυφά τους επισκέπτες από αξιόπιστες ιστοσελίδες σε επιβλαβείς. Το κακόβουλο λογισμικό που τοποθετείται στις συσκευές των χρηστών (το οποίο μπορεί να αλλάξει το τοπικό αρχείο ξενιστή - local host file) ή οι αλλαγές στις ρυθμίσεις DNS (Domain Name System) (πχ. μέσω επίθεσης DNS Poisoning) είναι δύο τρόποι για να επιτευχθεί αυτό. Η πρόθεση είναι να εξαπατήσουν τα θύματα ώστε να επισκεφθούν επικίνδυνους ιστότοπους που υποδύονται τους αξιόπιστους, προκειμένου να αποκτήσουν προσωπικά δεδομένα, συμπεριλαμβανομένων πληροφοριών τραπεζικών λογαριασμών και κωδικών πρόσβασης. Για παράδειγμα, εκτρέποντας τους πελάτες από τον δικτυακό τόπο μιας τράπεζας σε μια κακόβουλη σελίδα που έχει εκπληκτική ομοιότητα με αυτήν, οι επιτιθέμενοι μπορούν να κλέψουν τα προσωπικά τους δεδομένα (Miller et al., 2019). Σε αντίθεση με τις επιθέσεις phishing, στις επιθέσεις pharming ο χρήστης δεν

ακολουθεί κάποιον κακόβουλο σύνδεσμο αλλά ανακατευθύνετε σε αυτόν (χωρίς να το αντιλαμβάνεται) όταν επιχειρεί να επισκεφθεί την αντίστοιχη “νόμιμη” ιστοσελίδα.

Επιπτώσεις: Επειδή τα θύματα μπορεί να παρέχουν άθελά τους τις προσωπικές τους πληροφορίες σε δόλιους ιστότοπους, οι επιθέσεις phishing έχουν τη δυνατότητα να προκαλέσουν μεγάλες οικονομικές απώλειες. Η κλοπή ταυτότητας και η παράνομη πρόσβαση σε ιδιωτικούς ή επαγγελματικούς λογαριασμούς αποτελούν περαιτέρω κινδύνους.

18. Whitelisting Poisoning: Οι επιτιθέμενοι μέσω της συγκεκριμένης τακτικής, εκμεταλλεύονται τον μηχανισμό της «λευκής λίστας» (whitelisting) για να αποκτήσουν πρόσβαση σε συστήματα ή εφαρμογές που διαφορετικά θα ήταν απαγορευμένη. Η λευκή λίστα είναι μια μέθοδος ελέγχου πρόσβασης που επιτρέπει μόνο εγκεκριμένες εφαρμογές, ιστοσελίδες ή IP διευθύνσεις να αλληλεπιδρούν με το σύστημα. Οι επιτιθέμενοι, λοιπόν, χειραγωγούν αυτή τη διαδικασία, καταφέρνοντας να «περάσουν» κακόβουλα στοιχεία ως εγκεκριμένα (πχ. αποκτώντας τα διαπιστευτήρια ενός διαχειριστή μέσω μιας phishing επίθεσης και χρησιμοποιώντας τα για να προσπελάσουν και να τροποποιήσουν τη λευκή λίστα). Έπειτα, μπορούν να εκμεταλλευτούν το κενό ασφαλείας που δημιουργήθηκε για να διεισδύσουν στο δίκτυο της εταιρείας-θύματος (Brown et al., 2022).

Επιπτώσεις: Οι επιθέσεις αυτού του είδους μπορούν να οδηγήσουν σε ανεξέλεγκτη πρόσβαση σε συστήματα, διαρροή δεδομένων ή ακόμα και σε πλήρη παραβίαση του δικτύου ενός οργανισμού. Αν καταφέρει ένας επιτιθέμενος να εισαγάγει μια κακόβουλη εφαρμογή στη λευκή λίστα, μπορεί αυτή να χρησιμοποιηθεί για επιθέσεις σε βάθος, χωρίς να εντοπιστεί.

19. Clone Phishing: Οι απόπειρες Clone Phishing, περιλαμβάνουν τη δημιουργία κλωνοποιημένων μηνυμάτων ηλεκτρονικού ταχυδρομείου που μοιάζουν πολύ πανομοιότυπα με πραγματικά μηνύματα ηλεκτρονικού ταχυδρομείου που έχουν ήδη παραδοθεί. Αυτές οι επιθέσεις τροποποιούν μικροσκοπικά χαρακτηριστικά, όπως διευθύνσεις URL ή συνημμένα αρχεία, με αποτέλεσμα να δημιουργείται επιβλαβές υλικό. Οι παραλήπτες συχνά παραπλανώνται και υποθέτουν την γνησιότητα του μηνύματος και της προέλευσής του (Smith et al., 2021). Παράδειγμα: Ένας επιτιθέμενος δημιουργεί ένα κλωνοποιημένο email που φαίνεται να προέρχεται από έναν αξιόπιστο προμηθευτή και το στέλνει σε έναν υπάλληλο της εταιρείας. Το μήνυμα περιλαμβάνει έναν σύνδεσμο που φαίνεται γνήσιος αλλά οδηγεί σε κακόβουλο ιστότοπο, με σκοπό να κλέψει διαπιστευτήρια σύνδεσης. Ο υπάλληλος, εμπιστευόμενος το φαινομενικά οικείο μήνυμα, εισάγει τις πληροφορίες σύνδεσής του, δίνοντας στον επιτιθέμενο πρόσβαση σε ευαίσθητα δεδομένα.

Επιπτώσεις: Εάν οι επιτιθέμενοι καταφέρουν να πείσουν τα θύματα να αλληλεπιδράσουν με δόλιο υλικό, αυτό μπορεί να οδηγήσει σε κλοπή διαπιστευτηρίων, λήψη κακόβουλου λογισμικού ή παραβίαση του συστήματος. Αυτές οι επιθέσεις είναι ιδιαίτερα επιβλαβείς, καθώς τα θύματα πιστεύουν ότι επικοινωνούν με έναν αξιόπιστο αποστολέα, γεγονός που καθιστά δυσκολότερο τον εντοπισμό τους. Οι επιτιθέμενοι που επιτυγχάνουν μια τέτοια κλωνοποίηση μπορούν να αποκτήσουν πρόσβαση σε κρίσιμα δεδομένα ή να

χρησιμοποιήσουν το ηλεκτρονικό ταχυδρομείο για πιο προηγμένες επιθέσεις (Smith et al., 2021).

20. Catfishing: Αποτελεί τακτική κατά την οποία τα θύματα εξαπατώνται με την κατασκευή προφίλ σε διαδικτυακούς ιστότοπους. Οι επιτιθέμενοι συχνά κατασκευάζουν σχέσεις προκειμένου να αποκτήσουν ευαίσθητες πληροφορίες ή χρήματα, χρησιμοποιώντας πλασματικές εικόνες και προσωπικά δεδομένα για να κερδίσουν την εμπιστοσύνη των θυμάτων **[38]**. Παράδειγμα: Ένας επιτιθέμενος δημιουργεί ένα ψεύτικο προφίλ στα μέσα κοινωνικής δικτύωσης, προσποιούμενος ότι είναι επιτυχημένος επιχειρηματίας. Χρησιμοποιώντας παραπλανητικές τακτικές, δημιουργεί μια ψεύτικη σχέση με ένα θύμα, αποκτώντας την εμπιστοσύνη του. Στη συνέχεια, εκμεταλλεύεται αυτή τη σχέση για να αποσπάσει οικονομική βοήθεια ή ευαίσθητες πληροφορίες από το θύμα.

Επιπτώσεις: Τα θύματα αυτών των επιθέσεων υφίστανται σοβαρή οικονομική και ψυχολογική ζημία. Εκμεταλλευόμενος την ψυχολογική αδυναμία των θυμάτων, ο επιτιθέμενος μπορεί να αποκτήσει χρήματα, ιδιωτικές πληροφορίες εταιρειών ή ακόμη και προσωπικές πληροφορίες. Η ζημία είναι συνήθως σοβαρή και ανεπανόρθωτη από τη στιγμή που τα θύματα μαθαίνουν για την απάτη (Johnson et al., 2020).

21. Malvertising: Οι επιτιθέμενοι εισάγουν κακόβουλο λογισμικό μέσα σε διαφημίσεις που εμφανίζονται σε αξιόπιστους ιστότοπους. Οι χρήστες που κάνουν κλικ σε αυτές τις διαφημίσεις εκτίθενται σε κακόβουλο λογισμικό, το οποίο μπορεί να εγκατασταθεί στις συσκευές τους **[39]**. Παράδειγμα: Σε έναν δημοφιλή ιστότοπο κοινωνικής δικτύωσης, μια κακόβουλη διαφήμιση εμφανίζεται στο πλάι της σελίδας. Όταν ο χρήστης κάνει κλικ στη διαφήμιση, εγκαθίσταται αθόρυβα κακόβουλο λογισμικό στον υπολογιστή του (μέσω μιας drive-by-download επίθεσης), το οποίο παρακολουθεί τις κινήσεις του και συλλέγει ευαίσθητες πληροφορίες, όπως στοιχεία σύνδεσης και οικονομικά δεδομένα.

Επιπτώσεις: Οι επιθέσεις Malvertising είναι ιδιαίτερα επικίνδυνες καθώς ο χρήστης μπορεί να εκτεθεί σε κακόβουλο λογισμικό απλώς με την επίσκεψη μιας ιστοσελίδας. Μόλις το λογισμικό εγκατασταθεί, ο επιτιθέμενος μπορεί να αποκτήσει πρόσβαση σε ευαίσθητα δεδομένα, να εγκαταστήσει λυτρισμικό/κακόβουλο λογισμικό, ή να παρακολουθεί τις δραστηριότητες του χρήστη (Davis et al., 2018).

22. Spear Phishing: Το spear phishing είναι ένας τύπος στοχευμένου phishing κατά τον οποίο ο επιτιθέμενος δημιουργεί εξατομικευμένα μηνύματα προσαρμοσμένα σε ένα συγκεκριμένο άτομο ή επιχείρηση. Για να κάνει ένα μήνυμα να φαίνεται αυθεντικό, ο επιτιθέμενος συχνά περιλαμβάνει προσωπικές πληροφορίες, όπως το όνομα του θύματος, τη θέση εργασίας ή ακόμη και στοιχεία από τα κοινωνικά του δίκτυα. Ο σκοπός είναι να

εξαπατήσει το θύμα ώστε να αποκαλύψει ευαίσθητες πληροφορίες, να κατεβάσει κακόβουλο λογισμικό ή να παράσχει πρόσβαση σε βασικά συστήματα **[40]**. Για παράδειγμα, ένας επιτιθέμενος μπορεί να στείλει ένα ψεύτικο email στοχεύοντας έναν υπάλληλο ενός οργανισμού, προσποιούμενος έναν ανώτερο υπάλληλο ή συνεργάτη και ζητώντας από τον παραλήπτη να ανοίξει ένα επισυναπτόμενο αρχείο ή να ακολουθήσει έναν σύνδεσμο. Αν το θύμα ανταποκριθεί, ο επιτιθέμενος μπορεί να αποκτήσει άμεση πρόσβαση σε διαβαθμισμένες πληροφορίες ή να εγκαταστήσει κακόβουλο λογισμικό στο εταιρικό δίκτυο.

Επιπτώσεις: Οι επιθέσεις Spear Phishing είναι ιδιαίτερα επικίνδυνες, καθώς είναι εξαιρετικά πειστικές και δύσκολα εντοπίζονται. Οι επιπτώσεις περιλαμβάνουν την κλοπή δεδομένων, την πρόσβαση σε εταιρικά δίκτυα ή τραπεζικούς λογαριασμούς, και, σε ορισμένες περιπτώσεις, την πλήρη παραβίαση κρίσιμων συστημάτων.

Ο αντίκτυπος των επιθέσεων κοινωνικής μηχανικής μπορεί να ποικίλλει ανάλογα με την πολυπλοκότητα της επίθεσης, την ευπάθεια του στόχου και τα μέτρα ασφαλείας που εφαρμόζονται. Οι οργανισμοί και τα άτομα θα πρέπει να εφαρμόζουν ισχυρές πρακτικές ασφαλείας, συμπεριλαμβανομένης της εκπαίδευσης ευαισθητοποίησης, του ελέγχου ταυτότητας πολλαπλών παραγόντων, της κρυπτογράφησης και των τακτικών αξιολογήσεων ασφαλείας, για να μετριάσουν τους κινδύνους που σχετίζονται με τις επιθέσεις κοινωνικής μηχανικής (Hatfield, 2019). Ακόμη και οι πιο προηγμένες τεχνολογίες ασφαλείας, μπορούν να παρακαμφθούν αν οι υπάλληλοι δεν είναι κατάλληλα εκπαιδευμένοι και ευαισθητοποιημένοι για τις απειλές της κοινωνικής μηχανικής.

2.2.7 Παραδείγματα πραγματικών επιθέσεων κοινωνικής μηχανικής

Ακολουθούν παραδείγματα πραγματικών επιθέσεων κοινωνικής μηχανικής που γράφτηκαν στην ιστορία και αποδεικνύουν πόσο ευάλωτες μπορεί να γίνουν ακόμα και οι πιο κορυφαίες/οι επιχειρήσεις και οργανισμοί στον κόσμο (ανεξάρτητα των μέτρων ασφάλειας που μπορεί να εφαρμόζουν).

1. Η επίθεση στο Twitter (2020), (Lanterman, 2020):

- Τι συνέβη: Το 2020, μια από τις μεγαλύτερες επιθέσεις κοινωνικής μηχανικής στόχευσε το Twitter. Οι επιτιθέμενοι χρησιμοποιώντας τεχνικές phishing εξαπάτησαν υπαλλήλους της εταιρείας, αποκτώντας πρόσβαση σε εσωτερικά εργαλεία διαχείρισης. Με αυτόν τον τρόπο, κατάφεραν να παραβιάσουν λογαριασμούς διασημοτήτων και επιφανών προσωπικοτήτων, όπως ο Elon Musk, ο Jeff Bezos, και ο Joe Biden. Έπειτα, οι χάκερς χρησιμοποίησαν τους παραβιασμένους λογαριασμούς για να δημοσιεύσουν

απάτες κρυπτονομισμάτων, προτρέποντας τους ακολούθους (των λογαριασμών) να στείλουν Bitcoin σε συγκεκριμένα πορτοφόλια (με το ψεύτικο αντίτιμο τον διπλασιασμό των απεσταλμένων νομισμάτων).

Αντίκτυπος: Η επίθεση κατέδειξε τις αδυναμίες των εταιρικών διαδικασιών ασφαλείας, θέτοντας ερωτήματα για την προστασία των ψηφιακών πλατφορμών που χρησιμοποιούν δισεκατομμύρια άνθρωποι παγκοσμίως. Το περιστατικό προκάλεσε μεγάλες ανησυχίες για την ασφάλεια των κοινωνικών δικτύων και είχε επιπτώσεις στην αξιοπιστία του Twitter.

2. Η επίθεση στον RSA (2011), (Gruber, 2021):

- Τι συνέβη: Η επίθεση στον οργανισμό RSA Security, έναν από τους κορυφαίους προμηθευτές ασφάλειας, το 2011 θεωρείται μία από τις πιο περίπλοκες επιθέσεις κοινωνικής μηχανικής. Οι επιτιθέμενοι έστειλαν phishing emails σε εργαζομένους του RSA με ένα συνημμένο αρχείο Excel, το οποίο περιείχε κακόβουλο λογισμικό. Όταν οι εργαζόμενοι άνοιξαν το αρχείο, οι επιτιθέμενοι απέκτησαν πρόσβαση στο δίκτυο της εταιρείας. Ο λόγος όμως που η συγκεκριμένη επίθεση θεωρήθηκε πολύπλοκη έγκειται στο γεγονός ότι συνεχίστηκε πέρα από την πρώτη παραβίαση. Χρησιμοποιώντας τα διαπιστευτήρια που είχαν αποκτήσει, οι επιτιθέμενοι κινήθηκαν πλευρικά στο δίκτυο για να αποκτήσουν πρόσβαση σε περαιτέρω διακομιστές. Κυνήγησαν ιδιαίτερα την seed warehouse (αποθήκη “σπόρων”), η οποία περιείχε τα δεδομένα που αφορούσαν τα SecurID tokens. Παρακάμπτοντας επομένως τα μέτρα ασφαλείας, όπως τα συστήματα «air gap», οι επιτιθέμενοι κατάφεραν να αποκτήσουν ιδιωτικά δεδομένα που χρησιμοποιούνταν για τον έλεγχο ταυτότητας σε εκατοντάδες μεγάλες εταιρείες και κυβερνητικά ιδρύματα.

  Αντίκτυπος: Η παραβίαση είχε τεράστιες επιπτώσεις για την ασφάλεια των SecurID tokens και πολλές εταιρείες αναγκάστηκαν να τα αντικαταστήσουν, τα οποία πιθανότατα ήταν φυσικές συσκευές (hardware tokens) (άρα εδώ σίγουρα είχαμε οικονομική ζημία από την επίθεση). Η επίθεση αυτή κατέδειξε τον κίνδυνο των επιθέσεων phishing για τις υποδομές ασφάλειας και την ευπάθεια των μεγαλύτερων οργανισμών στον κόσμο.

3. Η επίθεση στη Target (2013), (Krebs, 2014):

- Τι συνέβη: Το 2013, η εταιρεία λιανικής Target υπέστη μία από τις μεγαλύτερες παραβιάσεις δεδομένων στην ιστορία της. Η επίθεση (τύπου εφοδιαστικής αλυσίδας) ξεκίνησε από μια επιτυχημένη επίθεση κοινωνικής μηχανικής σε έναν από τους παρόχους υπηρεσιών θέρμανσης και αερισμού της εταιρείας. Οι χάκερς απέκτησαν πρόσβαση στο δίκτυο του παρόχου μέσω phishing και, ακολουθώντας τεχνικές πλευρικής κίνησης, διείσδυσαν στο δίκτυο της Target. Εκεί εγκατέστησαν κακόβουλο λογισμικό στα συστήματα POS της εταιρείας, καταφέρνοντας έτσι να κλέψουν

προσωπικά δεδομένα και πληροφορίες πιστωτικών καρτών περίπου 40 εκατομμυρίων πελατών.

Αντικτυπος: Η επίθεση είχε τεράστιο οικονομικό αντίκτυπο στην Target, η οποία πλήρωσε εκατοντάδες εκατομμύρια δολάρια σε πρόστιμα και αποζημιώσεις. Το περιστατικό κατέδειξε τον κίνδυνο που συνιστούν οι τρίτοι πάροχοι και οι αλυσίδες εφοδιασμού για την ασφάλεια μιας εταιρείας.

2.2.8 Προστασία από επιθέσεις κοινωνικής μηχανικής

Η προστασία από επιθέσεις κοινωνικής μηχανικής απαιτεί από τους οργανισμούς την υιοθέτηση μιας πολυεπίπεδης στρατηγικής που συνδυάζει τεχνολογικές λύσεις, διαδικασίες και εκπαίδευση προσωπικού.

Ακολουθούν οι βασικές πρακτικές και τεχνικές για την προστασία από τις επιθέσεις κοινωνικής μηχανικής:

1. Εκπαίδευση και Ευαισθητοποίηση Προσωπικού

Η εκπαίδευση των εργαζομένων είναι από τα πιο σημαντικά αντίμετρα για την αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής. Οι εργαζόμενοι πρέπει να είναι σε θέση να αναγνωρίζουν τις τεχνικές/επιθέσεις που χρησιμοποιούν οι επιτιθέμενοι, όπως το phishing, το pretexting, και το vishing. Η συνεχής ενημέρωση για νέες απειλές και τα τακτικά εκπαιδευτικά προγράμματα ευαισθητοποίησης είναι απαραίτητα για να εκπαιδευτούν οι εργαζόμενοι ώστε να αναγνωρίζουν πότε γίνονται επιθέσεις κοινωνικής μηχανικής και να τις αποκρούουν, ενισχύοντας με αυτό τον τρόπο την ασφάλεια του οργανισμού τους.

2. Πολυπαραγοντική Αυθεντικοποίηση (MFA)

Η χρήση πολυπαραγοντικής αυθεντικοποίησης αποτελεί κρίσιμο μέτρο ασφάλειας, καθώς μειώνει την πιθανότητα μη εξουσιοδοτημένης πρόσβασης σε συστήματα, ακόμη και αν οι επιτιθέμενοι αποκτήσουν κωδικούς πρόσβασης μέσω επιθέσεων κοινωνικής μηχανικής. Η MFA απαιτεί από τους χρήστες να παρέχουν περισσότερες από μία μορφές ταυτοποίησης, όπως κωδικούς μιας χρήσης ή βιομετρικά δεδομένα.

3. Φιλτράρισμα Ηλεκτρονικού Ταχυδρομείου

Οι περισσότερες επιθέσεις κοινωνικής μηχανικής ξεκινούν μέσω ηλεκτρονικού ταχυδρομείου. Το φιλτράρισμα email είναι μια από τις πιο αποτελεσματικές τεχνολογικές λύσεις για την αποτροπή μηνυμάτων phishing και άλλων κακόβουλων email πριν φτάσουν στους τελικούς

χρήστες. Τα φιλτραρισμένα μηνύματα συχνά μεταφέρονται αυτόματα σε φακέλους όπως "Ανεπιθύμητη Αλληλογραφία" (Spam) ή "Αποκλεισμένα Μηνύματα" (Junk Mail). Αυτοί οι φάκελοι είναι προσβάσιμοι από τον χρήστη, αλλά καλό είναι να αποφεύγεται η επίσκεψη σε αυτούς, εκτός αν είναι απολύτως αναγκαίο, καθώς μπορεί να περιέχουν κακόβουλα μηνύματα (όπου κάποιος μπορεί πχ. να ανοίξει κατά λάθος κάποιο κακόβουλο συνημμένο αρχείο). Ορισμένα συστήματα φιλτραρίσματος διαγράφουν αυτόματα τα πιο επικίνδυνα μηνύματα ή τα αποθηκεύουν προσωρινά για έλεγχο από τους διαχειριστές δικτύου.

4. Συστήματα Ανίχνευσης και Αποτροπής Εισβολών (IDS/IPS)

Τα συστήματα αυτά μπορούν να εντοπίζουν ύποπτες δραστηριότητες σε πραγματικό χρόνο, αναγνωρίζοντας προσπάθειες παραβίασης των συστημάτων και, μέσω ανάλυσης πακέτων δεδομένων και ανίχνευσης μοτίβων συμπεριφοράς, αποτρέπουν τη συνέχιση επιθέσεων κοινωνικής μηχανικής, ειδικότερα όταν εμπλέκεται κακόβουλο λογισμικό. Αν και οι επιθέσεις κοινωνικής μηχανικής είναι συνήθως μη τεχνικές και βασίζονται στην ανθρώπινη αλληλεπίδραση, συχνά συνδέονται με τη χρήση κακόβουλου λογισμικού, όπως keyloggers (προγράμματα καταγραφής πληκτρολογήσεων που αποτυπώνουν τις πληκτρολογήσεις των χρηστών, κλέβοντας διαπιστευτήρια ή προσωπικές πληροφορίες) ή Trojan horses (κακόβουλα προγράμματα που “μεταμφιέζονται” ως νόμιμο λογισμικό και επιτρέπουν στους επιτιθέμενους να αποκτήσουν απομακρυσμένη πρόσβαση ή να κλέψουν δεδομένα από το θύμα), τα οποία εγκαθίστανται μετά από επιτυχημένες επιθέσεις τύπου phishing, vishing ή άλλων ειδών. Ένα παράδειγμα είναι η περίπτωση ενός χρήστη που έλαβε ένα phishing email με έναν κακόβουλο σύνδεσμο, οπότε εφόσον κάνει κλικ στον σύνδεσμο, αυτό μπορεί να οδηγήσει σε μια προσπάθεια εγκατάστασης κακόβουλου λογισμικού μέσω drive-by-download. Τα συστήματα θα ανιχνεύσουν την ασυνήθιστη δραστηριότητα (εγκατάστασης) και μπορούν να αποκλείσουν την προσπάθεια αυτή πριν το κακόβουλο λογισμικό εγκατασταθεί και προκαλέσει ζημιά στο σύστημα του χρήστη.

5. Διαχείριση Πληροφοριών και Συμβάντων Ασφάλειας (SIEM)

Η χρήση εργαλείων SIEM βοηθά στην καταγραφή και ανάλυση των δεδομένων ασφαλείας σε πραγματικό χρόνο. Τα συστήματα αυτά επιτρέπουν στους οργανισμούς να εντοπίζουν ανωμαλίες και να αντιδρούν άμεσα σε πιθανές επιθέσεις κοινωνικής μηχανικής, πριν αυτές προκαλέσουν ζημιές, ή να τις περιορίζουν. Για παράδειγμα, θα μπορούσε να σταλεί ένα email ηλεκτρονικού «ψαρέματος» σε έναν υπάλληλο, το οποίο τον οδηγεί σε έναν ψεύτικο ιστότοπο όπου πρέπει να εισάγει τα στοιχεία σύνδεσής του. Αυτά τα διαπιστευτήρια χρησιμοποιούνται από τον επιτιθέμενο για να εισέλθει στο δίκτυο της εταιρείας. Οι λύσεις SIEM (στην προηγούμενη περίπτωση αλλά και εν γένει σε πολλές περιπτώσεις) μπορούν να εντοπίζουν την εκάστοτε ανώμαλη δραστηριότητα, συμπεριλαμβανομένης της μαζικής αποστολής μηνυμάτων ηλεκτρονικού ταχυδρομείου ή της πρόσβασης από μη αναγνωρισμένους ιστότοπους, και να ειδοποιούν τους διαχειριστές, ώστε να μπορούν να αναλάβουν δράση προτού προκληθεί μεγαλύτερη ζημιά.

6. Περιορισμός Πρόσβασης και Έλεγχος Προνόμιων

Η αρχή της ελάχιστης δυνατής πρόσβασης (least privilege) πρέπει να εφαρμόζεται σε όλα τα επίπεδα ενός οργανισμού. Οι εργαζόμενοι πρέπει να έχουν πρόσβαση μόνο στα δεδομένα που χρειάζονται για την εκτέλεση των καθηκόντων τους. Αυτός ο περιορισμός μειώνει τον κίνδυνο όταν οι επιτιθέμενοι αποκτήσουν πρόσβαση μέσω ενός χαμηλόβαθμου υπαλλήλου διότι δεν θα μπορέσουν να προκαλέσουν σημαντικά ζημία στον οργανισμό. Επιπρόσθετα, όπως αναφέρθηκε και στην περίπτωση των αντιμέτρων για επιθέσεις κυβερνοασφάλειας, η χρήση μοντέλων πρόσβασης τύπου RBAC & ABAC μπορεί να δυσκολέψει τη μη εξουσιοδοτημένη πρόσβαση σε αγαθά του συστήματος σε έναν επιτιθέμενο που μόλις έχει διεισδύσει σε αυτό.

7. Ενημέρωση και Διορθώσεις Ευπαθειών

Οι οργανισμοί πρέπει να διαχειρίζονται ενεργά τις πληροφορίες τους, να παρακολουθούν τυχόν ευπάθειες και να διασφαλίζουν ότι όλα τα συστήματα και λογισμικά είναι πλήρως ενημερωμένα. Η τακτική ενημέρωση των συστημάτων και οι διορθώσεις των ευπαθειών ασφαλείας μειώνουν τις πιθανότητες εκμετάλλευσης από επιτιθέμενους. Να τονιστεί πως αυτό πρέπει να γίνεται όχι μόνο στα λογισμικά αλλά και στις εξαρτήσεις που αυτά εμπεριέχουν. Αν, επομένως, ένα λογισμικό χρησιμοποιεί έκδοση εξάρτησης που είναι τρωτή, θα πρέπει να μεταβαίνει σε μια νέα έκδοση της εξάρτησης όπου δεν εμφανίζονται πλέον οι σχετικές τρωτότητες. Αν δεν υπάρχει μια τέτοια νέα έκδοση, θα πρέπει να διερευνηθεί η χρήση μιας εναλλακτικής εξάρτησης.

Ακολουθεί ο Πίνακας 1 όπου πραγματοποιείται μια σύγκριση των αντιμέτρων με βάση τον άνθρωπο σε σχέση με τα αντίμετρα που βασίζονται στον υπολογιστή (δηλ. τα τεχνικά αντίμετρα).

Πίνακας 1: Σύγκριση αντιμέτρων με βάση τον άνθρωπο έναντι αντιμέτρων με βάση τον υπολογιστή. Πηγή: Salahdine & Kaabouch, 2019.

Αντίμετρα	Περιγραφή	Πλεονεκτήματα	Περιορισμοί
Βασισμένα στον άνθρωπο	Εκπαίδευση, Κατάρτιση, Ευαισθητοποίηση	- Εύκολη εκπαίδευση των ανθρώπων για το τι πρέπει να κάνουν - Μικρός αριθμός θυμάτων (στόχος αποτελεί η πλήρης εξάλειψη των θυμάτων μετά την κατάλληλη εκπαίδευση, κάνοντας τα άτομα ανθεκτικότερα στην συναισθηματική χειραγώγηση)	- Η απληστία (επιθυμία των ατόμων για κέρδος ή άλλα προσωπικά ωφέλη) δεν μπορεί εύκολα να καταπολεμιστεί - Ο επηρεασμός των ανθρώπινων αποφάσεων (όπου οι επιτιθέμενοι εκμεταλλεύονται ανθρώπινα συναισθήματα & ψυχολογικούς παράγοντες) δεν μπορεί εύκολα να μετριαστεί ή να εκμηδενιστεί
Βασισμένα σε υπολογιστή	Λογισμικό, συστήματα και εργαλεία	- Αποτελεσματικά (πολύ υψηλά επίπεδα προστασίας, με ταχύτητα & ακρίβεια που δεν είναι εύκολα εφικτή από τους ανθρώπους)	- Ακριβά προϊόντα (Ωστόσο, υπάρχουν και επιλογές ανοικτού κώδικα ή λύσεις που προσφέρονται ως υπηρεσίες υπολογιστικού νέφους, με μικρότερο κόστος σε σχέση με την αγορά παραδοσιακού λογισμικού) - Περιορισμοί στην ενοποίηση / χρήση εργαλείων - Απαιτούνται πρόσθετοι πόροι για την εγκατάσταση & εκτέλεσή τους - Μπορεί να υπάρχουν προβλήματα ασυμβατότητας μεταξύ εργαλείων - Περιορισμένα από την άγνοια των ανθρώπων - Τα άτομα θα πρέπει να εκπαιδευθούν ώστε να τα χρησιμοποιούν σωστά.

Ο ακρογωνιαίος λίθος της προστασίας των οργανισμών από επιθέσεις κοινωνικής μηχανικής είναι η χρήση προληπτικών μέτρων ασφαλείας. Σε αυτά περιλαμβάνονται η χρήση ισχυρών, μοναδικών κωδικών πρόσβασης που ενημερώνονται τακτικά και τα πρωτόκολλα διπλής αυθεντικοποίησης (2FA) που απαιτούν μια δεύτερη επιβεβαίωση ταυτότητας μέσω χρήσης βιομετρικών δεδομένων. Η αυστηρή διαχείριση δικαιωμάτων πρόσβασης εγγυάται περαιτέρω ότι μόνο τα άτομα με την κατάλληλη εξουσιοδότηση μπορούν να έχουν πρόσβαση σε ζωτικά συστήματα και δεδομένα. Οι συχνές ενημερώσεις λογισμικού και οι έλεγχοι ασφαλείας βοηθούν επιπλέον στην άμυνα κατά των ευπαθειών και των επιθέσεων που χρησιμοποιούν γνωστά ελαττώματα των συστημάτων. Σε συνδυασμό με τη γνώση και την εκπαίδευση των χρηστών, όλες αυτές οι προφυλάξεις παρέχουν ένα ισχυρό φράγμα κατά των επιθέσεων κοινωνικής μηχανικής.

Έτσι, η προστασία από επιθέσεις κοινωνικής μηχανικής απαιτεί μια ολοκληρωμένη προσέγγιση που συνδυάζει εκπαίδευση, τεχνολογικές λύσεις, και προληπτικά μέτρα ασφαλείας, όπως αυτά που προαναφέρθηκαν. Οι επιθέσεις αυτές είναι ιδιαίτερα επικίνδυνες καθώς βασίζονται στη χειραγώγηση του ανθρώπινου παράγοντα. Η συνεχής ευαισθητοποίηση και οι σωστές διαδικασίες, όπως τα πρωτόκολλα επαλήθευσης ταυτότητας και ο έλεγχος πρόσβασης, αποτελούν μόνο κάποιες από τις βέλτιστες πρακτικές ασφαλείας (security best practices), που μπορούν να ενισχύσουν σημαντικά την ασφάλεια ενός οργανισμού.

2.2.9 Ηθικές και νομικές διαστάσεις της κοινωνικής μηχανικής

Οι επιθέσεις κοινωνικής μηχανικής όχι μόνο παραβιάζουν την ασφάλεια των πληροφοριών, αλλά θέτουν επίσης ηθικά και νομικά ζητήματα που αφορούν την προσωπική και επαγγελματική ζωή των θυμάτων. Λαμβάνοντας υπόψιν τις νομικές διαστάσεις, οι επιθέσεις αυτές παραβιάζουν συχνά νομικούς κανόνες και νόμους που αφορούν την προστασία των δεδομένων, την κλοπή ταυτότητας και την παραβίαση της ιδιωτικής ζωής. Πολλές χώρες έχουν εφαρμόσει νομικά πλαίσια για να τιμωρούν τους δράστες των επιθέσεων κοινωνικής μηχανικής, ιδιαίτερα όταν αυτές οι επιθέσεις οδηγούν σε οικονομικές απώλειες ή παραβιάσεις ευαίσθητων πληροφοριών. Ωστόσο, σε περιπτώσεις όπου οι πελάτες ενός οργανισμού εμπλέκονται σε διαρροή δεδομένων, ο εν λόγω οργανισμός μπορεί επίσης να θεωρηθεί υπεύθυνος, ιδίως εφόσον αποδειχθεί ότι τα δεδομένα αυτά δεν προστατεύονταν σωστά. Η νομοθεσία περί προστασίας δεδομένων, συμπεριλαμβανομένου του Γενικού Κανονισμού για την Προστασία Δεδομένων (ΓΚΠΔ), επισύρει αυστηρές κυρώσεις για απρόσεκτη ή ανεπαρκή διαχείριση δεδομένων. Σε τέτοιες περιπτώσεις, οι οργανισμοί ενδέχεται να υπόκεινται σε πρόστιμα ή νομικές κυρώσεις.

Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR): Ο Γενικός Κανονισμός για την Προστασία Δεδομένων είναι ένας σχετικά πρόσφατος νόμος που ψηφίστηκε από την

Ευρωπαϊκή Επιτροπή τον Απρίλιο του 2016. Ο GDPR επιβάλλει αυστηρούς κανονισμούς σχετικά με την προστασία των προσωπικών δεδομένων στην Ευρωπαϊκή Ένωση. Οι οργανισμοί που δεν προστατεύουν επαρκώς τα δεδομένα τους ή τα δεδομένα ατόμων ή πελατών που διαχειρίζονται μπορεί να υποστούν σοβαρά πρόστιμα και νομικές συνέπειες. (Εικόνα 3)

Εικόνα 3: Βασικά στοιχεία του ΓΚΠΔ (Ισχύει πλήρως από τις 25 Μαΐου 2018). Πηγή: bitsnbytes.cybersecurity, 2018.

Οι οργανισμοί φέρουν μεγάλη νομική ευθύνη σε περίπτωση μη επαρκής εφαρμογής μέτρων ασφαλείας για την προστασία των δεδομένων από επιθέσεις κοινωνικής μηχανικής, ενώ μπορούν να θεωρηθούν υπεύθυνοι για την αμέλεια στην προστασία των προσωπικών δεδομένων των πελατών τους.

Ηθική Ευθύνη των Οργανισμών: Οι οργανισμοί έχουν την ηθική ευθύνη να διασφαλίζουν ότι οι υπάλληλοι και οι πελάτες τους είναι προστατευμένοι από τις απειλές της κοινωνικής μηχανικής. Αυτό περιλαμβάνει την ανάπτυξη και εφαρμογή προγραμμάτων εκπαίδευσης και την υιοθέτηση αυστηρών πρωτοκόλλων ασφαλείας. Η αδυναμία να ληφθούν προληπτικά μέτρα για την προστασία των θυμάτων μπορεί να οδηγήσει σε ηθικές και νομικές συνέπειες για τους οργανισμούς.

2.2.10 Εξελίξεις και μελλοντικές τάσεις στην κοινωνική μηχανική

Η εξάπλωση της απομακρυσμένης εργασίας, ιδίως μετά την πανδημία COVID-19, έχει οδηγήσει σε αυξημένη στόχευση απομακρυσμένων εργαζομένων (δηλ. εργαζομένων που δουλεύουν εξ αποστάσεως σε μια εταιρεία) μέσω επιθέσεων κοινωνικής μηχανικής. Οι

απομακρυσμένοι εργαζόμενοι συχνά εργάζονται εκτός των ελεγχόμενων περιβαλλόντων ασφαλείας των εταιρειών, κάνοντάς τους πιο ευάλωτους σε τέτοιου είδους επιθέσεις.

Οι βασικοί λόγοι που καθιστούν τους απομακρυσμένους εργαζόμενους ευάλωτους περιλαμβάνουν τα εξής:

- Αδύναμες Αμυντικές Υποδομές

Ενώ τα εταιρικά δίκτυα έχουν αυστηρά μέτρα ασφαλείας, τα οικιακά δίκτυα των εργαζομένων δεν διαθέτουν τα ίδια επίπεδα προστασίας. Στη προκειμένη περίπτωση (των οικιακών δικτύων), οι επιτιθέμενοι εκμεταλλεύονται τα μη ασφαλή δίκτυα Wi-Fi, συσκευές που δεν είναι πλήρως ενημερωμένες ή προστατευμένες, και αδύναμους κωδικούς πρόσβασης.

- Αυξημένη Χρήση Ψηφιακών Επικοινωνιών

Οι απομακρυσμένοι εργαζόμενοι βασίζονται περισσότερο σε ψηφιακά μέσα, όπως email, μηνύματα και τηλεδιάσκεψη, για τις καθημερινές τους εργασίες. Αυτή η αυξημένη ψηφιακή αλληλεπίδραση παρέχει περισσότερες ευκαιρίες για επιθέσεις phishing, πλαστοπροσωπίας και άλλες μορφές παραπλάνησης, όπως πλαστές προσκλήσεις για τηλεδιασκέψεις μέσω πλατφορμών, όπως το Zoom.

- Απομόνωση και Λιγότερη Υποστήριξη σε Πραγματικό Χρόνο

Οι απομακρυσμένοι εργαζόμενοι συχνά βρίσκονται απομονωμένοι από τις IT ομάδες υποστήριξης. Αυτό καθυστερεί την άμεση αναφορά ύποπτων δραστηριοτήτων και αυξάνει τον χρόνο αντίδρασης σε περιστατικά ασφαλείας. Η έλλειψη τακτικής φυσικής επικοινωνίας με συναδέλφους μπορεί επίσης να οδηγήσει σε χαμηλότερη ευαισθητοποίηση απέναντι στις απειλές.

- Εκμετάλλευση Ανθρώπινων Παραγόντων

Οι επιτιθέμενοι στοχεύουν το φόβο, την αβεβαιότητα και το αίσθημα επείγοντος που συχνά συνοδεύουν την απομακρυσμένη εργασία, ειδικά σε συνθήκες κρίσης. Οι εργαζόμενοι μπορεί να υποκύψουν σε ψυχολογική πίεση από τους επιτιθέμενους που προσποιούνται ότι είναι ανώτερα στελέχη, ζητώντας πληροφορίες ή ενέργειες που σε κανονικές συνθήκες θα αμφισβητούνταν.

Σήμερα, οι επιθέσεις κοινωνικής μηχανικής προσαρμόζονται και εκμεταλλεύονται αυτές τις νέες συνθήκες. Η έλλειψη φυσικής παρουσίας σε ασφαλή περιβάλλοντα γραφείου, σε συνδυασμό με την αυξανόμενη χρήση ψηφιακών εργαλείων, δημιουργούν νέες ευκαιρίες για τους επιτιθέμενους. Παράλληλα, η απομόνωση των απομακρυσμένων εργαζομένων, η

ελλιπής εκπαίδευση και η αδυναμία άμεσης υποστήριξης καθιστούν τις επιθέσεις αυτές ιδιαίτερα επιτυχημένες.

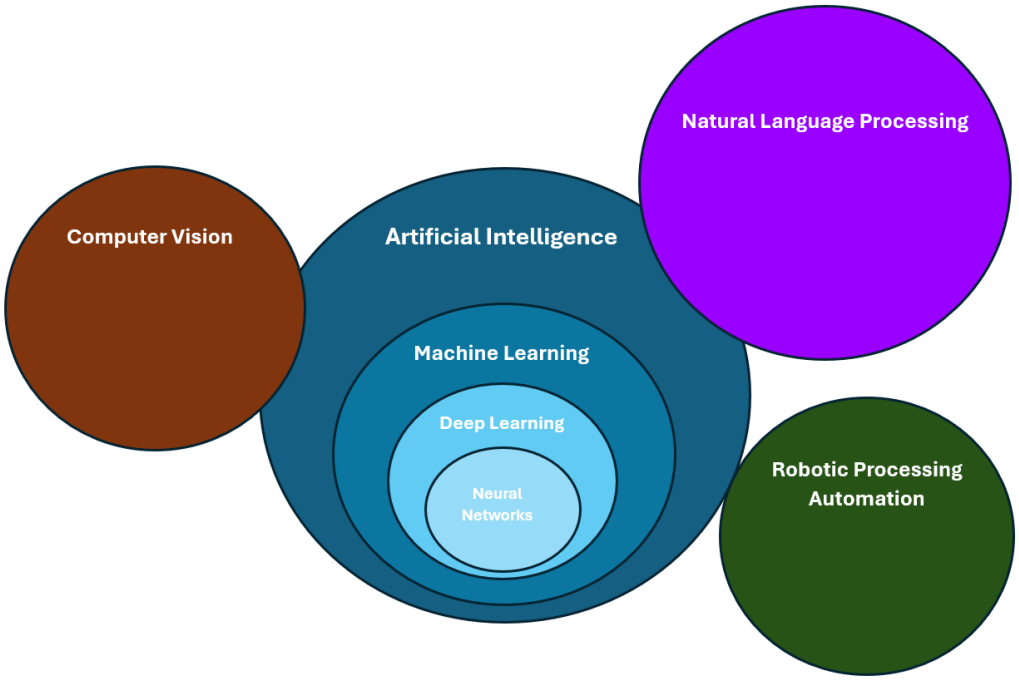
Στο μέλλον, οι επιθέσεις κοινωνικής μηχανικής αναμένεται να γίνουν ακόμα πιο σύνθετες, καθώς οι επιτιθέμενοι θα αξιοποιούν τις τεχνολογίες τεχνητής νοημοσύνης και αυτοματισμού για να προσαρμόζουν τις επιθέσεις τους σε πραγματικό χρόνο. Η αυξημένη στόχευση απομακρυσμένων εργαζομένων και η εκμετάλλευση των αδυναμιών σε οικιακά δίκτυα θα συνεχίσουν να αποτελούν απειλή για τους οργανισμούς. Ωστόσο, οι εξελίξεις στην κυβερνοασφάλεια, όπως η ανάπτυξη πιο προηγμένων εργαλείων παρακολούθησης, θα παίξει καθοριστικό ρόλο στην προστασία από αυτές τις απειλές.

2.3 Τεχνητή Νοημοσύνη & τεχνολογίες

2.3.1 Ορισμός Τεχνητής Νοημοσύνης

Η Τεχνητή Νοημοσύνη (ΤΝ) αποτελεί έναν τομέα της επιστήμης των υπολογιστών που εστιάζει στην ανάπτυξη συστημάτων και λογισμικού που επιδεικνύουν νοημοσύνη παρόμοια με εκείνη του ανθρώπου. Στόχος της ΤΝ είναι να επιτρέψει στις μηχανές να εκτελούν γνωστικές εργασίες, όπως η μάθηση, η αντίληψη, η επίλυση προβλημάτων και η λήψη αποφάσεων, χωρίς την άμεση ανθρώπινη παρέμβαση. Η ανάπτυξη της Τεχνητής Νοημοσύνης περιλαμβάνει τη χρήση μεθόδων, όπως η μηχανική μάθηση (machine learning), η οποία περιλαμβάνει υποκατηγορίες, όπως τα νευρωνικά δίκτυα και τα βαθιά νευρωνικά δίκτυα (deep learning), που επιτρέπουν στα συστήματα να βελτιώνουν τις επιδόσεις τους μέσω της εμπειρίας της συλλογής δεδομένων και της εκπαίδευσης μέσω αυτών (Russell, S., & Norvig, P. 2021).

2.3.2 Τεχνολογίες Τεχνητής Νοημοσύνης



Εικόνα 4: Μια οπτικοποίηση που δείχνει τα βασικά πεδία στον τομέα της Τεχνητής Νοημοσύνης.

Όπως φαίνεται από την Εικόνα 4:

- *Η τεχνητή νοημοσύνη*, είναι ο γενικός τομέας, με όλους τους άλλους κύκλους να αντιπροσωπεύουν πιο εξειδικευμένους τομείς, είτε εντός αυτού (ML, DL, NNs), είτε να τέμνονται-επικαλύπτονται με αυτόν (NLP, CV, RPA).
- *Η μηχανική μάθηση*, είναι ένα υποσύνολο της ΤΝ που επικεντρώνεται στα συστήματα που μαθαίνουν από δεδομένα.
- *Η βαθιά μάθηση*, είναι μια πιο εξειδικευμένη μορφή μηχανικής μάθησης, βασισμένη σε νευρωνικά δίκτυα.
- *Τα νευρωνικά δίκτυα*, αποτελούν τη ραχοκοκαλιά της βαθιάς μάθησης.
- *Οι τομείς NLP και Όραση Υπολογιστών* εφαρμόζουν τεχνικές Τεχνητής Νοημοσύνης, όπως Μηχανικής Μάθησης και Βαθιάς Μάθησης.
- *Ο τομέας RPA,* Αν και δεν είναι μέρος της ΤΝ, ενσωματώνει διάφορες τεχνικές της για την αυτοματοποίηση εργασιών. Έτσι, αφού δεν είναι υποκατηγορία της ΤΝ, εμφανίζεται στην εικόνα να τέμνεται με αυτήν, αλλά όχι να την επικαλύπτει.

Επομένως, με βάση την Εικόνα 4, η τεχνητή νοημοσύνη καλύπτει ένα ευρύ φάσμα τεχνολογιών, οι οποίες αναλύονται στη συνέχεια

1. *Μηχανική Μάθηση (Machine Learning - ML)*: Η Μηχανική Μάθηση (ΜΜ) είναι ένα υποσύνολο της ΤΝ που περιλαμβάνει την εκπαίδευση αλγορίθμων ώστε οι μηχανές να μαθαίνουν/εκπαιδεύονται από σύνολα δεδομένων και να πραγματοποιούν προβλέψεις ή να παίρνουν αποφάσεις χωρίς να προγραμματίζονται ρητά. Περιλαμβάνει τεχνικές, όπως η μάθηση με επίβλεψη (supervised learning), η μάθηση χωρίς επίβλεψη (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning) (Hadnagy, 2020).

Μάθηση με επίβλεψη (Supervised Learning):

Αποτελεί μέθοδο μηχανικής μάθησης όπου το σύστημα εκπαιδεύεται χρησιμοποιώντας δεδομένα εισόδου και τις αντίστοιχες (σωστές / αναμενόμενες) εξόδους τους, γνωστά ως ζεύγη εισόδου-εξόδου (input-output pairs). Στην εκπαίδευση αυτή, το μοντέλο μαθαίνει να συνδέει τα δεδομένα εισόδου με τις επιθυμητές εξόδους με βάση το σύνολο παραδειγμάτων (εκπαίδευσης). Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που μπορεί να προβλέψει σωστά την έξοδο για νέες, άγνωστες εισόδους. Οι πιο κοινές εργασίες μάθησης με επίβλεψη είναι η ταξινόμηση και η παλινδρόμηση, όπου τα δεδομένα εισόδου αντιστοιχίζονται σε διακριτές κατηγορίες ή σε συνεχείς τιμές, αντίστοιχα (Russell & Norvig, 2021). Η μάθηση με επίβλεψη μπορεί να χρησιμοποιηθεί σε διάφορους τομείς. Μάλιστα, πολλές φορές συνδυάζονται δύο ή παραπάνω αλγόριθμοι ΜΜ για την παραγωγή εξαιρετικά ακριβών προβλέψεων μέσω της χρήσης μεθόδων σύνθεσης μοντέλων/αλγορίθμων που ονομάζονται ensemble. Συνεπώς, έχουμε τις εξής εργασίες ή μεθόδους σύνθεσης που υποστηρίζονται από την μάθηση με επίβλεψη:

α. *Ταξινόμηση (Classification)*: Χρησιμοποιούνται αλγόριθμοι μάθησης με επίβλεψη, όπως Random Forest, Μηχανές διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) ή Μηχανές ενίσχυσης κλίσης (Gradient Boosting Machines - GBM), για εργασίες ταξινόμησης, όπως η ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, η ανάλυση συναισθήματος σε δεδομένα κειμένου ή η αναγνώριση εικόνων.

β. *Παλινδρόμηση (regression)*: Χρησιμοποιούνται αλγόριθμοι, όπως Γραμμική Παλινδρόμηση (Linear Regression), Δέντρα Αποφάσεων (Decision Trees) ή Νευρωνικά Δίκτυα (Neural Networks), για εργασίες παλινδρόμησης (δηλ, ουσιαστικά εκμάθησης ενός γραμμικού ή μη μοντέλου για την πρόβλεψη τιμών), όπως η πρόβλεψη τιμών κατοικιών με βάση χαρακτηριστικά όπως τετραγωνικά μέτρα, αριθμός υπνοδωματίων και τοποθεσία, η πρόβλεψη τιμών μετοχών ή η εκτίμηση εσόδων από πωλήσεις με βάση τις δαπάνες μάρκετινγκ (Rajasekharaiah et al., 2020).

γ. *Μέθοδοι Ensemble*: Εξερευνώνται μέθοδοι ensemble, όπως Bagging, Boosting (*πχ*. AdaBoost) ή Stacking, για να παράγουν ένα σύνθετο μοντέλο με βελτιωμένη απόδοση, συνδυάζοντας προβλέψεις από πολλά βασικά μοντέλα (μηχ. μάθησης που έχουν παραχθεί με

επίβλεψη). Αυτές οι τεχνικές είναι ιδιαίτερα χρήσιμες για πολύπλοκα σύνολα δεδομένων ή όταν τα μεμονωμένα μοντέλα έχουν ασυνεπή ή μη ικανοποιητική απόδοση (Rajasekharaiah et al., 2020).

Μάθηση χωρίς επίβλεψη (Unsupervised Learning):

Η μάθηση χωρίς επίβλεψη είναι μια τεχνική ΜΜ στην οποία το σύστημα αποκτά γνώση από δεδομένα που είτε δεν έχουν σαφή ζεύγη εισόδου-εξόδου (προς χρήση στην εκπαίδευση), είτε δεν είναι επισημασμένα. Χωρίς αναφορά σε ετικέτες ή στόχους, ο αλγόριθμος αναζητά στα δεδομένα μοτίβα, συσχετίσεις ή δομές. Όταν η υιοθέτηση της παλινδρόμησης ή της ταξινόμησης δεν είναι εφικτή (λόγω των προαναφερόμενων ιδιαιτεροτήτων των δεδομένων) και οι στόχοι της ανάλυσης είναι περισσότερο διερευνητικοί -όπως ο εντοπισμός συστάδων (clusters) ή η αποκάλυψη κρυφών συνδέσμων στα δεδομένα-, αυτό το είδος ΜΜ χρησιμοποιείται συχνά (Murphy, 2012). Με βάση αυτή την προσέγγιση, η μάθηση χωρίς επίβλεψη εφαρμόζεται σε διάφορους τομείς και περιλαμβάνει τεχνικές, όπως η ομαδοποίηση, η μείωση διαστάσεων και η μάθηση κανόνων συσχέτισης:

α. *Ομαδοποίηση (Clustering)*: Χρησιμοποιούνται αλγόριθμοι, όπως K-Means, Ιεραρχική ομαδοποίηση (hierarchical clustering), ή DBSCAN, για εργασίες ομαδοποίησης (clustering) (δεδομένων), όπως η τμηματοποίηση πελατών, η ανίχνευση ανωμαλιών στην κίνηση δικτύου, ή η ομαδοποίηση παρόμοιων εγγράφων στην ανάλυση κειμένου.

β. *Μείωση διαστάσεων (dimensionality reduction)*: Εφαρμόζονται τεχνικές, όπως η ανάλυση κύριων συνιστωσών (Principal Component Analysis - PCA), η t-διανεμημένη στοχαστική ενσωμάτωση γειτόνων (t-SNE) ή οι αποκωδικοποιητές, για τη μείωση των διαστάσεων των δεδομένων, όταν αυτές είναι πολλές σε αριθμό, διατηρώντας παράλληλα σημαντικές πληροφορίες. Αυτό μπορεί να είναι χρήσιμο για την οπτικοποίηση, την εξαγωγή χαρακτηριστικών ή την επιτάχυνση μεταγενέστερων εργασιών ΜΜ.

γ. *Μάθηση κανόνων συσχέτισης (Association Rule Learning)*: Εξερευνώνται αλγόριθμοι, όπως ο Apriori ή ο FP-Growth, για την εξόρυξη κανόνων συσχέτισης από δεδομένα (πχ. συναλλαγών). Αυτές οι τεχνικές χρησιμοποιούνται συνήθως στην ανάλυση καλαθιού αγοράς για την ανακάλυψη μοτίβων όπως "οι άνθρωποι που αγοράζουν το προϊόν Α είναι πιο πιθανό να αγοράσουν και το προϊόν Β" (Rajasekharaiah et al., 2020).

Ενισχυτική μάθηση (Reinforcement learning):

Αποτελεί υποπεδίο της μηχανικής μάθησης που ασχολείται με τη βελτιστοποίηση των αποφάσεων ενός πράκτορα (agent) μέσω αλληλεπίδρασης με το περιβάλλον του. Ο πράκτορας μαθαίνει να επιλέγει ενέργειες που μεγιστοποιούν την αθροιστική ανταμοιβή με την πάροδο του χρόνου, βασιζόμενος σε μη καθορισμένη εκ των προτέρων πληροφορία για τη σωστή ενέργεια. Η διαδικασία αυτή βασίζεται στην εξερεύνηση και εκμετάλλευση του περιβάλλοντος, όπου ο πράκτορας ακολουθεί στρατηγικές με σκοπό να βελτιώνει συνεχώς

την απόδοσή του. Εφαρμογές της ενισχυτικής μάθησης περιλαμβάνουν τον έλεγχο ρομποτικών συστημάτων, τη βελτιστοποίηση διαδικασιών και τη στρατηγική λήψη αποφάσεων σε παιχνίδια (Sutton & Barto, 2018; Kober et al., 2013). Με αυτή τη μεθοδολογία, η ενισχυτική μάθηση χρησιμοποιείται για την εκπαίδευση πρακτόρων σε περιβάλλοντα με στόχο την εκμάθηση βέλτιστων στρατηγικών, μέσω τεχνικών όπως το Q-learning, τα βαθιά δίκτυα Q (DQN) και οι μέθοδοι πολιτικής κλίσης, οι οποίες αναλύονται παρακάτω:

α. *Q-Learning*: Εφαρμογή αλγορίθμων Q-learning για την εκπαίδευση πρακτόρων σε περιβάλλοντα με διακριτούς χώρους καταστάσεων και ενεργειών, όπως το παίξιμο επιτραπέζιων παιχνιδιών (*πχ*. Tic-Tac-Toe, σκάκι) ή η πλοήγηση σε περιβάλλοντα gridworld. Αυτή η προσέγγιση ΤΝ επιτρέπει στους πράκτορες να μαθαίνουν να λαμβάνουν αποφάσεις βασισμένες σε εμπειρία, βελτιώνοντας την ικανότητά τους να μεγιστοποιούν ανταμοιβές.

β. *Βαθιά δίκτυα Q (DQN)*: Χρησιμοποιούνται τα Deep Q-Networks, τα οποία συνδυάζουν την εκμάθηση Q (Q-learning) με βαθιά νευρωνικά δίκτυα, για να αντιμετωπίσουν πιο σύνθετες εργασίες ενισχυτικής μάθησης, όπως η αναπαραγωγή παιχνιδιών Atari από ακατέργαστες εισόδους εικονοστοιχείων ή ο έλεγχος ρομποτικών συστημάτων σε προσομοιωμένα περιβάλλοντα (Shackleford, 2020). Τα DQN αποτελούν παράδειγμα χρήσης της ΤΝ για την κατανόηση πολύπλοκων σχέσεων στα δεδομένα και την απόκτηση προσαρμοστικών στρατηγικών.

γ. *Μέθοδοι κλίσης πολιτικής (Policy Gradient)*: Διερευνώνται μέθοδοι κλίσης πολιτικής, όπως η REINFORCE ή η βελτιστοποίηση εγγύς πολιτικής (PPO), για την εκπαίδευση πρακτόρων σε περιβάλλοντα με συνεχείς χώρους δράσης, όπως ο ρομποτικός έλεγχος ή η αυτόνομη πλοήγηση οχημάτων. Αυτοί οι αλγόριθμοι ΤΝ μαθαίνουν απευθείας μια πολιτική για τη μεγιστοποίηση των αναμενόμενων ανταμοιβών χωρίς ρητό υπολογισμό συναρτήσεων αξίας, αυξάνοντας την ικανότητα των πρακτόρων να προσαρμόζονται σε δυναμικά περιβάλλοντα.

2. *Βαθιά μάθηση (Deep Learning - DL)*: Η βαθιά μάθηση είναι ένας τύπος ΜΜ που χρησιμοποιεί νευρωνικά δίκτυα (neural networks) με πολλαπλά επίπεδα (δηλ. βαθιά νευρωνικά δίκτυα) για την εξαγωγή σύνθετων μοτίβων και χαρακτηριστικών από δεδομένα. Είναι ιδιαίτερα αποτελεσματική για εργασίες, όπως η αναγνώριση εικόνας και ομιλίας, η επεξεργασία φυσικής γλώσσας (NLP) και η Υπολογιστική Όραση, , καθώς οι αλγόριθμοι ΤΝ μαθαίνουν να αναγνωρίζουν σύνθετες σχέσεις που είναι δύσκολο να ανιχνευτούν με παραδοσιακές μεθόδους (Shackleford, 2020).

3. *Επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP)*: Η NLP επικεντρώνεται στην αλληλεπίδραση μεταξύ υπολογιστών και ανθρώπινης γλώσσας. Περιλαμβάνει εργασίες, όπως η ανάλυση κειμένου, η ανάλυση συναισθήματος, η γλωσσική μετάφραση κι η αναγνώριση ομιλίας. Η χρήση ΤΝ στην NLP επιτρέπει την κατανόηση και παραγωγή ανθρώπινης γλώσσας με αυξανόμενη ακρίβεια (Wang et al., 2021).

4. *Υπολογιστική Όραση (Computer Vision)*: Η Υπολογιστική Όραση αφορά τη δυνατότητα των μηχανών να ερμηνεύουν και να κατανοούν οπτικές πληροφορίες από εικόνες ή βίντεο. Αυτό περιλαμβάνει την ανίχνευση αντικειμένων, την ταξινόμηση εικόνων, την αναγνώριση προσώπου και την κατάτμηση εικόνων, επιτρέποντας στα συστήματα να αναλύουν και να ερμηνεύουν οπτικά δεδομένα σε πραγματικό χρόνο (Wang et al., 2021).

5. *Ρομποτική (Robotics)*: Η τεχνητή νοημοσύνη χρησιμοποιείται στη ρομποτική για την ανάπτυξη αυτόνομων ρομπότ που μπορούν να εκτελούν καθήκοντα σε διάφορους τομείς, όπως η μεταποίηση, η υγειονομική περίθαλψη, η γεωργία, η εφοδιαστική αλυσίδα και η εξερεύνηση του διαστήματος. Με τη χρήση αλγορίθμων ΤΝ, όπως η ενισχυτική μάθηση (RL) και τα CNNs, τα ρομπότ μπορούν να βελτιώσουν την απόδοσή τους σε σύνθετα και δυναμικά περιβάλλοντα, επεκτείνοντας τις δυνατότητες αυτοματισμού και αυτονομίας τους (Hadnagy, 2020).

2.3.3 Πλεονεκτήματα και Μειονεκτήματα Τεχνητής Νοημοσύνης

Παρακάτω παραθέτουμε τα κύρια πλεονεκτήματα και μειονεκτήματα της ΤΝ, προσδιορίζοντας και την επίδρασή τους στο πλαίσιο της κυβερνοασφάλειας. Η ανάλυση περιλαμβάνει τόσο τα οφέλη της υιοθέτησης της ΤΝ, συμπεριλαμβανομένης της αναγνώρισης προτύπων και της αυτοματοποίησης της ασφάλειας, όσο και τα μειονεκτήματά της, όπως το ενδεχόμενο ψευδών ειδοποιήσεων και η περίπλοκη διαχείριση (Prianto, Y., Sumantri, V. K., & Sasmita, P. Y., 2020).

Κυριότερα Πλεονεκτήματα ΤΝ:

- *Αυτοματοποιημένη ανίχνευση και πρόληψη (Automated Detection and Prevention):* όπως κακόβουλα μηνύματα ηλεκτρονικού ταχυδρομείου ή ύποπτη συμπεριφορά χρηστών, κυρίως μέσω αλγορίθμων μηχανικής μάθησης και ανίχνευσης ανωμαλιών.
- *Αντιδράσεις σε πραγματικό χρόνο (Real-time Responses)*: Όταν η τεχνητή νοημοσύνη ενσωματώνεται στα συστήματα ασφαλείας, μπορεί να προσφέρει γρήγορη ανάλυση και δράση που μειώνει τον αντίκτυπο των απειλών.
- *Εξατομίκευση και επεκτασιμότητα (Customization and Scalability)*: Τα προγράμματα AI είναι σε θέση να προσαρμόζουν τα προληπτικά μέτρα στις απαιτήσεις μιας εταιρείας και να επεκτείνονται με ευκολία.

Κυριότερα Μειονεκτήματα ΤΝ:

- *Μεροληψία και ψευδώς θετικά αποτελέσματα (Bias and False Positives):* Τα συστήματα τεχνητής νοημοσύνης είναι επιρρεπή σε προκαταλήψεις που μπορεί να οδηγήσουν σε λανθασμένα συμπεράσματα ή ψευδώς θετικά αποτελέσματα, μειώνοντας έτσι την αποτελεσματικότητα του συστήματος.

- *Υπερβολική εξάρτηση από την αυτοματοποίηση (Over-reliance on Automation):* Η εξάρτηση μόνο από την τεχνητή νοημοσύνη μπορεί να προκαλέσει μείωση της ανθρώπινης επίγνωσης και υπερβολική εξάρτηση από τις μηχανές.
- *Εκλεπτυσμός των επιθέσεων (Sophistication of Attacks):* Οι επιτιθέμενοι μπορούν να χρησιμοποιήσουν την τεχνητή νοημοσύνη (AI) για να ενισχύσουν τις επιθέσεις κοινωνικής μηχανικής με τη δημιουργία πιο ρεαλιστικών ηλεκτρονικών μηνυμάτων ηλεκτρονικού «ψαρέματος» ή ψευδών προσωπικοτήτων.

2.4 Τεχνητή Νοημοσύνη σε Σχέση με την Κοινωνική Μηχανική

Στην ενότητα αυτή, θα εξεταστεί η σχέση μεταξύ των επιθέσεων κοινωνικής μηχανικής και της τεχνητής νοημοσύνης. Ειδικότερα, θα μιλήσουμε και για τις μελλοντικές προοπτικές της ΤΝ, όπως η υπερνοημοσύνη (superintelligence), η πρωτοτυπία (singularity) και η έννοια του υπερανθρωπισμού (transhumanism), και πως αυτές μπορούν να επιδράσουν στις ικανότητες των επιθέσεων κοινωνικής μηχανικής και στον βαθμό αντιμετώπισής τους. Ένα ακόμα σημαντικό και επίκαιρο κομμάτι που θα διερευνήσουμε είναι οι σύγχρονες εφαρμογές της ΤΝ, με έμφαση στη γενεσιουργό/παραγωγική τεχνητή νοημοσύνη (generative AI) και τα μεγάλα γλωσσικά μοντέλα (LLMs), που όχι μόνο μπορούν να υποστηρίξουν την ανάλυση και πρόληψη επιθέσεων, αλλά επίσης θέτουν νέες προκλήσεις όσον αφορά τη χρήση τους από επιτιθέμενους.

Πριν προχωρήσουμε στην ανάλυση της εφαρμογής της ΤΝ στην κοινωνική μηχανική, ειδικότερα υπό το πρίσμα των τελευταίων της τεχνολογικών εξελίξεων, είναι σημαντικό να εξετάσουμε κάποιες θεμελιώδεις έννοιες που αφορούν την εξέλιξή της και τις πιθανές μελλοντικές της επιπτώσεις (Krüger, O., 2021):

- Με τον όρο «υπερνοημοσύνη» (superintelligence), περιγράφεται η ικανότητας της τεχνητής νοημοσύνης να ξεπερνά πλήρως την ανθρώπινη νοημοσύνη σε όλες τις εκφάνσεις της. Μπορεί να δημιουργηθεί για να προβλέψει και να σταματήσει κάθε απόπειρα επίθεσης στον τομέα της κοινωνικής μηχανικής, αλλά αν δεν γίνει σωστή διαχείριση, υπάρχει περίπτωση να γίνει υπερβολικά ισχυρή. Από την άλλη μεριά, η «υπερνοημοσύνη» μπορεί να οδηγήσει σε πιο εξελιγμένες επιθέσεις κοινωνικής μηχανικής που δεν μπορούν να αποτραπούν ή να αντιμετωπιστούν από οποιαδήποτε εκπαίδευση και να έχει γίνει πάνω στα άτομα ή εξέλιξη σε μηχανισμούς ασφάλειας.
- Με τον όρο «μοναδικότητα» (singularity), περιγράφεται η χρονική στιγμή κατά την οποία σημαντικές και μόνιμες αλλαγές στην ανθρωπότητα και τον πολιτισμό θα επέλθουν από την τεχνική πρόοδο, ιδίως στην τεχνητή νοημοσύνη, που θα οδηγήσει στην αυτοβελτίωση των μηχανών. Μάλιστα, θα είναι τόσο υψηλός ο ρυθμός της

προόδου ώστε θα πραγματοποιηθούν σημαντικές μετασχηματιστικές αλλαγές στον πολιτισμό όπου μελλοντικές εξελίξεις θα είναι τελείως απρόβλεπτες. Αυτό θα μπορούσε να συνεπάγεται την πλήρη ασφάλεια στην κοινωνική μηχανική καθώς και την ανάπτυξη ασύλληπτα ισχυρών και επικίνδυνων επιθέσεων.

- Με τον όρο «υπερανθρωπισμός» (transhumanism), περιγράφεται η εφαρμογή της τεχνολογίας για τη βελτίωση του ανθρώπινου δυναμικού (δηλ. των ανθρώπινων δυνατοτήτων μέσω συνένωσης του ανθρώπου με τις μηχανές). Από τη σκοπιά της κοινωνικής μηχανικής, ο υπερανθρωπισμός μπορεί να ενισχύσει την ανθρώπινη νοημοσύνη και την άμυνα κατά των κακόβουλων προθέσεων, από διαφορετική οπτική όμως θέτει επίσης νέους κινδύνους για την ασφάλεια (λόγω της ισχυροποίησης των επιτιθέμενων).

Η χρήση της τεχνητής νοημοσύνης στην κοινωνική μηχανική επικεντρώνεται σήμερα τόσο σε επιθετικές όσο και σε αμυντικές στρατηγικές. Η ΤΝ χρησιμοποιείται αμυντικά στην ανίχνευση ανωμαλιών και στην ανάλυση της συμπεριφοράς των χρηστών σε πραγματικό χρόνο για τον εντοπισμό αμφισβητούμενων/κακόβουλων δραστηριοτήτων. Η ΤΝ, για παράδειγμα, μπορεί να φιλτράρει τις επιβλαβείς πληροφορίες από τα μηνύματα ηλεκτρονικού ταχυδρομείου και να παρακολουθεί για μοτίβα που παραπέμπουν σε απόπειρες phishing (Aldawood & Skinner, 2018). Ωστόσο, η ΤΝ μπορεί να αξιοποιηθεί επίσης από τους επιτιθέμενους για τη δημιουργία πειστικών & προηγμένων επιθέσεων κοινωνικής μηχανικής, συμπεριλαμβανομένων των deepfakes και του spear-phishing, καθιστώντας δυσκολότερη τη διάκριση μεταξύ νόμιμων και κακόβουλων επικοινωνιών (Aldawood, H. A., & Skinner, G.,2018)

Μία από τις πιο πρόσφατες και σημαντικές εξελίξεις στον τομέα της ΤΝ είναι η γενεσιουργός/παραγωγική τεχνητή νοημοσύνη (Generative AI), η οποία παράγει περιεχόμενο που μιμείται αυτό του ανθρώπου. Η εξέλιξη μεγάλων γλωσσικών μοντέλων (LLMs), όπως το GPT-3 (Generative Pretrained Transformer 3) και το GPT-4 (Generative Pretrained Transformer 4), δηλαδή μοντέλων που χρησιμοποιούν τεχνικές βαθιάς μάθησης και τεράστιες βάσεις δεδομένων κειμένων για να μάθουν μοτίβα και σχέσεις στη γλώσσα, έχει και εδώ διττό ρόλο στον τομέα της κοινωνικής μηχανικής. Από τη μία πλευρά, διευκολύνει τους επιτιθέμενους να διεξάγουν πιο πειστικές και εξατομικευμένες επιθέσεις σε μεγαλύτερη κλίμακα, ενώ από την άλλη παρέχει ισχυρά εργαλεία για την ενίσχυση των αμυντικών μηχανισμών.

Τα μεγάλα γλωσσικά μοντέλα (LLM) έχουν τη δυνατότητα να παράγουν περιεχόμενο που μοιάζει πολύ με την ανθρώπινη γλώσσα, γεγονός που δημιουργεί σημαντικές τεχνικές και ηθικές ανησυχίες για το μέλλον. Η διάδοση ψεύτικων πληροφοριών αποτελεί μείζον πρόβλημα, καθώς τα LLM μπορούν να παρέχουν δεδομένα που είναι ανακριβή αλλά και πειστικά, υποστηρίζοντας επιβλαβείς ή δόλιες πληροφορίες. Επιπλέον, η δημιουργία ψεύτικου περιεχομένου μέσω των LLM -γνωστή ως deepfakes- εγείρει ανησυχίες σχετικά με την ακεραιότητα της ταυτότητας και την παραβίαση της ιδιωτικής ζωής. Ακόμα, καθώς τα μοντέλα μπορεί να επαναλαμβάνουν τις προκαταλήψεις που βρέθηκαν στα δεδομένα εκπαίδευσής

τους, η έλλειψη διαφάνειας και κατανόησης σχετικά με τον τρόπο λήψης αποφάσεων από τα LLM ενδέχεται να οδηγήσει σε προβλήματα δικαιοσύνης και διακρίσεων. Τέλος, εξακολουθεί να υπάρχει αβεβαιότητα σχετικά με το ποιος είναι υπεύθυνος για τη χρήση αυτών των μοντέλων και την εποπτεία της ηθικής εφαρμογής τους, γεγονός που καθιστά αναγκαία την απαίτηση για αυστηρότερες πολιτικές και κανονισμούς που θα διασφαλίζουν τη δίκαιη και υπεύθυνη χρήση τους.

Κεφάλαιο 3: Μεθοδολογία Βιβλιογραφικής Επισκόπησης

Σε αυτό το κεφάλαιο, παρουσιάζεται ο τρόπος με τον οποίο μεθοδολογικά υλοποιήθηκε η συστηματική ανασκόπηση (Systematic Literature Review - SLR). Αρχικά, τέθηκαν ορισμένα βασικά ερευνητικά ερωτήματα που ταίριαζαν με τον σκοπό της έρευνας. Αυτά τα ερωτήματα καθοδήγησαν την αναζήτηση των σχετικών επιστημονικών άρθρων και τη μελέτη τους καθώς και υπογράμμισαν τα σημαντικά σημεία στα οποία θα έπρεπε να εστιάσει η ανάλυση. Στη συνέχεια, παρατίθενται οι πηγές και οι λέξεις-κλειδιά που χρησιμοποιήθηκαν για την συγκέντρωση του σωστού υλικού. Ακολούθως, ορίζονται τα κριτήρια για τη συμπερίληψη και το φιλτράρισμα άρθρων. Μέσω αυτών των κριτηρίων, φιλτράρονται όσα άρθρα είναι εκτός της εμβέλειας της ανασκόπησης και δεν πληρούν τα πρότυπα αξιόπιστης, επιστημονικά τεκμηριωμένης έρευνας. Τέλος, παρέχεται μια ποσοτική ανάλυση όσον αφορά τα άρθρα που τελικά επιλέχθηκαν.

3.1 Ερευνητικά Ερωτήματα

Στόχος της παρούσας μελέτης είναι να εξεταστούν και να συγκριθούν οι τρόποι με τους οποίους η Τεχνητή Νοημοσύνη (ΤΝ) μπορεί να υποστηρίξει την ανίχνευση, αποτροπή, και μετρίαση επιθέσεων κοινωνικής μηχανικής. Εστιάζοντας στην κατανόηση του προβλήματος και καθοδηγώντας την έρευνα, τίθενται κάποια κρίσιμα ερωτήματα προς απάντηση.

- **Ερώτημα 1**: Ποια είναι τα αντίμετρα βασιζόμενα στην ΤΝ που χρησιμοποιούνται για την πρόληψη και αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής και πως μπορούν να κατηγοριοποιηθούν;
- **Ερώτημα 2**: Ποιες τεχνικές/μέθοδοι ΤΝ χρησιμοποιούνται κατά κόρον σε τέτοια αντίμετρα;
- **Ερώτημα 3**: Ποια αντίμετρα μπορούν να θεωρηθούν ως τα καλύτερα και με βάση ποια κριτήρια (σύγκρισης/αξιολόγησης);
- **Ερώτημα 4**: Ποιες προκλήσεις και περιορισμοί υπάρχουν στην εφαρμογή της ΤΝ για την ανίχνευση και αποτροπή επιθέσεων κοινωνικής μηχανικής;
- **Ερώτημα 5**: Ποιες μελλοντικές εξελίξεις και πολλά υποσχόμενες ερευνητικές κατευθύνσεις στην ΤΝ ή στον τρόπο που αυτή γίνεται εκμεταλλεύσιμη αναμένονται να βελτιώσουν την ικανότητα των συστημάτων ασφαλείας να μετριάσουν/αντιμετωπίσουν επιθέσεις κοινωνικής μηχανικής;

Με βάση αυτά τα ερωτήματα, η έρευνα θα εξετάσει την αποτελεσματικότητα και αποδοτικότητα των αντιμέτρων βασιζόμενων στην Τεχνητή Νοημοσύνη, ενώ θα αναλύσει τους

περιορισμούς και τις προκλήσεις που σχετίζονται με τη χρήση αυτών των τεχνολογιών για την προστασία από τις επιθέσεις κοινωνικής μηχανικής.

3.2 Προσδιορισμός Πηγών Αναζήτησης

Για τη διεξαγωγή της βιβλιογραφικής επισκόπησης σχετικά με την εφαρμογή της Τεχνητής Νοημοσύνης στην ανίχνευση και αντιμετώπιση επιθέσεων κοινωνικής μηχανικής, αναπτύχθηκε μια στρατηγική αναζήτησης που περιλαμβάνει τη χρήση κατάλληλων λέξεων-κλειδιών. Σκοπός αυτής της στρατηγικής ήταν η ακριβής στόχευση σε εξειδικευμένα αποτελέσματα που αφορούν τις πιο σύγχρονες τεχνικές, μεθόδους και αλγοριθμικές προσεγγίσεις που αξιοποιούν την ΤΝ στον τομέα της κυβερνοασφάλειας, και συγκεκριμένα για την αντιμετώπιση επιθέσεων κοινωνικής μηχανικής. Η αναπτυγμένη συμβολοσειρά αναζήτησης περιλαμβάνει τις εξής λέξεις-κλειδιά: “(Artificial Intelligence | AI | Machine Learning | ML) & (Social Engineering | Phishing | Pretexting | Impersonation | Piggybacking | Quid Pro Quo | Tailgating | Smishing | Catfishing) & (Scanning | Spotting | Monitoring | Detection | Avoidance | Obviation | Prevention | Mitigation | Alleviation | Reduction)”.

Η παραπάνω αναζήτηση πραγματοποιήθηκε σε έγκυρες και παγκόσμια αναγνωρισμένες ψηφιακές βιβλιοθήκες και επιστημονικές βάσεις δεδομένων, όπως οι **ScienceDirect**, **SpringerOpen**, **IEEE Xplore**, και **Google Scholar**. Οι βάσεις αυτές παρέχουν επιστημονικά τεκμηριωμένες πηγές που διασφαλίζουν την εγκυρότητα και αξιοπιστία των δεδομένων που χρησιμοποιούνται στην παρούσα μελέτη. Με αυτήν τη διαδικασία αναζήτησης, διασφαλίστηκε η συλλογή δεδομένων που συμβάλλουν ουσιαστικά στην κατανόηση των αντιμέτρων βασιζόμενων στην ΤΝ και των εξελήξεων τους για την ανίχνευση, πρόληψη και μετριασμό επιθέσεων κοινωνικής μηχανικής.

3.3 Κριτήρια Ένταξης και Αποκλεισμού Άρθρων

Η διαδικασία επιλογής των άρθρων που χρησιμοποιούνται για τη βιβλιογραφική επισκόπηση είναι ιδιαίτερα κρίσιμη για την εγκυρότητα της έρευνας. Για την αποφυγή υπερβολικά μεγάλου όγκου δεδομένων και για να διασφαλιστεί η ποιότητα των πηγών, εφαρμόστηκαν συγκεκριμένα κριτήρια ένταξης / αποκλεισμού και ποιότητας, τα οποία προσδιορίζονται ως εξής:

Κριτήρια Ένταξης / Αποκλεισμού:

1. **Πρόσβαση στο πλήρες κείμενο**: Προκειμένου να διευκολυνθεί η ολοκληρωμένη αξιολόγησή τους, επιλέγονται μόνο άρθρα που είναι πλήρως προσβάσιμα για ανάγνωση και ανάλυση.
2. **Γλωσσική επιλογή**: Προτιμώνται άρθρα που έχουν δημοσιευτεί στην αγγλική ή την ελληνική γλώσσα, ώστε να διασφαλίζεται η κατανόηση του περιεχομένου και η ορθή αξιοποίηση των δεδομένων από τους συγγραφείς της αναφοράς.
3. **Σύγχρονη βιβλιογραφία**: Επιλέγονται άρθρα που κυκλοφόρησαν μετά το 2015, ώστε να διασφαλιστεί ότι χρησιμοποιούνται σύγχρονα δεδομένα και τεχνολογίες, ιδίως στους τομείς της κοινωνικής μηχανικής και της τεχνητής νοημοσύνης.
4. **Συνάφεια με το θέμα**: Η περίληψη και το περιεχόμενο αξιολογούνται για να επιβεβαιωθεί η συνάφεια του άρθρου με τα ερευνητικά ερωτήματα και το θέμα/αντικείμενο της μελέτης.
5. **Αξιόπιστες πηγές**: Η συμπερίληψη επιστημονικών άρθρων που είτε έχουν δημοσιευτεί σε περιοδικά ή σε συνέδρια με κριτές είτε αποτελούν διδακτορικές διατριβές εγγυάται τόσο την αξιοπιστία των δεδομένων όσο και την επιστημονική εγκυρότητά τους.
6. **Μοναδικότητα πηγών**: Άρθρα που εντοπίζονται επανειλημμένα σε διαφορετικές βάσεις δεδομένων κατά τη διάρκεια της έρευνας ή άρθρα με διαφορετικούς τίτλους, αλλά παρόμοιο περιεχόμενο και ίδιους συγγραφείς, αποκλείονται, καθώς θεωρούνται πως περιγράφουν την ίδια συνεισφορά. Έτσι, αποφεύγεται η πολλαπλή χρήση των ίδιων δεδομένων και εξασφαλίζεται η μοναδικότητα των πηγών.

Κριτήρια Ποιότητας:

1. **Ανεπαρκές περιεχόμενο**: Άρθρα που δεν παρέχουν επαρκή πληροφορία ή δεν καλύπτουν το θέμα σε βάθος, με κανένα είδος αποτίμησης/αξιολόγησης (review/assessment) της συνεισφοράς τους αποκλείονται.
2. **Ποιότητα περιεχομένου**: Απορρίπτονται άρθρα με δυσνόητο και κακώς δομημένο περιεχόμενο.
3. **Αξιοπιστία μεθοδολογίας**: Άρθρα που δεν τεκμηριώνουν σαφώς τη μεθοδολογία τους ή δεν παρέχουν επαρκή δεδομένα για την αναπαραγωγή των αποτελεσμάτων τους απορρίπτονται.

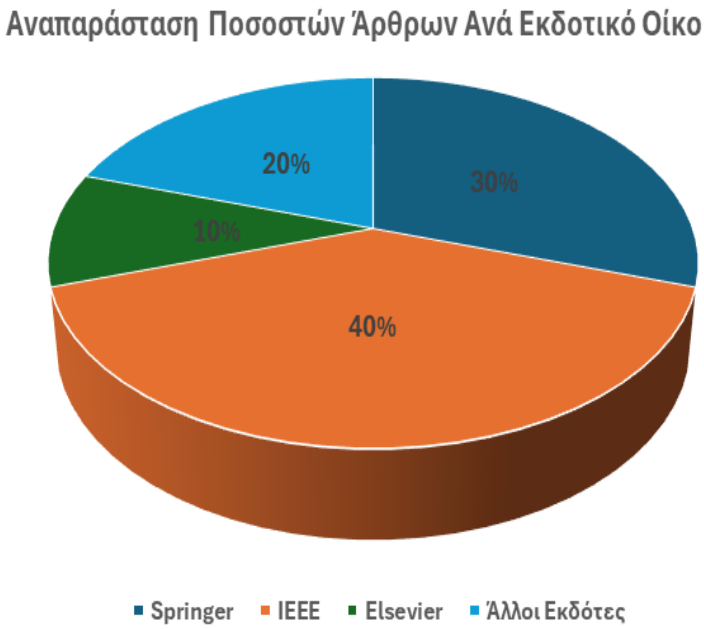
3.4 Ποσοτική Ανάλυση

Κατά τη διαδικασία βιβλιογραφικής επισκόπησης, εντοπίστηκαν αρχικά περίπου 250 άρθρα. Από αυτά, τελικά επιλέχθηκαν 100 άρθρα για τη διατριβή, εφαρμόζοντας τα κριτήρια

ένταξης και ποιότητας που καθορίστηκαν. Το χρονικό εύρος των άρθρων που επιλέχθηκαν εκτείνεται από το 2015 έως το 2024, διασφαλίζοντας ότι οι πληροφορίες είναι επικαιροποιημένες και σχετικές με τις πιο πρόσφατες εξελίξεις στην Τεχνητή Νοημοσύνη και την αντιμετώπιση επιθέσεων κοινωνικής μηχανικής. Από τα επιλεγμένα άρθρα, το 70% προέρχεται από δημοσιεύσεις σε επιστημονικά περιοδικά, ενώ το υπόλοιπο 30% προέρχεται από δημοσιεύσεις σε διεθνή συνέδρια. Αυτή η κατανομή ενισχύει την επιστημονική εγκυρότητα των πηγών και διασφαλίζει την ολοκληρωμένη κάλυψη του αντικειμένου της έρευνας.

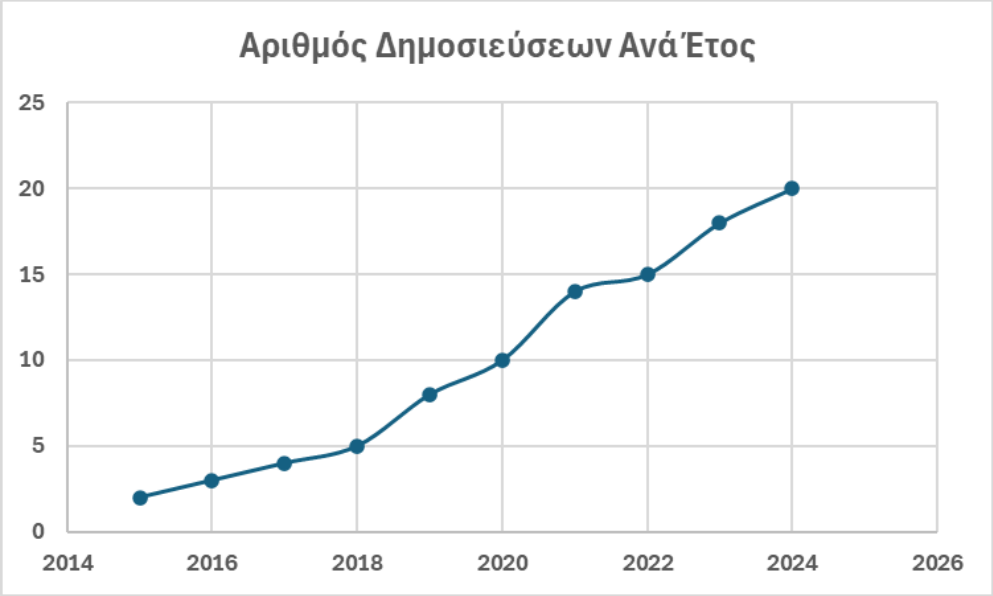
Τα ποσοστά των άρθρων ανά εκδοτικό οίκο είναι τα εξής (δείτε και Εικόνα 5):

- ***IEEE: 40%.*** Δέκα άρθρα από διεθνή συνέδρια και τριάντα από περιοδικά βρέθηκαν χρησιμοποιώντας τις παραμέτρους αναζήτησης τίτλου και περίληψης στη βάση δεδομένων των εκδοτών του IEEE.
- ***Springer: 30%***. Βρέθηκαν τριάντα αντιστοιχίες στην Springer, στην κατηγορία Journals (επιστημονικών περιοδικών), χρησιμοποιώντας τη συμβολοσειρά αναζήτησης SpringerOpen (η οποία επιστρέφει μόνο άρθρα που είναι διαθέσιμα για ανοικτή πρόσβαση).
- ***Elsevier: 10%.*** Στον εκδοτικό οίκο Elsevier τα αποτελέσματα της αναζήτησης αφορούσαν 10 επιστημονικά περιοδικά (journals).
- ***Άλλοι Εκδότες: 20%.*** Βρέθηκαν 20 αποτελέσματα κατά την αναζήτηση από το 2015 έως το 2024 σε άλλους εκδότες.



Εικόνα 5: Γραφική Αναπαράσταση των άρθρων βάση του εκδότη τους.

Παρακάτω παραθέτουμε ένα γράφημα που παρουσιάζει το πλήθος δημοσιεύσεων ανά έτος:



Εικόνα 6: Γραφική Αναπαράσταση πλήθους άρθρων ανά έτος.

Για την απεικόνιση της Εικόνας 6, επιλέχθηκε η χρήση ενός γραμμικού διαγράμματος, καθώς αυτό καθιστά πιο ευδιάκριτη την οπτικοποίηση του τρόπου με τον οποίο οι δημοσιεύσεις έχουν αλλάξει με την πάροδο του χρόνου, δίνοντας μια σαφή και συνοπτική εικόνα της ετήσιας τάσης των ερευνητικών άρθρων. Επειδή εμφανίζει την εξέλιξη των δεδομένων με την πάροδο του χρόνου και διευκολύνει τον εντοπισμό μοτίβων ή αξιοσημείωτων αλλαγών, το γραμμικό διάγραμμα είναι ιδιαίτερα κατάλληλο για την απεικόνιση χρονοσειρών. Στην προκειμένη περίπτωση, το γράφημα στον άξονα x περιλαμβάνει τις χρονολογίες των τελευταίων 10 χρόνων, ενώ ο άξονας y το πλήθος των δημοσιεύσεων. Έτσι, μας επιτρέπει να δούμε την αύξηση του πλήθους των άρθρων στους τομείς της τεχνητής νοημοσύνης (ΤΝ) και της αποτροπής/αντιμετώπισης επιθέσεων κοινωνικής μηχανικής από το 2015 έως το 2024.

Τα ευρήματα του γραμμικού διαγράμματος επιδεικνύουν μια σταθερή τάση αύξησης της ποσότητας των δημοσιεύσεων κατά τα προηγούμενα δέκα έτη. Η σταθερή αύξηση του ενδιαφέροντος και της ερευνητικής δραστηριότητας στο θέμα αυτό αναδεικνύεται από την προοδευτική αύξηση των δημοσιευμένων άρθρων από μόλις 2 το 2015, σε 20 το 2024. Η σημαντική αύξηση των δημοσιεύσεων μετά το 2018 (από 5, σε 20 το 2024) μπορεί να αποδοθεί σε διάφορους παράγοντες, όπως η αυξημένη ανησυχία γύρω από την ασφάλεια στον κυβερνοχώρο, ιδίως κατά τη διάρκεια της πανδημίας COVID-19, η οποία οδήγησε σε αύξηση των επιθέσεων κοινωνικής μηχανικής που προέρχονται από την απομακρυσμένη απασχόληση. Ο εντοπισμός αυτών των μοτίβων παρέχει σημαντικές πληροφορίες σχετικά με

τον τρόπο με τον οποίο αλλάζει ο κλάδος και υπογραμμίζει την ανάγκη για περισσότερη μελέτη και εστίαση στους κινδύνους κοινωνικής μηχανικής σήμερα περισσότερο από ποτέ.

3.5 Απαντήσεις στα Ερευνητικά Ερωτήματα

Τα ερευνητικά ερωτήματα που τέθηκαν στην αρχή της έρευνας, θα εξεταστούν στα επόμενα κεφάλαια ως εξής:

Στην Ενότητα 4.1, θα δοθούν απαντήσεις στο δεύτερο σκέλος του Ερευνητικού Ερωτήματος 1, για την κατηγοριοποίηση των αντίμετρων, που βασίζονται στην ΤΝ, κατά των επιθέσεων κοινωνικής μηχανικής.

Το Ερευνητικό Ερώτημα 2 θα εξεταστεί σε όλη την Ενότητα 4.2, καθώς θα γίνει ανάλυση των τεχνικών και μεθόδων ΤΝ που χρησιμοποιούνται κατά κόρον σε αυτά τα αντίμετρα.

Το πρώτο σκέλος από το Ερευνητικό Ερώτημα 1 θα καλυφθεί στην Ενότητα 4.3, αναλύοντας τις τεχνικές και τα αντίμετρα που χρησιμοποιούνται για την πρόληψη και αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής. Έτσι θα δοθεί μια ολοκληρωμένη εικόνα των διαθέσιμων λύσεων που χρησιμοποιούν την ΤΝ.

Στην Ενότητα 4.4, θα πραγματοποιείται η σύγκριση των αντιμέτρων, απαντώντας στο Ερευνητικό Ερώτημα 3 και καθορίζοντας ποια αντίμετρα θεωρούνται τα καλύτερα, με βάση τα προκαθορισμένα κριτήρια σύγκρισης, για όλες τις κατηγορίες.

Τα Ερευνητικά Ερωτήματα 4 και 5 θα εξεταστούν στο Κεφάλαιο 5, όπου θα συζητηθούν οι προκλήσεις και περιορισμοί της εφαρμογής της ΤΝ για την ανίχνευση και αποτροπή επιθέσεων κοινωνικής μηχανικής καθώς και θα παρουσιαστούν οι μελλοντικές εξελίξεις και οι πολλά υποσχόμενες ερευνητικές κατευθύνσεις στην ΤΝ που αναμένεται να βελτιώσουν την ικανότητα των συστημάτων ασφαλείας να μετριάζουν/αντιμετωπίζουν επιθέσεις κοινωνικής μηχανικής.

Κεφάλαιο 4: Ανάλυση

Το τέταρτο κεφάλαιο της παρούσας διατριβής παρέχει μια διεξοδική εξέταση των σύγχρονων αντιμέτρων που βασίζονται στην Τεχνητή Νοημοσύνη (ΤΝ) για την καταπολέμηση των επιθέσεων κοινωνικής μηχανικής. Στόχος του κεφαλαίου είναι να δώσει μια ενδελεχή κατηγοριοποίηση και μελέτη των αντιμέτρων και των αλγορίθμων ΤΝ που χρησιμοποιούνται για την αντιμετώπιση των διαφόρων ειδών επιθέσεων κοινωνικής μηχανικής, καθώς και μια εξέταση των προβλημάτων που αντιμετωπίζουν τα αντίμετρα αυτά. Η μέθοδος που ακολουθείται σε αυτό το κεφάλαιο επιδιώκει να εμβαθύνει διεξοδικά στις θεμελιώδεις έννοιες του τρόπου λειτουργίας των αλγορίθμων ΤΝ, δίνοντας έμφαση στα πλεονεκτήματα και τους περιορισμούς τους. Παράλληλα, επιδιώκει να αναδείξει τους πιο αποτελεσματικούς αλγορίθμους ΤΝ, συγκρίνοντας και αξιολογώντας τους με βάση συγκεκριμένα κριτήρια. Τέλος, παρουσιάζονται τα αντίμετρα που ενσωματώνουν όλους αυτούς τους αλγορίθμους, τα οποία, αφού αναλυθούν βάσει του τύπου προσέγγισης που ακολουθούν, συγκρίνονται, προσφέροντας μια πιο ολοκληρωμένη εικόνα της απόδοσης και της καταλληλότητάς τους στην αποτροπή και αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής.

Η δομή του κεφαλαίου επιτρέπει μια οργανωμένη και συνεκτική παρουσίαση των αποτελεσμάτων, ενισχύοντας τη συνολική συνεισφορά της έρευνας στην κατανόηση και αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής. Ειδικότερα, το κεφάλαιο αυτό αποτελείται από τις εξής 4 ενότητες:

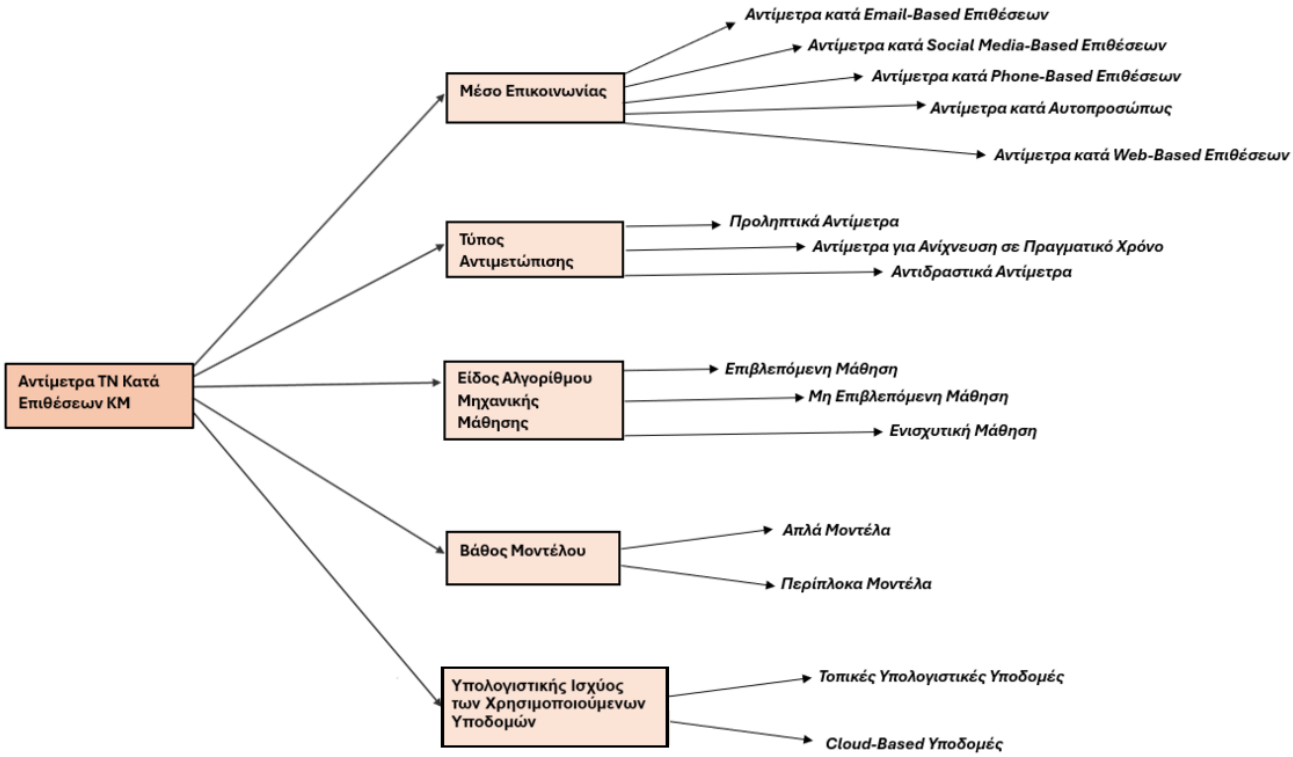
- 4.1 όπου παρουσιάζεται μια ιεραρχική κατηγοριοποίηση των αντιμέτρων και παρέχεται μια σύντομη περιγραφή των σχετικών κατηγοριών.
- 4.2 όπου γίνεται η ανάλυση των τεχνικών ΤΝ που χρησιμοποιούνται στα αντίμετρα καθώς και η σύγκρισή τους με βάση ένα συγκεκριμένο σύνολο από κριτήρια.
- 4.3 όπου γίνεται ανάλυση των αντιμέτρων με βάση τον τύπο της προσέγγισης που ακολουθούν
- 4.4 όπου γίνεται σύγκριση των αντιμέτρων και παρουσίαση των αποτελεσμάτων που προκύπτουν από αυτή τη σύγκριση.

4.1 Κατηγοριοποίηση Αντιμέτρων

Η Εικόνα 7 παρουσιάζει την ιεραρχική κατηγοριοποίηση των αντίμετρων που βασίζονται στην ΤΝ, κατά των επιθέσεων κοινωνικής μηχανικής. Η ιεραρχική κατηγοριοποίηση των αντιμέτρων επιτυγχάνεται βάσει των εξής πέντε διαστάσεων:

1. Του **μέσου επικοινωνίας** των επιθέσεων που τα αντίμετρα στοχεύουν, όπως οι επιθέσεις μέσω email, κοινωνικών δικτύων ή τηλεφωνικών κλήσεων. Αυτή η προσέγγιση επιτρέπει την ανάπτυξη εξειδικευμένων τεχνικών ΤΝ που ανταποκρίνονται στις μοναδικές προκλήσεις κάθε μέσου, διευκολύνοντας έτσι την ανάλυση και τελικά σύγκρισή τους.
2. Του **τύπου αντιμετώπισης**. Η διάσταση αυτή επιτρέπει την οργάνωση των αντιμέτρων με βάση τον τύπο αντιμετώπισης των επιθέσεων κοινωνικής μηχανικής επίθεσης που εφαρμόζουν. Οι τύποι αντιμετώπισης είναι οι εξής: (α) *πρόληψη*: λαμβάνονται προληπτικά μέτρα για την αποτροπή των επιθέσεων, (β) *ανίχνευση*: παρέχεται ικανότητα ανίχνευσης των επιθέσεων και (γ) *αντίδραση*: εφόσον ανιχνευθούν οι επιθέσεις, το αντίστοιχο αντίμετρο μπορεί είτε να τις αντιμετωπίσει επιτυχώς πριν επιτύχουν το σκοπό τους είτε να μετριάσει τις επιπτώσεις τους. Η χρήση της διάστασης αυτής διευκολύνει την ανάλυση των καλύτερων αντιμέτρων για τον κάθε τύπο αντιμετώπισης αλλά και με βάση τη δυνατότητά τους να υποστηρίξουν πολλαπλούς τύπους αντιμετώπισης, αποτελώντας με αυτό τον τρόπο πιο ολοκληρωμένες προσεγγίσεις αντιμετώπισης επιθέσεων. Ως αποτέλεσμα, προσφέρει ένα σαφές πλαίσιο και καθιστά δυνατή την αποτελεσματικότερη στόχευση των αντιμέτρων με βάση τις απαιτήσεις του κάθε τύπου αντιμετώπισης.
3. Του **είδους αλγορίθμου μηχανικής μάθησης.** Η συγκεκριμένη κατηγοριοποίηση είναι κρίσιμη για την κατανόηση της προσέγγισης ΤΝ που χρησιμοποιείται για την ανίχνευση και αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής από το κάθε αντίμετρο, καθώς κάθε τύπος αλγορίθμου ΤΝ έχει διαφορετική εφαρμογή, ισχύ και περιορισμούς.
4. Του **Βάθους Μοντέλου (Complexity of Model):** οι επιδόσεις και η αποτελεσματικότητα των μοντέλων μηχανικής μάθησης επηρεάζονται από τα διαφορετικά επίπεδα πολυπλοκότητάς τους. Τα αντίμετρα που βασίζονται σε απλά μοντέλα μπορεί να είναι πιο φιλικά προς τον χρήστη, αλλά τα περίπλοκα/σύνθετα μοντέλα μπορεί να λειτουργούν καλύτερα σε πιο περίπλοκα σενάρια.
5. **Της Υπολογιστικής Ισχύος των Χρησιμοποιούμενων Υποδομών (Computational Power of Used Infrastructure).** Από τους σημαντικότερους παράγοντες για την ταξινόμηση των αντιμέτρων είναι η διαθεσιμότητά τους που σχετίζεται με τις απαιτήσεις τους γα υπολογιστικούς πόρους αλλά και τους αλγόριθμους ΤΝ που αυτά χρησιμοποιούν. Συνεπώς, αν ένα αντίμετρο χρησιμοποιεί έναν απλό αλγόριθμο ΤΝ, τότε δεν θα έχει μεγάλες απαιτήσεις για πόρους και μπορεί να υποστηριχθεί από μια επιτόπια υποδομή. Προφανώς, όσο περισσότεροι πόροι βρίσκονται σε αυτή την υποδομή, τόσο πιο υψηλή θα είναι η διαθεσιμότητα του αντιμέτρου. Από την άλλη μεριά, αν ένα αντίμετρο χρησιμοποιεί πιο σύνθετους αλγορίθμους ΤΝ, τότε θα πρέπει να υποστηριχθεί από μια υποδομή υψηλής υπολογιστικής ισχύος, όπως είναι μια υποδομή υπολογιστικού νέφους. Μια τέτοια υποδομή μπορεί να προσφέρει έναν

μεγάλο αριθμό από πόρους και να τους κλιμακώνει αυξητικά ανάλογα με τον φόρτο εργασίας, ενισχύοντας την διαθεσιμότητα του αντιμέτρου.



Εικόνα 7: Γραφική Αναπαράσταση Κατηγοριοποίησης Αντίμετρων Βασισμένων στην ΤΝ κατά των Επιθέσεων ΚΜ.

Στη συνέχεια αναλύουμε σύντομα τις κατηγορίες της κάθε διάστασης σε ξεχωριστές παραγράφους.

Η διάσταση του μέσου επικοινωνίας (που χρησιμοποιείται από τις επιθέσεις που αντιμετωπίζονται από τα αντίμετρα) περιλαμβάνει τις εξής κατηγορίες:.

- **Αντίμετρα κατά Email-Based Επιθέσεων:** Περιλαμβάνει αντίμετρα που στοχεύουν στον εντοπισμό και την αντιμετώπιση επιθέσεων, όπως phishing, spear phishing, και clone phising, δηλαδή επιθέσεων μέσω ηλεκτρονικού ταχυδρομείου (email).

- **Αντίμετρα κατά Social Media-Based Επιθέσεων:** Εστιάζει σε αντίμετρα που χρησιμοποιούνται για τον εντοπισμό επιθέσεων ή κατηγοριών επιθέσεων που βασίζονται στα κοινωνικά δίκτυα, όπως catfishing, πλαστοπροσωπία (impersonation) και συλλογή διαπιστευτηρίων (credential harvesting).

- **Αντίμετρα κατά Phone-Based Επιθέσεων:** Περιλαμβάνει αντίμετρα για την αντιμετώπιση τηλεφωνικών επιθέσεων, όπως vishing, smishing και robo-calls.

- **Αντίμετρα κατά Αυτοπροσώπως (In-Person) Επιθέσεων:** Περιλαμβάνει αντίμετρα που χρησιμοποιούνται για την αντιμετώπιση αυτοπροσώπως επιθέσεων, όπως το tailgating και pretexting, που βασίζονται είτε σε μια φυσική επαφή είτε σε μια άμεση ψηφιακή επαφή.

- **Αντίμετρα κατά Web-Based Επιθέσεων:** Εστιάζει σε αντίμετρα που χρησιμοποιούνται για την ανίχνευση διαδικτυακών επιθέσεων (web-based attacks), όπως watering hole, malvertising και clickjacking, που έχουν ως μέσο επίθεσης το διαδίκτυο (πχ. μολυσμένες ή κακόβουλες ιστοσελίδες).

Η διάσταση του τύπου αντιμετώπισης περιλαμβάνει τις εξής κατηγορίες:

- **Προληπτικά Αντίμετρα**:
  - *Στόχος*: Ανίχνευση δυνητικών επιθέσεων και πρόληψή τους πριν εκτελεστούν.
  - *Παραδείγματα*: Τεχνικές ανάλυσης συμπεριφοράς, φίλτρα ανίχνευσης phishing (επιθέσεων).
- **Αντίμετρα για Ανίχνευση σε Πραγματικό Χρόνο**:
  - *Στόχος*: Συνεχής παρακολούθηση για ανίχνευση επιθέσεων ή εισβολών καθώς συμβαίνουν.
  - *Παραδείγματα*: Ανίχνευση ανωμαλιών δικτύου, συστήματα ανίχνευσης εισβολών (Intrusion Detection Systems - IDS).
- **Αντιδραστικά Αντίμετρα**:
  - *Στόχος*: Αντίδραση σε επιθέσεις μόλις εντοπιστούν ή καταγραφούν για την αντιμετώπισή τους ή μετριασμό των επιπτώσεών τους.
  - *Παραδείγματα*: Αυτοματοποιημένη αντίδραση σε επιθέσεις, συστήματα απομόνωσης παραβιασμένων περιοχών.

Η διάσταση του είδους αλγορίθμου μηχανικής μάθησης (3) περιλαμβάνει τις εξής κατηγορίες:

- **Επιβλεπόμενη Μάθηση (Supervised Learning)**: Αφορά την χρήση αλγορίθμων ΤΝ που έχουν εκπαιδευτεί σε δομημένα δεδομένα με ετικέτες και χρησιμοποιούνται συχνά για πρόβλεψη και ταξινόμηση. Οι Support Vector Machines (SVM), Random Forests, και Logistic Regression είναι μερικά παραδείγματα τέτοιων αλγορίθμων. Αυτοί οι αλγόριθμοι μπορεί να μην είναι σε θέση να εντοπίσουν νέες ή μη αναγνωρισμένες επιθέσεις, αλλά είναι αποτελεσματικοί στον εντοπισμό γνωστών επιθέσεων.

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)**: Τα δεδομένα χωρίς ετικέτες (labels) αναλύονται με τη χρήση αυτής της οικογένειας αλγορίθμων για την εύρεση ανωμαλιών και κρυφών μοτίβων. Η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA) και η Ανάλυση Συστάδων (Cluster Analysis) είναι δύο μέθοδοι που χρησιμοποιούνται για την εύρεση περίεργων μοτίβων που μπορούν να υποδείξουν κακόβουλη δραστηριότητα. Είναι ιδιαίτερα χρήσιμες για τον εντοπισμό νέων επιθέσεων κοινωνικής μηχανικής.

- **Ενισχυτική Μάθηση (Reinforcement Learning)**: Αυτό το είδος (αλγορίθμων) ΤΝ χρησιμοποιείται για τη βελτιστοποίηση αποφάσεων σε διαρκώς μεταβαλλόμενες συνθήκες. Οι αλγόριθμοι ενισχυτικής μάθησης, όπως οι Q-learning και Deep Q-Networks (DQN), μαθαίνουν μέσω αλληλεπίδρασης με το περιβάλλον τους, προσαρμοζόμενοι σε νέες απειλές καθώς αυτές εμφανίζονται, καθιστώντας τους κατάλληλους για εφαρμογές, όπως η ανίχνευση εισβολών.

Η διάσταση του βάθους μοντέλου (Complexity of Model) που χρησιμοποιείται από ένα αντίμετρο (4) περιλαμβάνει τις εξής κατηγορίες:

- **Απλά Μοντέλα (Simple Models)**: Η χρήση αλγορίθμων ΤΝ, όπως Naive Bayes, Logistic Regression και Decision Trees, οδηγεί στην παραγωγή σχετικά απλών μοντέλων με χαμηλό υπολογιστικό κόστος. Αυτά τα μοντέλα είναι εύκολα να εκπαιδευτούν και να υλοποιηθούν, καθιστώντας τα καταλληλα για περιπτώσεις που απαιτούν γρήγορη απόκριση σε επιθέσεις με περιορισμένους πόρους. Ωστόσο, η απλότητά τους μπορεί να περιορίσει την ακρίβεια και την αντοχή σε πιο σύνθετες επιθέσεις.

- **Περίπλοκα Μοντέλα (Complex Models)**: Τα μοντέλα που παράγονται από αλγορίθμους βαθιάς μάθησης (Deep Learning) ή από σύνθεση πολλαπλών μοντέλων είναι πολύ πιο περίπλοκα/σύνθετα, απαιτώντας περισσότερο χρόνο και κόστος για την κατασκευή και εφαρμογή τους, λόγω και της ανάγκης χρήσης περισσότερων υπολογιστικών πόρων. Τα μοντέλα αυτά λειτουργούν ιδιαίτερα καλά στον εντοπισμό πιο σύνθετων επιθέσεων κοινωνικής μηχανικής, όπως οι απόπειρες deepfake ή πλαστοπροσωπίας. Τα μοντέλα αυτά που είναι «μαύρου κουτιού» (black-box models) είναι πιο ακριβή, αλλά συνήθως είναι πιο δύσκολο να κατανοηθεί ο τρόπος λειτουργίας τους.

Η διάσταση της χρήση υποδομών υψηλής υπολογιστικής ισχύος (Computational Resources) στα αντίμετρα (5) περιλαμβάνει τις εξής κατηγορίες:

- **Τοπικές Υπολογιστικές Υποδομές (On-Premise Systems)**: Τα συστήματα αυτά ανιχνεύουν και αντιδρούν στις απειλές χρησιμοποιώντας την ικανότητα επεξεργασίας της τοπικής υποδομής, όπως ένας ή παραπάνω διακομιστές της επιχείρησης. Παρόλο που οι γρήγοροι χρόνοι αντίδρασης και ο έλεγχος των δεδομένων είναι δυνατοί με μια τέτοια τεχνολογία (υπό κανονικούς φόρτους εργασίας), η αξιοποίησή της περιορίζεται από τους πόρους που διαθέτει η εν λόγω επιχείρηση. Λόγω του χαμηλού υπολογιστικού τους κόστους, απλοί αλγόριθμοι ΤΝ, όπως τα δέντρα αποφάσεων (Decision Trees) και οι Naive Bayes, είναι κατάλληλοι για τέτοιου είδους υλοποιήσεις.

- **Cloud-Based Υποδομές:** Μεγάλες ποσότητες δεδομένων μπορούν να υποβληθούν σε επεξεργασία με τη χρήση λύσεων που βασίζονται στο υπολογιστικό νέφος (cloud), λόγω της ισχυρής υπολογιστικής ισχύος που προσφέρουν οι πλατφόρμες νέφους. Για την αποτελεσματική λειτουργία πολύπλοκων αλγορίθμων σε πραγματικό χρόνο, όπως τα μοντέλα βαθιάς μάθησης (Deep Learning Models) και τα Recurrent Neural Networks (RNNs), απαιτούνται συχνά αυτές οι υποδομές. Παρόλο που οι λύσεις αυτές είναι αποδοτικές και κλιμακούμενες, έχουν υψηλότερο λειτουργικό κόστος και ενδέχεται να υπάρχουν ζητήματα ασφαλείας με αυτές (ειδικότερα λόγω της χρήσης δημόσιων νεφών με υψηλή επιφάνεια επιθέσεων).

4.2 Ανάλυση Τεχνικών ΤΝ που Χρησιμοποιούνται στα Αντίμετρα

Η ώθηση για την εισαγωγή της ΤΝ στην ασφάλεια του κυβερνοχώρου ξεκίνησε πραγματικά επειδή οι παραδοσιακές μέθοδοι προστασίας από επιθέσεις κοινωνικής μηχανικής δεν μπορούσαν να ανταποκριθούν επαρκώς ενάντια στις εξελιγμένες και πολυδιάστατες επιθέσεις κοινωνικής μηχανικής. Με το phishing και άλλες μεθόδους επίθεσης σε άνοδο, και την απόλυτη πολυπλοκότητα των δεδομένων που αντιμετωπίζουμε, έπρεπε να στραφούμε σε αλγόριθμους ΤΝ. Η ΤΝ και η Μηχανική Μάθηση που ενσωματώνει μπορούν να οδηγήσουν στην παραγωγή μοντέλων, τα οποία είναι ικανά να χειριστούν τόνους δεδομένων, να εντοπίζουν μοτίβα και ανωμαλίες σε πραγματικό χρόνο με μεγάλη ακρίβεια. Για αυτό το λόγο, ενσωματώνονται ολοένα και περισσότερο σε συστήματα ασφάλειας, εκσυγχρονίζοντάς τα και κάνοντάς τα πιο επιτυχή. Επιπλέον, με την βοήθεια της συνεχούς μάθησης, η σύνδεση αυτή επιτρέπει στα συστήματα να προσαρμόζονται σε νέες απειλές, παρέχοντας μια αυτοματοποιημένη, δυναμική μέθοδο ασφάλειας.

Στην ενότητα αυτή, έπειτα από την παρουσίαση των τρεχόντων ζητημάτων και των αντίστοιχων τεχνικών / αλγορίθμων ΤΝ προς επίλυσή τους, θα παρουσιαστεί ένας συγκριτικός πίνακας των αλγορίθμων αυτών, με βάση συγκεκριμένα κριτήρια. Πιο συγκεκριμένα, η δομή της ενότητας 4.2 βασίζεται σε τρεις κύριες υπο-ενότητες. Αρχικά, η υπο-ενότητα 4.2.1 παρουσιάζει τα προβλήματα που πρέπει να αντιμετωπιστούν, κατηγοριοποιημένα ανάλογα με

το μέσο επικοινωνίας που χρησιμοποιούν οι επιθέσεις κοινωνικής μηχανικής. Στη συνέχεια, η υπο-ενότητα 4.2.2 αναλύει τους αλγόριθμους Τεχνητής Νοημοσύνης που μπορούν να χρησιμοποιηθούν για την αντιμετώπιση των συγκεκριμένων προβλημάτων, οργανώνοντάς τους επίσης με βάση το μέσο επικοινωνίας των επιθέσεων. Τέλος, η υπο-ενότητα 4.2.3 περιλαμβάνει έναν συγκριτικό πίνακα των αλγορίθμων αυτών, χρησιμοποιώντας ένα συγκεκριμένο σύνολο κριτηρίων. Από την ανάλυση αυτή προκύπτουν παρατηρήσεις και συμπεράσματα που ισχύουν τόσο στο επίπεδο του εκάστοτε μέσου επικοινωνίας των επιθέσεων όσο και σε καθολικό επίπεδο.

4.2.1 Προβλήματα προς αντιμετώπιση

Τα κύρια προβλήματα στην αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής, περιστρέφονται γύρω από τον εντοπισμό επιθέσεων που εκμεταλλεύονται την ανθρώπινη ψυχολογία, τη πλαστοπροσωπία και τη χρήση ψευδών σεναρίων για την απόκτηση προσωπικών πληροφοριών. Λόγω της πολύπλευρης φύσης αυτών των επιθέσεων, κάθε μία από αυτές, ανάλογα με το μέσο επικοινωνίας, μπορεί να εκμεταλλεύεται και να βασίζεται σε διαφορετικές περιστάσεις. Συνεπώς, ακολουθεί μια ανάλυση, που αφορά τα κύρια ζητήματα που εντοπίζονται στην αντιμετώπιση επιθέσεων Κοινωνικής Μηχανικής (ΚΜ), βασισμένη στο μέσο επικοινωνίας των εκάστοτε επιθέσεων.

4.2.1.1 Email-Based Επιθέσεις

Τα ζητήματα που προκύπτουν από τις επιθέσεις **phishing** είναι πολύπλευρα. Πρώτον, το γεγονός ότι οι επιτιθέμενοι χρησιμοποιούν συνδέσμους που φαίνονται αυθεντικοί αλλά στην πραγματικότητα οδηγούν τους χρήστες σε κακόβουλους ιστότοπους, καθιστά τον εντοπισμό των δόλιων διευθύνσεων URL ένα από τα πιο δύσκολα καθήκοντα. Ο βαθμός δυσκολίας αυξάνεται ακόμη περισσότερο όταν οι κακόβουλοι ιστότοποι μοιάζουν πάρα πολύ με τους αντίστοιχους αυθεντικούς/νόμιμους. Επίσης, οι επιτιθέμενοι χρησιμοποιούν μια διεύθυνση αποστολέα που μοιάζει με την επίσημη (δηλαδή με αυτή ενός νόμιμου αποστολέα), οπότε είναι ζωτικής σημασίας να αποκαλύπτονται μόνο οι ψεύτικοι αποστολείς, χωρίς όμως να διαγράφονται ή να μπαίνουν σε καραντίνα μηνύματα που προέρχονται από νόμιμους, αυθεντικούς αποστολείς.

Στις **spear phishing** επιθέσεις, τα προβλήματα γίνονται ακόμα πιο εξειδικευμένα, καθώς οι επιτιθέμενοι προσαρμόζουν τα μηνύματά τους για να μοιάζουν απολύτως φυσιολογικά στον παραλήπτη, βασισμένα σε προσωπικές πληροφορίες του στόχου. Δεδομένου ότι το περιεχόμενο συχνά δεν έχει εμφανώς παραπλανητικά χαρακτηριστικά (σε αυτό θα μπορούσε να βοηθήσει και η ΤΝ), ο εντοπισμός αυτών των στοχευμένων επιθέσεων

- που χρησιμοποιούν προσωπικές πληροφορίες για να αυξήσουν την πειθώ - αποτελεί σημαντική πρόκληση.

Στις επιθέσεις **clone phishing**, η πρόκληση έγκειται περισσότερο στην ανίχνευση πολύ μικρών τροποποιήσεων σε ήδη γνωστά και αποδεκτά μηνύματα (που μπορεί να έχουν σταλθεί προσφάτως). Οι επιτιθέμενοι εδώ εισάγουν κακόβουλους συνδέσμους ή παραποιημένο περιεχόμενο σε ένα μήνυμα που κατά τα άλλα είναι πανομοιότυπο με ένα προηγουμένως νόμιμο email.

4.2.1.2 Social Media-Based Επιθέσεις

Το πρωταρχικό ζήτημα με τις επιθέσεις **catfishing** είναι η κατασκευή ψευδών προφίλ που δεν γνωρίζει ο στόχος-θύμα και η επακόλουθη εξαπάτηση του θύματος με τη χρήση πλασματικών ή κατασκευασμένων ταυτοτήτων. Επειδή οι επιτιθέμενοι χρησιμοποιούν συχνά ονόματα, εικόνες και άλλες λεπτομέρειες που φαίνονται αυθεντικές, μπορεί να είναι δύσκολο να εντοπιστεί αν ένα προφίλ είναι όντως ψευδές. Η διατήρηση της επαφής με τα θύματα είναι ένα δεύτερο ζήτημα, καθώς οι επιτιθέμενοι χρησιμοποιούν ψεύτικες αφηγήσεις ή συναισθηματικές συνδέσεις για να ξεγελάσουν τον στόχο τους και να κερδίσουν την εμπιστοσύνη του.

Στις επιθέσεις **Impersonation** (πλαστοπροσωπίας), το βασικό πρόβλημα είναι η ακριβής αντιγραφή της ταυτότητας ενός γνήσιου ατόμου ή οργανισμού στα κοινωνικά δίκτυα. Για να εξαπατήσουν τους χρήστες, οι επιτιθέμενοι κατασκευάζουν προφίλ ή λογαριασμούς που μιμούνται τις πραγματικές ταυτότητες. Η δυσκολία εντοπισμού λεπτών διαφοροποιήσεων μεταξύ αυθεντικών και απατηλών προφίλ, ιδίως όταν οι επιτιθέμενοι έχουν πρόσβαση στις προσωπικές πληροφορίες των πλαστογραφημένων οντοτήτων, και η αδυναμία επιβεβαίωσης της αυθεντικότητας των προφίλ είναι τα κύρια ζητήματα. Μια άλλη δυσκολία είναι να προσδιοριστεί ποιες χρήσεις μιας ταυτότητας είναι νόμιμες και ποιες όχι.

Ο θεμελιώδης στόχος των επιθέσεων **Credential Harvesting** (αυτές θεωρούνται ως μια κατηγορία επιθέσεων που περιλαμβάνει επιθέσεις, όπως phishing, vishing, baiting, κλπ.) είναι η κλοπή των διαπιστευτηρίων των χρηστών, η οποία συχνά επιτυγχάνεται μέσω ψευδών σελίδων σύνδεσης ή μηνυμάτων που παραπέμπουν σε ιστότοπους phishing. Ένα από τα πιο δύσκολα καθήκοντα είναι ο εντοπισμός των ψεύτικων ιστοσελίδων σύνδεσης και η διασφάλιση των διαπιστευτηρίων των χρηστών από την αποκάλυψη. Ο εντοπισμός ψευδών μηνυμάτων ή προσπαθειών σύνδεσης που οδηγούν σε επικίνδυνους ιστότοπους είναι μια σημαντική πρόκληση που περιπλέκει την προστασία των διαπιστευτηρίων.

4.2.1.3 Phone-Based Επιθέσεις

Στις επιθέσεις **Vishing**, το πρώτο ζήτημα που πρέπει να αντιμετωπιστεί είναι ο εντοπισμός των παραπλανητικών κλήσεων, καθώς οι επιτιθέμενοι χρησιμοποιούν συχνά τεχνολογίες παραποίησης αριθμών για να κάνουν τις κλήσεις να φαίνονται ότι προέρχονται από αξιόπιστες οργανώσεις ή επιχειρήσεις. Δεύτερον, υπάρχει ζήτημα με την ανάλυση φωνής και τόνου, καθώς οι επιτιθέμενοι χρησιμοποιούν στρατηγικές ψυχολογικής πίεσης (πχ. δημιουργία αίσθησης του επείγοντος) για να πείσουν τα θύματα. Οπότε, μπορεί να είναι δύσκολο να εντοπιστούν αυτά τα ψευδή φωνητικά μοτίβα. Τέλος, η ανίχνευση παρατυπιών στις συνομιλίες (πχ. ασυνήθιστα αιτήματα ή ασάφεια στις πληροφορίες) είναι ένα κρίσιμο ζήτημα, καθώς οι επιτιθέμενοι προσπαθούν να αποκτήσουν ζωτικές πληροφορίες κατά τη διάρκεια μιας φαινομενικά έγκυρης συνομιλίας, γεγονός που καθιστά αναγκαία τη συνεχή εξέταση των αιτημάτων κλήσης.

Οι επιθέσεις **Smishing** περιλαμβάνουν επιτιθέμενους που χρησιμοποιούν μηνύματα SMS για να εξαπατήσουν τους καταναλωτές ώστε να αποκαλύψουν προσωπικές πληροφορίες ή να κάνουν κλικ σε επικίνδυνους ιστότοπους. Το κύριο ζήτημα εδώ είναι ο εντοπισμός των κακόβουλων SMS, τα οποία μοιάζουν να προέρχονται από αξιόπιστες πηγές και περιέχουν συνδέσμους ή αιτήματα για ευαίσθητες πληροφορίες. Ένα άλλο ζήτημα είναι η ανίχνευση επικίνδυνων συνδέσμων σε SMS που ανακατευθύνουν σε ιστότοπους phishing. Οι επιτιθέμενοι χρησιμοποιούν συνδέσμους που φαίνονται αυθεντικοί αλλά στην πραγματικότητα είναι επιβλαβείς. Τέλος, υπάρχει το ζήτημα του εντοπισμού παραπλανητικού περιεχομένου και τόνου, στο οποίο οι επιτιθέμενοι χρησιμοποιούν γλώσσα που μεταδίδει μια αίσθηση επείγοντος, όπως προειδοποιήσεις για ζητήματα/προβλήματα σε τράπεζες ή λογαριασμούς, καθιστώντας δυσκολότερο τον εντοπισμό παραπλανητικών μηνυμάτων.

4.2.1.4 In-Person Επιθέσεις

Στις επιθέσεις **Tailgating/Piggybacking**, το πρόβλημα που πρέπει να αντιμετωπιστεί είναι η μη εξουσιοδοτημένη πρόσβαση σε ελεγχόμενες περιοχές μέσω της φυσικής παραβίασης. Οι επιτιθέμενοι ακολουθούν νόμιμους χρήστες για να αποκτήσουν πρόσβαση σε ασφαλείς περιοχές χωρίς να έχουν τα κατάλληλα διαπιστευτήρια. Το κύριο πρόβλημα είναι η δυσκολία στην ανίχνευση αυτών των φυσικών παραβιάσεων σε πραγματικό χρόνο, καθώς οι επιτιθέμενοι συχνά εισέρχονται χωρίς να γίνονται αντιληπτοί. Εφόσον η επιτήρηση των σημείων εισόδων πραγματοποιείται με κάμερες ασφαλείας, ένα σχετικό ζήτημα είναι η ποιότητα της εικόνας, η οποία μπορεί να είναι κακή ή να δυσχαιρένεται από την ώρα της ημέρας, πράγμα το οποίο κάνει πιο δύσκολη την ανίχνευση.

Οι **Pretexting** επιθέσεις περιλαμβάνουν επιτιθέμενους που δημιουργούν μια φανταστική προσωπικότητα ή μια ιστορία για να πείσουν τα θύματα να αποκαλύψουν ευαίσθητες πληροφορίες. Το ζήτημα εδώ είναι να εντοπίζονται οι ψεύτικες ιστορίες και η παραπλανητική συμπεριφορά. Οι επιτιθέμενοι χρησιμοποιούν την κοινωνική μηχανική για να κερδίσουν την εμπιστοσύνη των θυμάτων τους, καθιστώντας δύσκολη την αναγνώριση της απάτης. Επιπλέον, η επιβεβαίωση της εγκυρότητας της ταυτότητας του επιτιθέμενου μπορεί

να είναι δύσκολη, επειδή συχνά αυτός χρησιμοποιεί πραγματικά δεδομένα για να υποστηρίξει την απατηλή του ταυτότητα.

4.2.1.5 Web-Based Επιθέσεις

Οι επιθέσεις **Watering Hole** στοχεύουν σε ιστότοπους που επισκέπτονται συχνά τα θύματα-στόχοι και τους μολύνουν με κακόβουλο λογισμικό. Ένα από τα βασικά ζητήματα που αντιμετωπίζονται είναι η αναγνώριση των παραβιασμένων νόμιμων ιστότοπων, καθώς οι επιτιθέμενοι εισάγουν επιβλαβή κώδικα σε αξιόπιστες πλατφόρμες που οι χρήστες θεωρούν ότι είναι ασφαλείς. Επιπλέον, το κακόβουλο λογισμικό που χρησιμοποιείται μπορεί να είναι μη αναγνωρίσιμο, καθιστώντας δύσκολη την ανίχνευση κακόβουλου κώδικα σε κατά τα άλλα ασφαλείς συνδέσεις. Ένα άλλο σημαντικό ζήτημα είναι η άμυνα κατά των στοχευμένων επιθέσεων, στις οποίες οι επιτιθέμενοι χρησιμοποιούν μοναδικά μοτίβα και συμπεριφορές για να στοχεύουν με ακρίβεια τα θύματά τους, καθιστώντας αδύνατη την αποφυγή αυτών των απειλών σε πραγματικό χρόνο. Να τονιστεί σε αυτό το σημείο πως η αδυναμία αυτή έγκειται στο γεγονός πως στις επιθέσεις αυτές η συμπεριφορά μιας ιστοσελίδας/ιστοτόπου είναι συνήθως κανονική και μόνο όταν γίνεται σύνδεση σε αυτήν ενός στοχευμένου θύματος, υπάρχει κάποια διαφοροποίηση. Συνεπώς, είναι πολύ δύσκολη η συμπεριφορική ανίχνευση για αυτόν ακριβώς το λόγο.

Στις επιθέσεις **Malvertising**, οι επιτιθέμενοι χρησιμοποιούν διαφημιστικά δίκτυα για να διασπείρουν κακόβουλο λογισμικό μέσω διαφημίσεων, που εμφανίζονται σε νόμιμες ιστοσελίδες. Ένα από τα βασικά προβλήματα είναι η ανίχνευση κακόβουλων διαφημίσεων σε πραγματικό χρόνο, καθώς αυτές μπορούν να εμφανιστούν σε πολλές πλατφόρμες πριν προλάβουν να εντοπιστούν. Επίσης, οι επιτιθέμενοι εκμεταλλεύονται τις αλληλεπιδράσεις των χρηστών με τις διαφημίσεις, απαιτώντας από τους χρήστες να επιλέξουν (μέσω του ποντικιού) τις διαφημίσεις για να ενεργοποιηθεί το κακόβουλο λογισμικό, κάτι που καθιστά την ανίχνευση αυτών των αλληλεπιδράσεων δύσκολη. Τέλος, οι κακόβουλες διαφημίσεις συχνά εμφανίζονται ως αθώες ή στοχευμένες διαφημίσεις που προσελκύουν συγκεκριμένες ομάδες χρηστών, κάνοντας την αναγνώρισή τους πριν την αλληλεπίδραση ακόμη πιο περίπλοκη.

4.2.2 Ανάλυση Σχετικών Αλγορίθμων ή τεχνικών ΤΝ

Σε αυτή την υποενότητα θα αναλυθούν οι πιο διαδεδομένοι αλγόριθμοι και οι τεχνικές ΤΝ που εφαρμόζονται για την αντιμετώπιση των προβλημάτων που αναλύθηκαν στην προηγούμενη υποενότητα (4.2.1). Βάση των άρθρων που μελετήθηκαν, διαχωρίστηκαν οι χρησιμοποιούμενες τεχνικές/αλγόριθμοι ΤΝ ανά μέσο επικοινωνίας των επιθέσεων που αντιμετωπίζουν, ανάλογα με την βέλτιστη συμβατότητα αλγορίθμου - επίθεσης, ενώ παράλληλα διατυπώθηκαν τα πλεονεκτήματα και μειονεκτήματα κάθε αλγορίθμου ξεχωριστά. Η συγκεκριμένη τακτική έχει ως σκοπό την πιο αντικειμενική αξιολόγηση τόσο μεμονωμένα (των διαφόρων τεχνικών ΤΝ αντιμετώπισης για τις υποκατηγορίες των επιθέσεων), όσο και

συνολικά για τους αλγόριθμους που κατατάσσονται στην αντιμετώπιση επιθέσεων ΚΜ, με παράμετρο/διάσταση κατηγοριοποίησης το μέσο επικοινωνίας. Ακολουθεί η σχετική ανάλυση, για κάθε κατηγορία ξεχωριστά:

4.2.2.1 Email-Based Επιθέσεις

- **Phishing:**

  - **Support Vector Machine (SVM) [59]**: Πρόκειται για ένα μοντέλο εποπτευόμενης μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Είναι ιδιαίτερα αποτελεσματικό στην ανίχνευση επιθέσεων phishing, ειδικότερα όταν αυτές περιλαμβάνουν συνδέσμους σε κακόβουλες ιστοσελίδες, αναλύοντας τα χαρακτηριστικά των ιστοσελίδων και ταξινομώντας τις είτε ως phishing (δηλ. κακόβουλες) είτε ως έγκυρες. Ο αλγόριθμος SVM διαχειρίζεται καλά χώρους υψηλής διάστασης και βοηθά στην εξαγωγή χαρακτηριστικών για την ακριβή αναγνώριση ιστοσελίδων ως phishing, ενώ παράλληλα είναι κατάλληλος για την αναγνώριση παραπλανητικών URL και συνδέσμων στα phishing emails, δημιουργώντας σαφή όρια μεταξύ κακόβουλων και μη κακόβουλων email. Για την ταξινόμηση phishing URLs βασίζεται σε εξαγόμενα χαρακτηριστικά, όπως το μήκος του URL, την ηλικία του domain και τα (χρησιμοποιούμενα) πιστοποιητικά SSL.

    Πλεονεκτήματα: Υψηλή ακρίβεια, ικανότητα γενίκευσης σε διαφορετικά σενάρια και ικανότητα διαχείρισης μη γραμμικών συνόλων δεδομένων μέσω της χρήσης kernel functions.

    Μειονεκτήματα: Απαιτεί πολλούς υπολογιστικούς πόρους, ιδιαίτερα σε μεγάλης κλίμακας περιβάλλοντα, είναι πιο αργός όταν εφαρμόζεται σε συστήματα ανίχνευσης σε πραγματικό χρόνο ενώ η επιλογή του κατάλληλου kernel και των παραμέτρων μπορεί να είναι πολύπλοκη και να απαιτεί (εκτεταμένες) δοκιμές.

  - **Αλγόριθμοι Δέντρων Αποφάσεων (Decision Tree) [58], [59]**: Αυτή η τεχνική ΤΝ/ΜΜ δημιουργεί ένα μοντέλο που προβλέπει την στοχευμένη τιμή (πχ. αν ένα email είναι phishing) με βάση διάφορα χαρακτηριστικά εισόδου. Οι Αλγόριθμοι Δέντρων Αποφάσεων βοηθούν στην αντιμετώπιση του προβλήματος της μαζικής ανίχνευσης παραπλανητικών email (phishing), κατατάσσοντας τα μηνύματα με βάση συγκεκριμένα χαρακτηριστικά (πχ., τη διεύθυνση του αποστολέα, τον τύπο του συνδέσμου). Έτσι, δημιουργούν ένα

σύνολο κανόνων που ταξινομούν τα email σε κατηγορίες, μαθαίνοντας από εκπαιδευτικά δεδομένα που περιλαμβάνουν παραδείγματα phishing και έγκυρων email [58]. Πρόκειται για μια απλή αλλά αποτελεσματική μέθοδο για την ανίχνευση επιθέσεων κοινωνικής μηχανικής, ειδικά του phishing, αναλύοντας μοτίβα στα δεδομένα [59] και εντοπίζοντας παραπλανητικούς συνδέσμους / ψευδείς αποστολείς.

Πλεονεκτήματα: Εύκολη τεχνική TN/MM στην ερμηνεία και την εφαρμογή. Μπορεί να ταξινομήσει γρήγορα URLs σε phishing ή έγκυρες κατηγορίες με βάση προκαθορισμένες συνθήκες (που έχουν προκύψει από την εκπαίδευση).

Μειονεκτήματα: Τείνει να υπερπροσαρμόζεται (overfitting) σε μεγάλα σύνολα δεδομένων ή όταν περιλαμβάνονται πολλά χαρακτηριστικά, με αποτέλεσμα τη μειωμένη γενίκευση, αν δεν χρησιμοποιηθούν κατάλληλες τεχνικές ρύθμισης.

- **Random Forest Classifier [59]**: Μια μέθοδος συλλογικής μάθησης που λειτουργεί με την κατασκευή πολλαπλών δέντρων αποφάσεων. Εφαρμόζοντας έπειτα τον μέσο όρο των αποτελεσμάτων πολλών δέντρων, αυξάνεται η ακρίβεια ανίχνευσης. Η μέθοδος Random Forest είναι ικανή να διαχειριστεί μεγάλο όγκο χαρακτηριστικών και να παράγει ακριβείς προβλέψεις κατά των ιστοσελίδων phishing. Χρησιμοποιώντας αυτή τη μέθοδο, οι διαδικτυακές σελίδες μπορούν να κατηγοριοποιηθούν ανάλογα με το περιεχόμενό τους, τη δομή της διεύθυνσης URL και άλλα χαρακτηριστικά που θεωρούνται ενδείξεις phishing. Η μέθοδος αυτή συνδυάζει πολλά δέντρα αποφάσεων για να ενισχύσει την ακρίβεια και να μειώσει την πιθανότητα υπερπροσαρμογής, κάτι που την καθιστά εξαιρετικά αποτελεσματική στη μαζική ανάλυση email. Με λίγα λόγια, επεξεργάζεται τα χαρακτηριστικά των email σε μεγάλο βάθος, αναλύοντας κάθε στοιχείο τους και συνδυάζοντας τα αποτελέσματα πολλαπλών δέντρων, για να εξασφαλίσει ακριβή ανίχνευση phishing μηνυμάτων σε μεγάλες λίστες.

  Πλεονεκτήματα: Η μέθοδος αυτήαρέχει υψηλή ακρίβεια, είναι ανθεκτική στον θόρυβο και ιδανική για μεγάλα σύνολα δεδομένων (datasets) και είναι λιγότερο επιρρεπής στην υπερπροσαρμογή σε σύγκριση με άλλους ταξινομητές, όπως τα Δέντρα Αποφάσεων.

  Μειονεκτήματα: Αν και είναι ελαφρύτερη σε υπολογιστικούς πόρους από την τεχνική SVM, εξακολουθεί να απαιτεί σημαντικούς πόρους και μπορεί να είναι

πιο αργή σε σχέση με απλούστερους τεχνικές, όπως την KNN (που αναλύεται στη συνέχεια).

- **K-Nearest Neighbor (KNN) [58]:** Είναι ένας απλός αλγόριθμος ΜΜ που βασίζεται σε παραδείγματα και χρησιμοποιείται για την ανίχνευση phishing συγκρίνοντας νέες ιστοσελίδες με τις πιο παρόμοιες από το σύνολο εκπαίδευσης. Συγκρίνοντας την εγγύτητα μιας ιστοσελίδας με γνωστές σελίδες phishing ή νόμιμες σελίδες με βάση συγκεκριμένα χαρακτηριστικά, ο KNN μπορεί να αναγνωρίσει αποτελεσματικά ιστοσελίδες phishing.

  Πλεονεκτήματα: Εύκολος στην εφαρμογή και δεν απαιτεί συγκεκριμένα είδη/μοντέλα κατανομής των δεδομένων.

  Μειονεκτήματα: Απαιτεί υψηλή υπολογιστική ισχύ σε πραγματικό χρόνο, καθώς πρέπει να συγκρίνει όλα τα σημεία δεδομένων. Επίσης, είναι λιγότερο ακριβής από τον SVM ή τον Random Forest όταν εφαρμόζεται σε μεγάλα σύνολα δεδομένων.

- **Logistic Regression [60]**: Είναι ένας βασικός αλγόριθμος ΜΜ, που χρησιμοποιείται κυρίως για δυαδική ταξινόμηση. Αν και δεν είναι τόσο σύνθετος όσο άλλοι αλγόριθμοι ΜΜ, θεωρείται κι αυτός μέρος της επιβλεπόμενης μάθησης (supervised learning). Ο Logistic Regression υπολογίζει την πιθανότητα ένα email να είναι phishing, με έμφαση στην αναγνώριση μαζικών αποστολών και τον εντοπισμό ψευδών διευθύνσεων αποστολέα. Οι συντελεστές που υπολογίζονται για κάθε χαρακτηριστικό (π.χ. το πλήθος των λανθασμένων συνδέσμων σε ένα email ή η ασυνήθιστη διεύθυνση αποστολέα) συνδέονται άμεσα με την πιθανότητα να είναι το email phishing, καθώς αυξάνουν ή μειώνουν την πιθανότητα αυτή ανάλογα με τη συμβολή τους στην εκτίμηση του κινδύνου.

  Πλεονεκτήματα: Επειδή είναι μαθηματικά απλός, είναι γρήγορος και αποδοτικός για μικρά και μεγάλα σύνολα δεδομένων αλλά και εύκολα ερμηνεύσιμος, καθώς προσφέρει σαφείς συντελεστές για κάθε χαρακτηριστικό (feature) που χρησιμοποιείται στην ταξινόμηση.

  Μειονεκτήματα: Υποθέτει ότι υπάρχει γραμμική σχέση μεταξύ των χαρακτηριστικών και του αποτελέσματος, κάτι που δεν ισχύει πάντα στα δεδομένα phishing, οπότε έχει χαμηλότερη ακρίβεια (στις περιπτώσεις μη γραμμικότητας).

- **AdaBoost [58]**: Είναι μια τεχνική ενίσχυσης (boosting) που ενσωματώνει την ΤΝ, καθώς συνδυάζει πολλούς (λογικά αδύναμους) ταξινομητές για να δημιουργήσει έναν ισχυρό ταξινομητή. Έτσι, εντοπίζει χαρακτηριστικά, όπως περίεργες λέξεις-κλειδιά, ασυνήθιστες δομές συνδέσμων, ή μη τυπικά μοτίβα email. Χρησιμοποιείται ευρέως σε εφαρμογές ΤΝ για phishing που βασίζονται σε email επικοινωνία, λόγω της δυνατότητας της να βελτιώνει συνεχώς την ακρίβειά της σε πραγματικά δεδομένα.

  Πλεονεκτήματα: Είναι εύκολο να εφαρμοστεί και έχει ισχυρή θεωρητική βάση. Λειτουργεί επαναληπτικά για να βελτιώνει αδύναμους ταξινομητές, μαθαίνοντας από τα λάθη τους.

  Μειονεκτήματα: Τα δεδομένα με πολλά λάθη ή μεγάλη ανισορροπία μπορεί εύκολα να μειώσουν την ακρίβεια και την αποτελεσματικότητα της τεχνικής αυτής.

- **Spear Phishing:**

  - **Markov Decision Process (MDP) [53]:** Ο MDP δεν αποτελεί αλγόριθμο μηχανικής μάθησης ή τεχνητής νοημοσύνης από μόνος του, αλλά χρησιμοποιείται συχνά σε συνδυασμό με τεχνικές reinforcement learning για τη μοντελοποίηση αλληλεπιδράσεων και τη λήψη αποφάσεων σε καταστάσεις αβεβαιότητας. Στην περίπτωση του Spear Phishing, ο MDP μπορεί να χρησιμοποιηθεί για την ανίχνευση εξατομικευμένων μηνυμάτων μέσω της ανάλυσης της ακολουθίας αλληλεπιδράσεων μεταξύ του χρήστη και του email. Συγκεκριμένα, μπορεί να βοηθήσει στην κατανόηση του προσαρμοσμένου περιεχομένου και στη δημιουργία ενός προτύπου για το τι θεωρείται ύποπτο email, ανιχνεύοντας αλλαγές στη ροή της συνομιλίας ή στο ύφος της. Έτσι, είναι ιδιαίτερα χρήσιμος στην αντιμετώπιση επιθέσεων που συγχωνεύονται με νόμιμες συνομιλίες, όπου οι επιτιθέμενοι εισχωρούν σε υφιστάμενες συνομιλίες για να προσδώσουν στα μηνύματά τους αληθοφάνεια, και στην προσωποποίηση των επιθέσεων.

    Πλεονεκτήματα: Καλή απόδοση σε ανάλυση αλληλεπιδράσεων και δυναμικών συστημάτων, συμβάλλοντας στην ανίχνευση μεταβαλλόμενων, περίπλοκων και προσαρμοσμένων επιθέσεων.

Μειονεκτήματα: Η πολυπλοκότητα των μοντέλων και η απαίτηση σε μεγάλο όγκο δεδομένων για την παραγωγή τους, μπορεί να επιβραδύνει την απόδοση. Αν και αποτελεί μια δυναμική προσέγγιση, απαιτεί υψηλό υπολογιστικό κόστος για να παραμείνει αποτελεσματική σε πραγματικό χρόνο.

- **Reinforcement Learning (RL) [53]:** Αποτελεσματικός στην αντιμετώπιση της εκμετάλλευσης ευάλωτων συναισθημάτων και της εξατομίκευσης των επιθέσεων. Ο RL μαθαίνει δυναμικά από τις αλληλεπιδράσεις με τα δεδομένα και προσαρμόζεται στις νέες τακτικές που χρησιμοποιούν οι επιτιθέμενοι. Χρησιμοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας (NLP), ο RL μπορεί να εντοπίσει συναισθηματικά φορτισμένα μηνύματα και να ανιχνεύσει μοτίβα συναισθηματικής χειραγώγησης (όπως φράσεις ή λέξεις που προκαλούν φόβο, αίσθηση επείγοντος ή πίεσης). Επιπλέον, μπορεί να προσαρμόσει τα μοντέλα του για να εντοπίζει νέες τεχνικές spear phishing.

  Πλεονεκτήματα: Υψηλή προσαρμοστικότητα σε νέες μορφές spear phishing και ικανότητα συνεχούς μάθησης και βελτίωσης της απόδοσης. Παρέχει εξαιρετικά αποτελέσματα σε δυναμικά περιβάλλοντα και είναι κατάλληλος για αντιμετώπιση επιθέσεων με διαδοχικά στάδια.

  Μειονεκτήματα: Απαιτεί μεγάλα σύνολα δεδομένων για εκπαίδευση και χρόνο για να φτάσει σε βέλτιστη απόδοση και σε κατάλληλο επίπεδο προσαρμοστικότητας, ώστε να μπορεί να ανταποκρίνεται σε ποικίλες μορφές επιθέσεων, ακόμη και σε νέες μορφές ή είδη επιθέσεων. Ως επακόλουθο αυτού, έχει υψηλό υπολογιστικό κόστος.

- **Νευρωνικά Δίκτυα (Neural Networks) [53]**: Αποτελούν έναν από τους πιο κλασικούς αλγόριθμους της ΤΝ. Τα Νευρωνικά Δίκτυα είναι εξαιρετικά για την ανίχνευση προσαρμοσμένων μηνυμάτων και κακόβουλων επισυναπτόμενων αρχείων ή συνδέσμων. Τα δίκτυα αυτά αναλύουν το περιεχόμενο, το ύφος και τις πληροφορίες του αποστολέα, ενώ μπορούν να εκπαιδευτούν για να ανιχνεύουν ανωμαλίες στις επικοινωνίες, όπως επισυναπτόμενα αρχεία που μπορεί να περιέχουν κακόβουλο λογισμικό ή συνδέσμους που οδηγούν σε phishing σελίδες. Χρησιμοποιούνται σε εξελιγμένα σενάρια τόσο για phishing όσο και για spear-phishing επιθέσεις.

  Πλεονεκτήματα: Πολύ υψηλή ακρίβεια στην ανάλυση δεδομένων και κατηγοριοποίηση μηνυμάτων σε σύνθετα προβλήματα.

Μειονεκτήματα: Απαιτούν σημαντικούς πόρους για εκπαίδευση και είναι ευαίσθητα σε περιπτώσεις θορυβώδους περιεχομένου με πολλές φορές δυσκολία στην ρύθμιση και ερμηνεία.

- **XGBoost [53]**: Μια πιο εξελιγμένη έκδοση του Gradient Boosting, το XGBoost αποτελεί μία από τις πιο δημοφιλείς τεχνικές στην ΤΝ για την ανίχνευση επιθέσεων κοινωνικής μηχανικής. Είναι ιδιαίτερα αποτελεσματικό για spear phishing επιθέσεις και μπορεί να κλιμακωθεί για μεγαλύτερα δεδομένα. Χρησιμοποιείται ευρέως για ανίχνευση email-based επιθέσεων κοινωνικής μηχανικής λόγω της ταχύτητας και ακρίβειάς του. Ο XGBoost αναγνωρίζει πλαστογραφημένες διευθύνσεις και προσαρμοσμένα μηνύματα, αναλύοντας πολλαπλά χαρακτηριστικά ταυτόχρονα, κάτι που βελτιώνει την ανίχνευση προσαρμοσμένων επιθέσεων spear phishing.

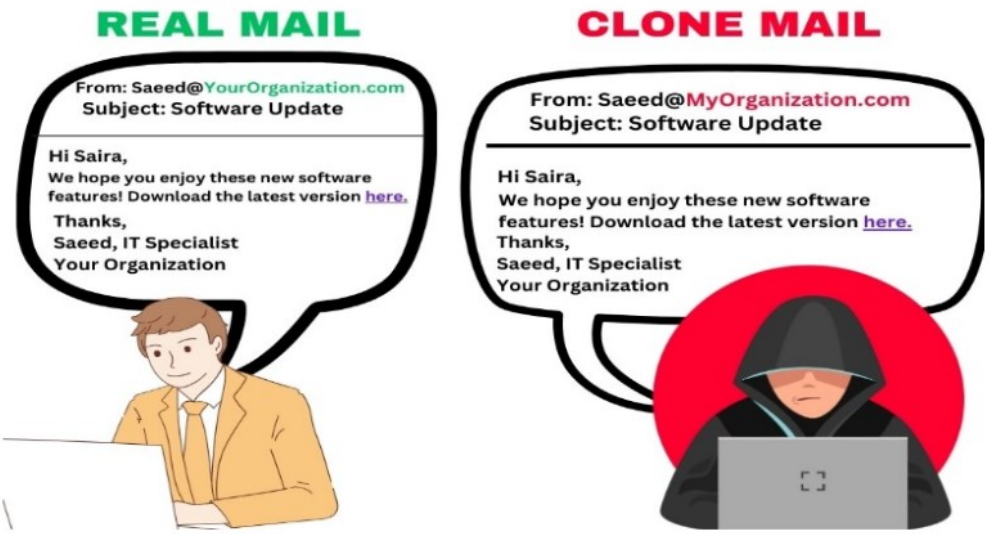
  Πλεονεκτήματα: Πολύ υψηλή ακρίβεια και ταχύτητα εκπαίδευσης, καθώς και δυνατότητα χειρισμού μεγάλων συνόλων δεδομένων (datasets) και περίπλοκων σχέσεων.

  Μειονεκτήματα: Μπορεί να γίνει περίπλοκος στη ρύθμιση και την εφαρμογή, ενώ η ερμηνεία των αποτελεσμάτων του μπορεί να είναι δυσκολότερη σε σχέση με άλλες μεθόδους. Επίσης, οδηγεί σε υψηλή κατανάλωση μνήμης.

- **Clone Phishing (Εικόνα 8):**

  - Αλγόριθμοι Υπολογιστικής Όρασης (Computer Vision Algorithms): Οι απόπειρες ηλεκτρονικού «ψαρέματος» ανιχνεύονται με την ανάλυση οπτικού περιεχομένου, όπως φωτογραφίες, λογότυπα και διατάξεις ιστοτόπων, μέσω της συνδυαστικής χρήσης αλγορίθμων υπολογιστικής όρασης και αλγορίθμων μηχανικής μάθησης (ΜΜ). Ειδικότερα, οι αλγόριθμοι υπολογιστικής όρασης, όπως οι HOG, DAISY, SIFT και SURF, χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών (feature extraction), επιτελώντας το έργο της αναγνώρισης σημαντικών οπτικών χαρακτηριστικών που μπορεί να υποδηλώνουν προσπάθεια απομίμησης. Στη συνέχεια, η ανίχνευση στηρίζεται στη χρήση αλγορίθμων μηχανικής μάθησης, όπως SVM ή k-NN, που αξιοποιούν τα εξαγόμενα χαρακτηριστικά για την αναγνώριση ύποπτων μοτίβων. Αυτή η συνδυαστική τεχνική ΤΝ μπορεί να εντοπίζει εμπορικά σήματα ή στοιχεία σχεδιασμού που μιμούνται σε επιθέσεις Clone Phishing. Ειδικότερα, με την ανάλυση των οπτικών συστατικών ενός ηλεκτρονικού μηνύματος ή ιστότοπου, η Υπολογιστική Όραση μπορεί να εντοπίσει ψεύτικους ιστότοπους ή μηνύματα

ηλεκτρονικού ταχυδρομείου που μιμούνται αυθεντικές πηγές, παρατηρώντας διαφορές στις δομές διάταξης, στα χρώματα της μάρκας ή στα εμβλήματα, που μπορεί να διαφύγουν από τον χρήστη **[55]**.



Εικόνα 8: Προστασία από ηλεκτρονικά μηνύματα - Διάκριση μεταξύ πραγματικών και ψεύτικων μηνυμάτων. Πηγή: Liaqat et al., 2024.

Όλοι οι αλγόριθμοι Υπολογιστικής Όρασης που πρόκειται να εξεταστούν είναι **στατικοί αλγόριθμοι**, καθώς δεν προσαρμόζονται με την πάροδο του χρόνου ή τα νέα δεδομένα **[56]**. Πιο συγκεκριμένα βάσει του **[57],** τέτοιοι στατικοί αλγόριθμοι είναι οι εξεις:

- **HOG (Histogram of Oriented Gradients):** Η κύρια εφαρμογή του περιλαμβάνει τον προσδιορισμό των γεωμετρικών πληροφοριών της ιστοσελίδας, εξετάζοντας τη δομή της εικόνας μέσω της ανάλυσης της κατεύθυνσης κλίσης. Ο HOG μπορεί να χρησιμοποιηθεί για την ανίχνευση κλωνοποιημένων ιστοσελίδων phishing, καθώς μπορεί να διακρίνει παραλλαγές στα οπτικά συστατικά των ιστοσελίδων που προσπαθούν να μιμηθούν αυθεντικές ιστοσελίδες.

  Πλεονεκτήματα: Ανίχνευση αντικειμένων και περιγραμμάτων, ακρίβεια στην οπτική ομοιότητα για την αναγνώριση ψεύτικων ιστοσελίδων που βασίζονται στη μίμηση της εμφάνισης. Εκτός αυτού, επειδή αναλύει οπτικά δεδομένα, ο HOG δεν επηρεάζεται από τη δομή ή το περιεχόμενο του κώδικα της ιστοσελίδας.

Μειονεκτήματα: Σε περιπτώσεις όπου η κλίμακα ή το μέγεθος των αντικειμένων αλλάζει, ο HOG ενδέχεται να μην αποδώσει καλά, ενώ επίσης η υπολογιστική πολυπλοκότητα του μπορεί να είναι υψηλή, ειδικότερα όταν αναλύει ιστοσελίδες σε πραγματικό χρόνο.

- **SIFT (Scale-Invariant Feature Transform):** Ο αλγόριθμος SIFT αποτελεί ένα χρήσιμο εργαλείο για την ανίχνευση ιστοσελίδων clone phishing με βάση την οπτική τους εμφάνιση. Ο σκοπός του είναι να εντοπίζει βασικά σημεία σε μια εικόνα και να παρέχει μια αναλλοίωτη σε κλίμακα και περιστροφή περιγραφή αυτών των σημείων, καθιστώντας τον ανθεκτικό στον εντοπισμό αντικειμένων ή μοτίβων, ακόμη και όταν αυτά κλιμακώνονται, περιστρέφονται ή παραμορφώνονται. Στο πλαίσιο των επιθέσεων clone phishing, ο SIFT χρησιμοποιείται για την ανάλυση οπτικών στοιχείων, όπως τα λογότυπα ή οι οπτικές υπογραφές μέσα σε μηνύματα ή ιστοτόπους. Αυτά τα χαρακτηριστικά μπορεί να έχουν μορφή εικόνας και να περιλαμβάνουν αναγνωριστικά στοιχεία που, αν τροποποιηθούν, μπορούν να παραπλανήσουν τους χρήστες. Η χρήση του SIFT μπορεί να βοηθήσει στην αναγνώριση τέτοιων αλλαγών, διευκολύνοντας την ανίχνευση ψευδών ιστοσελίδων ή μηνυμάτων που αποσκοπούν στην παραπλάνηση θυμάτων.

  Πλεονεκτήματα: Μπορεί να ανιχνεύσει βασικά σημεία σε εικόνες ανεξάρτητα από αλλαγές στο μέγεθος ή τον προσανατολισμό τους, καθιστώντας τον ανθεκτικό στην αναγνώριση ιστοσελίδων clone phishing που μιμούνται νόμιμες. Επιπλέον, όπως αναφέρεται στο άρθρο **[57]**, όταν συνδυάζεται με μοντέλα όπως το XGBoost, ο SIFT μπορεί να επιτύχει πολύ υψηλή ακρίβεια (έως και 89.34% στη μελέτη), καθιστώντας τον ένα πολύ αξιόπιστο εργαλείο.

  Μειονεκτήματα: Απαιτεί σημαντική μνήμη, ειδικά όταν εξάγει χαρακτηριστικά από εικόνες υψηλής ανάλυσης ή επεξεργάζεται μεγάλα σύνολα δεδομένων. Επίσης, σε ορισμένες περιπτώσεις μπορεί να ανιχνεύσει θόρυβο ή άσχετα οπτικά στοιχεία, οδηγώντας σε ψευδώς θετικά αποτελέσματα, ενώ παράλληλα απαιτεί σημαντικούς υπολογιστικούς πόρους.

- **DAISY (Dense Descriptor):** Αποτελεί έναν αλγόριθμο ανίχνευσης χαρακτηριστικών, ίδια φιλοσοφίας με τον SIFT, όμως λειτουργεί με πυκνή δειγματοληψία, που σημαίνει ότι καταγράφει χαρακτηριστικά σε ολόκληρη την εικόνα, αντί να επικεντρώνεται μόνο σε λίγα βασικά σημεία. Ο αλγόριθμος αυτός είναι ιδιαίτερα κατάλληλος για την

ανίχνευση οπτικής ομοιότητας σε ιστοσελίδες clone phishing όπου μπορεί να αντιγράφεται ολόκληρη η διάταξη της σελίδας.

- *Πλεονεκτήματα*: Εξασφαλίζει ότι τα χαρακτηριστικά συλλέγονται σε ολόκληρη την εικόνα, καθιστώντας τον ανθεκτικό για συγκρίσεις πλήρους σελίδας, ενώ αντιμετωπίζει ικανοποιητικά τις περιστροφικές παραμορφώσεις. Ακόμη, είναι Ιδανικός για την ανίχνευση τοπικών διαφοροποιήσεων σε χαρακτηριστικά συνδέσμων και αρχείων.

  Μειονεκτήματα: Πιο αργός όσον αφορά τις εφαρμογές σε πραγματικό χρόνο, καθώς επεξεργάζεται πολύ περισσότερα σημεία δεδομένων λόγω της πυκνής δειγματοληψίας.

- **SURF (Speeded Up Robust Features)**: Είναι ένας αλγόριθμος εντοπισμού χαρακτηριστικών, παρόμοιος με τον SIFT, αλλά σχεδιάστηκε για να απαιτεί μικρότερη υπολογιστική ισχύ. Επικεντρώνεται στην ανίχνευση βασικών σημείων σε εικόνες που είναι λιγότερο επηρεασμένα από θόρυβο.

  Πλεονεκτήματα: Κατάλληλος για ανίχνευση σε πραγματικό χρόνο, λόγω της υψηλής του ταχύτητας. Επιπλέον, είναι αμετάβλητος στην κλίμακα και ανθεκτικός στις αλλαγές μεγέθους ή ζουμ μιας εικόνας.

  Μειονεκτήματα: Έχει χαμηλότερη ακρίβεια για την ανίχνευση σύνθετων οπτικών ομοιοτήτων.

4.2.2.2 Social Media-Based Επιθέσεις

- **Catfishing:**
  - **Generative Adversarial Networks (GANs):** Τα GANs μπορούν να χρησιμοποιηθούν για την ανίχνευση πλαστοπροσωπίας και ψεύτικων προφίλ σε catfishing επιθέσεις. Μέσω της εκπαίδευσης, τα GANs μαθαίνουν να δημιουργούν συνθετικά προφίλ που μοιάζουν με ψεύτικα προφίλ, ώστε να αποκτήσουν την ικανότητα αναγνώρισης χαρακτηριστικών των παραποιημένων προφίλ. Όταν εμφανίζεται ένα ύποπτο προφίλ, τα GANs μπορούν να το συγκρίνουν με τα συνθετικά ψεύτικα προφίλ που έχουν εκπαιδευτεί να παράγουν, διευκολύνοντας έτσι την ανίχνευση παραποιημένων ή κλεμμένων στοιχείων. Αυτή η διαδικασία επιτρέπει την αναγνώριση και αποκάλυψη ψευδών ταυτοτήτων, συγκρίνοντας τα προφίλ των χρηστών με δεδομένα εικόνας και αναγνωρισμένα πρότυπα ανθρώπινης συμπεριφοράς.

Για παράδειγμα, είναι σε θέση να εντοπίσουν τη χρήση ψεύτικων φωτογραφιών ή κλεμμένου περιεχομένου σε προφίλ, βασισμένα στη σύγκριση μεταξύ των εκπαιδευμένων ψεύτικων προφίλ και των προφίλ των επιτιθέμενων **[71].**

Πλεονεκτήματα: Τα GANs είναι δυναμικά και μαθαίνουν να βελτιώνονται με την πάροδο του χρόνου καθώς παράγουν περισσότερα δεδομένα, γεγονός που τα καθιστά ιδιαίτερα χρήσιμα για τον εντοπισμό συνθετικών ταυτοτήτων στο catfishing.

Μειονεκτήματα: Απαιτούν σημαντικούς υπολογιστικούς πόρους και χρόνο για εκπαίδευση **[72]**.

- **Natural Language Processing (NLP):** Η NLP επιτρέπει την ανίχνευση ψυχολογικής χειραγώγησης μέσω της ανάλυσης του ύφους των συνομιλιών και της συναισθηματικής φόρτισης. Μπορεί να εντοπίσει συναισθηματικά ευαίσθητα σημεία σε συνομιλίες που χρησιμοποιούνται για να παγιδεύσουν το θύμα. Τα μοντέλα που παράγονται από την NLP μπορούν να αναλύουν τις αλληλεπιδράσεις των χρηστών για να εντοπίζουν πρότυπα γλώσσας που δεν είναι συνεπή με την πραγματική ανθρώπινη συμπεριφορά **[73].**

  Πλεονεκτήματα: Η NLP τεχνική είναι εξαιρετικά προσαρμοστική, γεγονός που σημαίνει ότι μπορεί να αναλύει τεράστιες ποσότητες περιεχομένου που δημιουργείται από χρήστες σε πραγματικό χρόνο. Μπορεί να ανιχνεύσει γλωσσικές ανωμαλίες ενδεικτικές παραπλανητικής συμπεριφοράς, κάτι ιδιαίτερα χρήσιμο στη διαδικτυακή επικοινωνία, καθιστώντας την κατάλληλη για την ανίχνευση του catfishing.

  Μειονεκτήματα: Ο σημαντικότερος περιορισμός των μοντέλων NLP είναι η εξάρτησή τους από μεγάλα, υψηλής ποιότητας σύνολα δεδομένων με ετικέτες. Η απόδοση τους μπορεί να μειωθεί σε πολύγλωσσα περιβάλλοντα ή όταν εξελιγμένοι επιτιθέμενοι χρησιμοποιούν μη-προδιαγεγραμμένη γλώσσα.

- **Impersonation:**

  - **Biometric Authentication and Deep Learning (BA & DL) [74]**: Τα βιομετρικά στοιχεία χρησιμοποιούνται στη ΤΝ για την υλοποίηση ορισμένων λειτουργιών ταυτοποίησης, όπως η αναγνώριση ομιλίας και προσώπου, ώστε να μπορούν να χρησιμοποιηθούν για να επιβεβαιώσουν ότι ένα άτομο είναι αυτό που λέει ότι είναι. Σε απόπειρες πλαστοπροσωπίας σε πραγματικό χρόνο, όπως κατά τη διάρκεια συνομιλιών βίντεο ή ήχου, τα μοντέλα βαθιάς μάθησης που εκπαιδεύονται σε τεράστια σύνολα δεδομένων ανθρώπινων προσώπων και

φωνών μπορούν να εντοπίσουν μικροσκοπικές διαφορές ή αναντιστοιχίες. Αυτό λειτουργεί ιδιαίτερα καλά κατά την επιβεβαίωση της ταυτότητας σε επικοινωνία σε πραγματικό χρόνο **[74].**
Πλεονεκτήματα: Πολύ ακριβής επαλήθευση ταυτότητας μέσω βιομετρικών δεδομένων, που καθιστά δύσκολο για τους επιτιθέμενους να μιμηθούν ακριβώς τα χαρακτηριστικά του ατόμου που προσπαθούν να πλαστογραφήσουν.

Μειονεκτήματα: Οι βιομετρικές λύσεις απαιτούν κατάλληλη υποδομή και συλλογή δεδομένων, γεγονός που μπορεί να επηρεάσει την ταχύτητα υιοθέτησής τους από πλατφόρμες ή εταιρείες. Επιπλέον, εάν κλαπούν βιομετρικά δεδομένα, όπως τα δεδομένα προσώπου, φωνής ή άλλων βιομετρικών χαρακτηριστικών, που χρησιμοποιούνται για την ταυτοποίηση των χρηστών, είναι αδύνατο να ανανεωθούν, δημιουργώντας κενά στην ασφάλεια.

- **Digital Watermarking and Blockchain [74]:** Οι τεχνικές ψηφιακής υδατογράφησης ενισχυμένες με ΤΝ (πχ. αλγόριθμοι όπως CNNs, GANs & RL) μπορούν να ενσωματώνουν αόρατα υδατογραφήματα σε έγκυρα προσωπικά διαπιστευτήρια (*πχ*. εικόνες ή έγγραφα), βοηθώντας στην πρόληψη μη εξουσιοδοτημένης αντιγραφής. Σε συνδυασμό με την αλυσίδα blockchain, αυτό εξασφαλίζει ότι τα προσωπικά αναγνωριστικά δεν μπορούν εύκολα να παραποιηθούν ή να πλαστογραφηθούν στο διαδίκτυο **[74].** Η χρήση του blockchain προσθέτει ένα επιπλέον επίπεδο ασφάλειας, καθώς κάθε αλλαγή στα δεδομένα ανιχνεύεται και καταγράφεται, διασφαλίζοντας την ακεραιότητα και την αυθεντικότητα των εγγράφων σε πραγματικό χρόνο.

  Πλεονεκτήματα: Ικανότητα διασφάλισης της αυθεντικότητας των δεδομένων μέσω υδατογράφησης και της διασφάλισης ακεραιότητας μέσω blockchain. Το blockchain παρέχει ένα επίπεδο διαφάνειας και αμεταβλητότητας, καθιστώντας την πλαστογράφηση σχεδόν αδύνατη.

  Μειονεκτήματα: Η Ψηφιακή Υδατογράφηση μπορεί να είναι ευάλωτη σε τεχνικές παραποίησης, ενώ το Blockchain απαιτεί υψηλή υπολογιστική ισχύ και πόρους για την πλήρη ενσωμάτωσή του σε συστήματα ανίχνευσης πλαστοπροσωπίας.

- **Credential Harvesting [70], [76]**:

  - **User Behavior Analytics (UBA) [70]:** Οι τεχνικές ανάλυσης συμπεριφοράς χρηστών βασίζονται σε ΤΝ για την παρακολούθηση της συμπεριφοράς ενός χρήστη στο διαδίκτυο. Αυτές οι τεχνικές μπορούν να εντοπίσουν μη

φυσιολογικές συμπεριφορές, όπως προσπάθειες σύνδεσης από άγνωστες τοποθεσίες ή μη φυσιολογικές ώρες χρήσης, που υποδεικνύουν ότι τα διαπιστευτήρια έχουν κλαπεί και χρησιμοποιούνται κακόβουλα. Ωστόσο, προκειμένου να ελαχιστοποιηθεί ο ανεπιθύμητος αποκλεισμός όταν ένας γνήσιος χρήστης ταξιδεύει ή αλλάζει τοποθεσία, το σύστημα μπορεί να ειδοποιήσει τον χρήστη. Ο χρήστης θα είναι σε θέση να επιβεβαιώσει ή να αποκλείσει τη δραστηριότητα, διασφαλίζοντας ότι η νόμιμη πρόσβαση δεν διαταράσσεται, διατηρώντας παράλληλα υψηλό βαθμό ασφάλειας.

- Ο **Isolation Forest [76]** αλγόριθμος ανήκει στις UBA τεχνικές και χρησιμοποιείται για την ανίχνευση ανωμαλιών διαχωρίζοντας τα δεδομένα που διαφέρουν σημαντικά από το υπόλοιπο σύνολο (όπως προσπάθειες σύνδεσης από μη εξουσιοδοτημένες τοποθεσίες ή κατά μία ασυνήθιστη ώρα). Αυτό βοηθά στην ανίχνευση παραβιασμένων λογαριασμών, καθώς οι δραστηριότητες που εντοπίζονται διαφέρουν από τη συνηθισμένη συμπεριφορά σύνδεσης του νόμιμου χρήστη, παρόλο που μπορεί να φαίνονται αρχικά νόμιμες.

  Πλεονεκτήματα: Δεν απαιτεί επισήμανση δεδομένων (labeling) ή την ύπαρξη σαφούς κατηγοριοποίησης. Αυτό σημαίνει ότι μπορεί να λειτουργήσει αυτόνομα σε περιβάλλοντα όπου οι ανωμαλίες είναι σπάνιες και τα δεδομένα είναι μεγάλα. Ο αλγόριθμος έχει χαμηλό υπολογιστικό κόστος και λειτουργεί εξαιρετικά καλά σε περιπτώσεις όπου οι παραδοσιακές μέθοδοι αποτυγχάνουν.

  Μειονεκτήματα: Ο αλγόριθμος μπορεί να εντοπίζει ψευδείς θετικές ανωμαλίες σε περιπτώσεις που υπάρχουν πολλές προσωρινές αλλαγές στη συμπεριφορά των χρηστών, και ενδέχεται να απαιτεί συνεχή προσαρμογή για την επίτευξη υψηλής ακρίβειας.

4.2.2.3 Phone-Based Επιθέσεις

- **Vishing [77]:**

  - **i-Vector** και **x-Vector**: Οι τεχνικές αυτές αποτελούν μεθόδους εξαγωγής χαρακτηριστικών, οι οποίες περιγράφουν τα χαρακτηριστικά της φωνής του ομιλητή. Αυτά τα μοντέλα μπορούν να εκπαιδευτούν με βάση τη φωνή πραγματικών χρηστών οπότε οποιαδήποτε απόκλιση από αυτά τα εκπαιδευμένα χαρακτηριστικά μπορεί να δημιουργήσει συναγερμό σε επιθέσεις vishing. Η εκπαίδευση πραγματοποιείται με αλγορίθμους ΜΜ όπως GMM (Gaussian Mixturel Models) (για i-Vector) και DNN (Deep Neural Networks) (για x-Vector).

Πλεονεκτήματα: Υψηλή ακρίβεια σε βιομετρικά δεδομένα και μεγάλη ανθεκτικότητα σε περιβαλλοντικό θόρυβο. Επίσης, μπορούν να εντοπίσουν διαφορές στη φωνή του καλούντος που δεν είναι εμφανείς στο ανθρώπινο αυτί.

Μειονεκτήματα: Το υψηλό υπολογιστικό κόστος. Αυτές οι τεχνικές απαιτούν σημαντική υπολογιστική ισχύ για την επεξεργασία και εκπαίδευση των δεδομένων. Προφανώς, δεν μπορεί να καλυφθεί το σύνολο των ατόμων που θα μπορούσαν να επικοινωνήσουν με ένα θύμα. Συνεπώς, η ακρίβεια των τεχνικών αυτών δεν μπορεί να είναι μεγάλη. Μπορεί να είναι καλή μόνο για τα άτομα για τα οποία έχουν εκπαιδευτεί.

- **Gaussian Mixture Models (GMM)**: Ο αλγόριθμος GMM (μόνος του σε σχέση με τη συνδυασμένη χρήση με i-Vector) χρησιμοποιείται συχνά στη βιομετρική φωνής για να μοντελοποιεί την κατανομή της φωνής ενός ομιλητή. Στην ανίχνευση επιθέσεων vishing, το GMM μπορεί να βοηθήσει στον εντοπισμό εφόσον η φωνή του ομιλητή δεν αντιστοιχεί στην ισχυριζόμενη ταυτότητα.

  Πλεονεκτήματα: Εξαιρετική απόδοση στην αναγνώριση μοτίβων φωνής, ακόμη και όταν υπάρχουν μικρές διαφορές στην εκφορά του λόγου. Μπορεί να εντοπίσει ανωμαλίες ακόμα και σε περιπτώσεις χαμηλής ποιότητας ήχου.

  Μειονεκτήματα: Ο GMM, λόγω περιορισμένης προσαρμοστικότητας, μπορεί να μην προσαρμόζεται καλά σε συνθήκες σύντομων φωνητικών δειγμάτων. Επίσης, όπως και στις προηγούμενες τεχνικές, δουλεύει καλά μόνο για τις ταυτότητες για τις οποίες έχει εκπαιδευτεί.

- **Recurrent Neural Networks (RNNs):** Τα RNN είναι νευρωνικά δίκτυα που ειδικεύονται στην επεξεργασία διαδοχικών δεδομένων, όπως σήματα ομιλίας και μοτίβα συνομιλίας. Στην περίπτωση των επιθέσεων Vishing, τα RNN μπορούν να εντοπίσουν αλλαγές στον τόνο της ομιλίας, το ύφος και τις παύσεις, κάτι που είναι ζωτικής σημασίας για την ανίχνευση πλαστοπροσωπίας ή αυτοματοποιημένων κλήσεων. Επιπλέον, τα RNN μπορούν να αποθηκεύουν πληροφορίες από προηγούμενες στιγμές σήματος ώστε να αναγνωρίζουν μοτίβα που σχετίζονται με την εξαπάτηση. Για παράδειγμα, στην ανίχνευση πλαστογράφησης της ταυτότητας ενός ομιλητή ή των robocalls, τα RNN μπορούν να αναλύσουν ολόκληρες ακολουθίες ομιλίας, κρατώντας στη μνήμη

προηγούμενα τμήματα της συνομιλίας. Έτσι, αν υπάρχουν μοτίβα που υποδεικνύουν εξαπάτηση, αυτά εντοπίζονται μέσω της αλληλουχίας των στιγμών στο σήμα και όχι απλώς σε μια μεμονωμένη χρονική στιγμή.

Πλεονεκτήματα: Πολύ καλή απόδοση στην ανίχνευση ακολουθιακών μοτίβων και ανωμαλιών στη φωνητική ροή, καθώς και δυνατότητα επεξεργασίας μεγάλων ακολουθιακών δεδομένων.

*Μειονεκτήματα***:** Απαιτούν εκτενή εκπαίδευση και μεγάλα σύνολα δεδομένων για να επιτύχουν υψηλή ακρίβεια, ενώ είναι ευαίσθητα στην ποιότητα των εισερχόμενων δεδομένων.

- **Smishing [78]:**

  - **Naive Bayes**: Αποτελεί έναν πιθανοτικό ταξινομητή βασισμένο στο θεώρημα Bayes. Επειδή έχει καλές επιδόσεις στην ταξινόμηση κειμένου, χρησιμοποιείται συχνά για την ανάλυση του περιεχομένου μηνυμάτων SMS με σκοπό την ανίχνευση επιθέσεων smishing. Προσδιορίζει την πιθανότητα ότι ένα μήνυμα είναι phishing με βάση την ύπαρξη συγκεκριμένων όρων ή φράσεων που συνδέονται συχνά με τέτοιες απόπειρες. Συνεπώς, δίνεται έμφαση στην αποτελεσματικότητα του αλγορίθμου Naive Bayes στο χειρισμό των μεγάλων συνόλων δεδομένων μηνυμάτων SMS, προσφέροντας μια λύση για ανίχνευση σε πραγματικό χρόνο.

    Πλεονεκτήματα: Είναι γρήγορος και αποτελεσματικός για την κατηγοριοποίηση μεγάλου αριθμού μηνυμάτων, ενώ παράλληλα παρέχει υψηλά ποσοστά ακρίβειας για απλούστερα σενάρια.

    *Μειονεκτήματα***:** Μπορεί να εμφανίσει ψευδώς θετικά αποτελέσματα, ειδικά όταν τα μηνύματα έχουν μικρή διαφοροποίηση στο περιεχόμενο. Αυτό σημαίνει ότι αθώα μηνύματα, όπως εκείνα που γράφτηκαν βιαστικά ή περιέχουν ορολογία που παραπέμπει σε smishing, μπορεί να θεωρηθούν ύποπτα από τον αλγόριθμο. Σε τέτοιες περιπτώσεις, η απλότητα του Naive Bayes μπορεί να οδηγήσει σε ανακριβή ταξινόμηση.

  - **Support Vector Machine (SVM):** Ο αλγόριθμος SVM είναι μια εξαιρετική επιλογή για την ανίχνευση Smishing επιθέσεων, καθώς μπορεί να εντοπίσει περίπλοκα μοτίβα σε μεγάλα σύνολα δεδομένων μηνυμάτων, ανιχνεύοντας

μηνύματα SMS τύπου Smishing που μπορεί να είναι δυσδιάκριτα για απλούστερους αλγορίθμους.

Πλεονεκτήματα: Πολύ ακριβής στην ανίχνευση σύνθετων μοτίβων και κατάλληλος για πολύπλοκα σύνολα δεδομένων.

Μειονεκτήματα: Έχει υψηλό υπολογιστικό κόστος και απαιτεί περισσότερο χρόνο για εκπαίδευση και ανάλυση σε σύγκριση με άλλους αλγόριθμους.

4.2.2.4 In-Person Επιθέσεις

- ***Tailgating/Piggybacking [81]*:**
    - **Haar Cascade Classifier:** Πρόκειται για μια κοινή μέθοδο Μηχανικής Όρασης για την ανίχνευση προσώπων και τη διάκριση μεταξύ κινούμενων αντικειμένων. Χρησιμοποιεί σύνολα δεδομένων εκπαίδευσης για τον εντοπισμό χαρακτηριστικών προσώπου, το οποίο είναι ιδιαίτερα αποτελεσματικό στις επιθέσεις tailgating/piggybacking, στις οποίες ένας επιτιθέμενος ακολουθεί έναν εξουσιοδοτημένο χρήστη προκειμένου να αποκτήσει παράνομη πρόσβαση. Αν το δεύτερο πρόσωπο δεν ταιριάζει με τα εκπαιδευμένα δεδομένα, μπορεί να ειδοποιηθεί η ασφάλεια (του κτιρίου του οργανισμού όπου γίνεται η παραβίαση).

      Πλεονεκτήματα: Υψηλή ακρίβεια στην ανίχνευση προσώπων, ιδανική για περιβάλλοντα με κάμερες ασφαλείας.

      Μειονεκτήματα: Μπορεί να έχει δυσκολίες σε συνθήκες χαμηλού φωτισμού ή όταν οι επιτιθέμενοι χρησιμοποιούν τεχνικές κάλυψης του προσώπου.

    - **Background Subtraction Algorithm:** Ο αλγόριθμος αφαιρεί το σταθερό φόντο και ανιχνεύει αντικείμενα που κινούνται στην εικόνα, κάτι που είναι χρήσιμο στην ανίχνευση παραπάνω από ενός ατόμου που περνάει ταυτόχρονα από ένα σημείο πρόσβασης. Μπορεί να ανιχνεύσει περισσότερα από ένα άτομα που κινούνται σε περιορισμένες περιοχές, όπως οι πόρτες ασφαλείας. Χρησιμοποιείται σε συστήματα παρακολούθησης για την καταγραφή εισόδων σε περιοχές με περιορισμένη πρόσβαση. Επεκτάσεις του αλγορίθμου περιλαμβάνουν τη χρήση CNNs, τα οποία αναγνωρίζουν πιο σύνθετα μοτίβα κίνησης, SVMs για προσαρμογή σε διάφορες συνθήκες φωτισμού, και Random Forests για αύξηση της στιβαρότητας σε περιβάλλοντα με υψηλή πολυπλοκότητα.

Πλεονεκτήματα: Πολύ καλή απόδοση στην ανίχνευση κίνησης σε πραγματικό χρόνο και με χαμηλό υπολογιστικό κόστος. Δυνατότητα επεκτάσεων με τεχνικές ΜΜ για περαιτέρω βελτίωση.

Μειονεκτήματα: Η απόδοση μπορεί να επηρεαστεί από την ποιότητα του βίντεο και χρειάζεται ενίσχυση σε περιπτώσεις με υψηλό θόρυβο.

- **Pretexting [82]:**
    - **K-Means Clustering [80]:** Ο αλγόριθμος ομαδοποιεί τη συμπεριφορά των χρηστών σε διάφορα "κέντρα" ή ομάδες με βάση τις κοινές τους δραστηριότητες (πχ. συνηθισμένες ώρες σύνδεσης, αιτήματα πρόσβασης σε δεδομένα, τύποι συσκευών). Αν κάποιος χρήστης ξεκινήσει να ενεργεί με μη συνηθισμένο τρόπο, π.χ. ζητώντας πρόσβαση σε δεδομένα που δεν συνάδουν με τον ρόλο του, ο αλγόριθμος μπορεί να εντοπίσει αυτή την απόκλιση από την ομάδα του και να τη χαρακτηρίσει ως πιθανή απειλή. Οι επιτιθέμενοι που χρησιμοποιούν pretexting συχνά ενεργούν με τρόπο που δεν ταιριάζει με τη φυσιολογική συμπεριφορά ενός χρήστη, οπότε το γεγονός αυτό καθιστά τον αλγόριθμο ιδανικό για την ανίχνευση τέτοιου είδους επιθέσεων.

      Πλεονεκτήματα: Ευκολία στην υλοποίηση, ο αλγόριθμος είναι απλός και μπορεί να υλοποιηθεί εύκολα με μικρή υπολογιστική ισχύ.

      Μειονεκτήματα: Σε πολύπλοκα περιβάλλοντα μπορεί να είναι δύσκολο να καθοριστεί εκ των προτέρων ο αριθμός των συστάδων (clusters). Επιπλέον, όταν τα δεδομένα είναι ασαφώς διαιρεμένα ή πολύπλοκα ή δεν έχουν γραμμικές συσχετίσεις μεταξύ τους, η απόδοση είναι κακή. Τέλος, τα ευρήματα μπορεί να ποικίλουν ανάλογα με το πόσο ευαίσθητες είναι οι αρχικές συνθήκες.

    - **Autoencoders [81]:** Οι επιθέσεις pretexting συχνά περιλαμβάνουν ασυνήθιστες ή σπάνιες ενέργειες από τον επιτιθέμενο που προσπαθεί να εξαπατήσει τον στόχο του. Επειδή οι Autoencoders (δηλαδή ένα είδος νευρωνικών δικτύων με δύο μέρη, τον κωδικοποιητή και τον αποκωδικοποιητή) μπορούν να ανιχνεύσουν συμπεριφορές που δεν έχουν εμφανιστεί στο παρελθόν ή που δεν συνάδουν με τα συνηθισμένα πρότυπα, είναι ιδανικοί για την ανίχνευση επιθέσεων pretexting.

      Πλεονεκτήματα: Έχουν τη δυνατότητα να ανιχνεύουν πολύπλοκα μοτίβα στη συμπεριφορά των χρηστών και να εντοπίζουν με ακρίβεια ανωμαλίες, χωρίς να απαιτούν προκαθορισμένα χαρακτηριστικά, αφού μαθαίνουν αυτόματα από τα

δεδομένα. Είναι κατάλληλοι για μεγάλα σύνολα δεδομένων, εντοπίζοντας μικρές αποκλίσεις που άλλοι αλγόριθμοι ενδέχεται να αγνοήσουν.

Μειονεκτήματα: Απαιτούν μεγάλα σύνολα δεδομένων για την εκπαίδευσή τους και η υπολογιστική ισχύς που απαιτείται είναι υψηλή.

4.2.2.5 Web-Based Επιθέσεις

- **Watering Hole [83]:**

  - **Intrusion Detection System (IDS) με Machine Learning:** Τα συστήματα ανίχνευσης εισβολών (IDS) με δυνατότητες ΜΜ μπορούν να σαρώνουν την κυκλοφορία του δικτύου και να εντοπίζουν ανωμαλίες, όπως η απροσδόκητη μεταφορά δεδομένων, που θα μπορούσαν να υποδηλώνουν μια επίθεση Watering Hole. Τα συστήματα IDS ανιχνεύουν αποκλίσεις από τα κανονικά πρότυπα αναλύοντας τις δικτυακές δραστηριότητες.

    Πλεονεκτήματα: Ανίχνευση σε πραγματικό χρόνο και δυνατότητα αυτόματης απόκρισης σε ύποπτη δραστηριότητα.

    Μειονεκτήματα: Τα IDS μπορεί να παράγουν ψευδώς θετικά αποτελέσματα ενώ απαιτούν συνεχή παρακολούθηση και βελτιστοποίηση.

  - **Neural Network-based Detection:** Γίνεται χρήση νευρωνικών δικτύων για τη ανάλυση της συμπεριφορά ενός ιστοτόπου και την κυκλοφορία του δικτύου. Οι επιθέσεις Watering Hole συχνά χρησιμοποιούν ελαττώματα σε αξιόπιστους ιστότοπους για να παραδώσουν κακόβουλο λογισμικό που οι χρήστες εκφορτώνουν εν αγνοία τους. Το νευρωνικό δίκτυο μπορεί να ανιχνεύσει παρατυπίες στην κυκλοφορία ιστοτόπων καθώς και αποκλίσεις συμπεριφοράς μεταξύ των χρηστών που επισκέπτονται συχνά αυτούς τους ιστότοπους.

    Πλεονεκτήματα: Υψηλή ακρίβεια και δυνατότητα ανίχνευσης λεπτών αποκλίσεων στη συμπεριφορά των ιστότοπων, με εξαιρετική ικανότητα να εντοπίζει τα μοτίβα που συσχετίζονται με κακόβουλες δραστηριότητες.

Μειονεκτήματα: Απαιτεί εκπαίδευση με μεγάλες ποσότητες δεδομένων και μπορεί να παρουσιάζει καθυστερήσεις στην αρχική ανίχνευση εάν τα δεδομένα δεν είναι επαρκή.

- **Malvertising [84]:**

  - **Extra Trees Classifier:** Αποτελεί παραλλαγή του Random Forest που δημιουργεί πολλαπλά δέντρα από τυχαία δείγματα των δεδομένων. Εφαρμόζεται στην ανίχνευση Malvertising μέσω της ανάλυσης συμπεριφοράς των διαφημίσεων της τρέχουσας και της ιστορικής συμπεριφοράς τους, εντοπίζοντας διαφορές που μπορεί να υποδηλώνουν κακόβουλη δραστηριότητα. Μπορεί να ανιχνεύσει κακόβουλες διαφημίσεις που κρύβονται σε μεγάλους διαφημιστικούς όγκους, εξετάζοντας την ανωμαλία στη συμπεριφορά των διαφημίσεων και στην αλληλεπίδραση με τους χρήστες.

    Πλεονεκτήματα: Πολύ καλή απόδοση σε μεγάλους όγκους δεδομένων και ταχύτητα στην ανάλυση, ενώ είναι λιγότερο επιρρεπής στην υπερ-προσαρμογή (overfitting).

    Μειονεκτήματα: Μπορεί να χρειάζεται περισσότερη προσαρμογή για να αποφευχθούν ψευδώς θετικά αποτελέσματα.

  - **Stochastic Gradient Descent (SGD):** Ο συγκεκριμένος αλγόριθμος εκπαιδεύεται πάνω σε σύνολα δεδομένων κακόβουλων διαφημίσεων για να εντοπίζει μοτίβα που συνδέονται με τέτοιου είδους επιθέσεις. Χρησιμοποιείται για την ανίχνευση ύποπτης συμπεριφοράς σε διαφημίσεις, όπως μη φυσιολογική αλληλεπίδραση με τους χρήστες, αλλαγές σε URLs ή άλλα στοιχεία που σχετίζονται με τη διαφημιστική καμπάνια.

    Πλεονεκτήματα: Γρήγορος και αποδοτικός σε μεγάλα σύνολα δεδομένων, ιδιαίτερα χρήσιμος όταν τα δεδομένα είναι πολυδιάστατα.

    Μειονεκτήματα: Μπορεί να απαιτεί πιο προσεκτική ρύθμιση των παραμέτρων του για να επιτύχει υψηλή ακρίβεια και να αποφευχθούν υπερβολικά ψευδώς θετικά αποτελέσματα.

4.2.3 Σύγκριση Αλγορίθμων ΤΝ

Η ενότητα αυτή πραγματοποιεί μια συγκριτική ανάλυση των τεχνικών ΤΝ που χρησιμοποιούνται κατά των επιθέσεων κοινωνικής μηχανικής με βάση ένα σύνολο από κριτήρια. Η ανάλυση πραγματοποιείται μέσω της χρήσης πινάκων που επιδεικνύουν τα αποτελέσματα της σύγκρισης ανά κατηγορία της διάστασης του μέσου επικοινωνίας των επιθέσεων. Με αυτό τον τρόπο, μπορεί να προκύψουν συμπεράσματα που να εστιάζουν σε κάθε κατηγορία της διάστασης αυτής ή να είναι καθολικά. Επιπλέον, για κάθε κατηγορία επιθέσεων, συμπεριλήφθηκε ένα γράφημα βασισμένο στους πίνακες, που επιτρέπει την οπτική σύγκριση των αλγορίθμων, με βάση τα κριτήρια σύγκρισης των τεχνικών ΤΝ, που προσδιορίζονται παρακάτω.

- **Τύπος**: Κατηγοριοποιεί μια τεχνική ΤΝ ως δυναμική ή στατική. Οι δυναμικές τεχνικές προσαρμόζονται σε πραγματικό χρόνο σε νέες πληροφορίες, ενώ οι στατικές τεχνικές βασίζονται στη παραγωγή και χρήση ενός σταθερού μοντέλου που δεν αλλάζει με την πάροδο του χρόνου ή τις νέες επιθέσεις εκτός και αν επανεκπαιδευτεί.

- **Αποτελεσματικότητα (Ακρίβεια)**: Αξιολογεί πόσο καλά η τεχνική ΤΝ μπορεί να ανιχνεύσει με ακρίβεια επιθέσεις κοινωνικής μηχανικής. Υψηλή ακρίβεια σημαίνει λιγότερα ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα. Χαμηλή ακρίβεια, από την άλλη πλευρά, σημαίνει ότι ο αλγόριθμος κάνει συχνά λάθος, είτε αποτυγχάνοντας να ανιχνεύσει πραγματικές επιθέσεις (ψευδώς αρνητικά) είτε εντοπίζοντας κανονικές δραστηριότητες ως επιθέσεις (ψευδώς θετικά). Ειδικότερα στο κριτήριο της αποτελεσματικότητας (το οποίο βάση προσωπικής κρίσης αποτελεί θεμέλιο λίθο της αξιολόγησης αλγορίθμων, αφού διασφαλίζει την αξιοπιστία των αποτελεσμάτων), η ποσοτική αντιστοίχιση που θα γίνει αργότερα, βασίζεται σε συγκεκριμένες τιμές ποσοστών ακρίβειας, καθώς αυτές προέρχονται από δημοσιευμένα αποτελέσματα σε επιστημονικά άρθρα. Για τα υπόλοιπα κριτήρια, η ποσοτική αντιστοίχιση γίνεται αρχικώς σε εύρη τιμών.

  Κλίμακα αξιολόγησης:

  - *Πολύ Υψηλή*: Ελάχιστα έως καθόλου ψευδώς θετικά/αρνητικά.
  - *Υψηλή*: Ικανοποιητικά καλός εντοπισμός με λίγα λάθη.
  - *Μέτρια*: Εμφάνιση μέτριου αριθμού από λάθη.
  - *Χαμηλή*: Συχνές αποκλίσεις στα αποτελέσματα.
  - *Πολύ χαμηλή*: Πολύ ασταθής ανίχνευση με πολλά λάθη.

➢ **Υπολογιστικό Κόστος**: Εκτιμά τον αριθμό των πόρων (υπολογιστική ισχύς, κύρια μνήμη) που χρειάζεται η τεχνική ΤΝ για να εκτελεστεί. Υψηλό κόστος σημαίνει περισσότερη κατανάλωση πόρων. Από την άλλη, χαμηλό υπολογιστικό κόστος σημαίνει ότι ο αλγόριθμος μπορεί να λειτουργήσει αποτελεσματικά με περιορισμένους πόρους.

Κλίμακα αξιολόγησης:

- *Πολύ Χαμηλό*: Ελάχιστοι υπολογιστικοί πόροι.
- *Χαμηλό*: Επαρκής απόδοση με περιορισμένους πόρους.
- *Μέτριο*: Λογική κατανάλωση πόρων.
- *Υψηλό*: Αυξημένες απαιτήσεις σε επεξεργαστική ισχύ και μνήμη.
- *Πολύ Υψηλό*: Πολύ μεγάλη κατανάλωση πόρων οπότε υπάρχει απαίτηση για ισχυρή υποδομή. Καταλληλότητα μόνο για εξειδικευμένα συστήματα (Παραδείγματα περιλαμβάνουν συστήματα ανίχνευσης σε πραγματικό χρόνο για μεγάλης κλίμακας δίκτυα ή πλατφόρμες με τεχνολογίες που απαιτούν υψηλή υπολογιστική ισχύ, όπως κέντρα δεδομένων υπολογιστικού νέφους (cloud data centers), κυβερνητικά συστήματα ασφαλείας ή μεγάλα κέντρα ελέγχου βιομηχανικών εγκαταστάσεων).

➢ **Κλιμάκωση**: Στο πλαίσιο αυτό, η κλιμάκωση δεν περιλαμβάνει μόνο την ποσότητα των δεδομένων αλλά και την πολυπλοκότητά τους καθώς και την ταχύτητα επεξεργασίας τους. Αναφέρεται συγκεκριμένα στην ικανότητα της μεθόδου να διαχειρίζεται αυξανόμενο όγκο δεδομένων και ολοένα και πιο περίπλοκες επιθέσεις χωρίς να υφίσταται αισθητή επιδείνωση των επιδόσεων. Η υψηλή κλιμακωσιμότητα υποδηλώνει ότι η προσέγγιση μπορεί να αντιδράσει γρήγορα καθώς και να χειριστεί καλά μεγαλύτερα σύνολα δεδομένων ακόμα και στην περίπτωση πιο περίπλοκων απειλών. Από την άλλη πλευρά, η χαμηλή κλιμακωσιμότητα καθιστά τη μέθοδο λιγότερο χρήσιμη για συστήματα μεγάλης κλίμακας, όπου η ταχύτητα επεξεργασίας είναι ζωτικής σημασίας, επειδή η απόδοσή της μειώνεται δραστικά καθώς αυξάνεται ο όγκος των δεδομένων ή η πολυπλοκότητα των επιθέσεων.

Κλίμακα αξιολόγησης:

- *Πολύ Υψηλή:* Ο αλγόριθμος διαχειρίζεται πολύ μεγάλα δεδομένα και σύνθετες επιθέσεις χωρίς επιπτώσεις στην απόδοση, με ταχύτατη επεξεργασία.
- *Υψηλή:* Ικανοποιητική απόδοση σε αυξανόμενα δεδομένα με λίγες αποκλίσεις. Ο αλγόριθμος διατηρεί καλή ταχύτητα και προσαρμόζεται σε πιο περίπλοκα δεδομένα.

- *Μέτρια*: Μειωμένη απόδοση σε μεγάλα δεδομένα ή σύνθετες επιθέσεις οπότε ο αλγόριθμος/τεχνική απαιτεί βελτιώσεις για να παραμείνει αποδοτικός. Διαφορετικά, υπάρχει λογική ταχύτητα επεξεργασίας.
- *Χαμηλή*: Σημαντική μείωση απόδοσης με περιορισμένη ταχύτητα (επεξεργασίας) καθώς αυξάνονται τα δεδομένα ή η πολυπλοκότητα των επιθέσεων.
- *Πολύ χαμηλή:* Ο αλγόριθμος ουσιαστικά δεν μπορεί να διαχειριστεί μεγάλα σύνολα δεδομένων ή περίπλοκες επιθέσεις.

➢ **Προσαρμοστικότητα**: Αναφέρεται στο κατά πόσο ένας αλγόριθμος μπορεί να προσαρμόζεται σε διαφορετικά περιβάλλοντα, απαιτήσεις και νέες επιθέσεις ή περιορισμούς.

Κλίμακα αξιολόγησης:

- *Πολύ Υψηλή*: Ο αλγόριθμος διαθέτει εξελιγμένα χαρακτηριστικά ευελιξίας που του επιτρέπουν να προσαρμόζεται αυτόματα ή πολύ εύκολα σε διαφορετικά περιβάλλοντα, ενώ παράλληλα ανιχνεύει και αντιμετωπίζει νέες απειλές.
- *Υψηλή*: Ο αλγόριθμος μπορεί να λειτουργεί αποτελεσματικά σε ποικιλία περιβαλλόντων και προσαρμόζεται με σχετική ευκολία σε νέες απειλές.
- *Μέτρια*: Ο αλγόριθμος έχει κάποια ευελιξία ώστε να χρησιμοποιείται σε έναν συγκεκριμένο αριθμό διαφορετικών περιβαλλόντων και να αντιμετωπίζει ορισμένες μόνο νέες απειλές. Μπορεί να προσαρμοστεί σε αλλαγές, αλλά απαιτεί κάποια περιοδική ρύθμιση και συντήρηση για να παραμένει αποτελεσματικός σε νέες συνθήκες.
- *Χαμηλή*: Ο αλγόριθμος είναι κατάλληλος για περιορισμένα περιβάλλοντα με συγκεκριμένες απαιτήσεις και αντιμετωπίζει δυσκολίες σε περιβάλλοντα με μεταβαλλόμενες συνθήκες. Χρειάζεται κάποια αναδιαμόρφωση για να προσαρμοστεί σε νέες απειλές.
- *Πολύ χαμηλή*: Ο αλγόριθμος ή η τεχνική είναι σχεδιασμένη για ένα συγκεκριμένο περιβάλλον και λειτουργεί μόνο υπό συγκεκριμένες προϋποθέσεις. Δυσκολεύεται να προσαρμοστεί σε νέες απειλές και απαιτεί σημαντικές τροποποιήσεις για να ενσωματώσει αλλαγές.

Για την καλύτερη εποπτεία των κριτηρίων αξιολόγησης, οι ποιοτικές τιμές που αφορούν όλους τους παραπάνω δείκτες των τεχνικών ΤΝ έχουν αρχικά αντιστοιχιστεί σε ποσοτικά εύρη τιμών. Με αυτόν τον τρόπο, καθίσταται δυνατή η ποσοτικοποίηση των δεδομένων και η οπτική σύγκριση των τεχνικών σε κάθε κριτήριο. Η κατεύθυνση της αντιστοίχισης βασίζεται στη φύση ή κατεύθυνση τιμών κάθε κριτηρίου:

- *θετική κατεύθυνση* (*πχ*. για το κριτήριο της κλιμάκωσης): Χαμηλές ποιοτικά τιμές αντιστοιχούν σε χαμηλά ποσοτικά εύρη (πχ. πολύ χαμηλή -> [0,20)) ενώ υψηλές ποιοτικά τιμές αντιστοιχούν σε υψηλά ποσοτικά εύρη (*πχ*. πολύ υψηλή -> (80,100]). Οπότε, έχουμε την εξής ολοκληρωμένη αντιστοίχιση σε αυτή την περίπτωση:
    - Πολύ Χαμηλή [0-20)
    - Χαμηλή [20-40)
    - Μέτρια [40-60)
    - Υψηλή [60-80)
    - Πολύ υψηλή [80-100]
- *αρνητική κατεύθυνση* (*πχ*. για το κριτήριο του υπολογιστικού κόστους): Χαμηλές ποιοτικά τιμές αντιστοιχούν σε υψηλά ποσοτικά εύρη (*πχ*. πολύ χαμηλή -> (80,100)) ενώ υψηλές ποιοτικά τιμές αντιστοιχούν σε χαμηλά ποσοτικά εύρη (*πχ*. πολύ υψηλή -> [0,20]). Οπότε, έχουμε την εξής ολοκληρωμένη αντιστοίχιση σε αυτή την περίπτωση:
    - Πολύ Χαμηλή [80-100]
    - Χαμηλή [60-80)
    - Μέτρια [40-60)
    - Υψηλή [20-40)
    - Πολύ υψηλή [0-20)

Με βάση την παραπάνω αντιστοίχιση ποιοτικών τιμών σε ποσοτικά εύρη, για να έχουμε μια πιο ακριβή σύγκριση των τεχνικών ΤΝ και να παράγουμε πιο εξειδικευμένα αποτελέσματα, αξιολογούμε κάθε τεχνική ΤΝ ως προς το εκάστοτε κριτήριο (εκτός από αυτό της αποτελεσματικότητας) με μια συγκεκριμένη τιμή από το αντιστοιχισμένο ποσοτικό εύρος. Η αξιολόγηση αυτή βασίζεται στην ακόλουθη λογική, η οποία επεξηγείται με βάση το παράδειγμα του ποσοτικού εύρους [0,20) για την “Πολύ χαμηλή” ποιοτική τιμή με θετική κατεύθυνση κλίμακα αξιολόγησης:

- Όταν ένας αλγόριθμος ΤΝ αποδεδειγμένα (με βάση τις διάφορες βιβλιογραφικές πηγές) βρίσκεται στο συγκεκριμένο εύρος χωρίς κάποια διακύμανση (δηλ. αναφορά για επίτευξη καλύτερης ποιοτικής τιμής από την τρέχουσα για το τρέχων κριτήριο από τον αλγόριθμο), επιλέγουμε τιμή κοντά στο κατώτατο όριο. Με βάση το τρέχον παράδειγμα, εφόσον ο αλγόριθμος θα έχει πάντοτε την “Πολύ χαμηλή” ποιοτική τιμή στην βιβλιογραφία, η ποσοτική τιμή αντιστοίχισής του θα είναι το 5 (ουσιαστικά κινούμαστε σε πολλαπλάσια του 5 στην αντιστοίχισή μας).
- Αν ο αλγόριθμος αναφέρεται με διαβαθμισμένη απόδοση ως προς το κριτήριο (ορισμένες αναφορές τον θέτουν στην τρέχουσα ποιοτική τιμή και άλλες στην αμέσως επόμενη), επιλέγεται η τιμή στο κέντρο του εύρους, συγκεκριμένα στο μέσον του. Με βάση το τρέχων παράδειγμα, η τιμή αυτή θα είναι το 10 (εφόσον ο αλγόριθμος έχει “Πολύ χαμηλή τιμή” σε ορισμένες αναφορές και “Χαμηλή τιμή” σε άλλες).
- Όταν η απόδοση τείνει προς την επόμενη ποιοτική τιμή αλλά υπάρχουν λίγες αναφορές που υποδεικνύουν την τρέχουσα, αντιστοιχίζεται τιμή κοντά στο ανώτατο όριο. Στο

τρέχων παράδειγμα, η τιμή θα είναι η 15 (εφόσον τις πιο πολλές φορές ο αλγόριθμος έχει “Χαμηλή τιμή” και “Χαμηλή ελάχιστη τιμή” σε ορισμένες).

Σε περιπτώσεις που εμφανίζονται διακυμάνσεις σε ποιοτικές τιμές, που αγγίζουν τρία επίπεδα και όχι δύο (πχ. έχουμε χαμηλή, μέτρια και υψηλή απόδοση σε μερικές περιπτώσεις/πηγές για μια τεχνική ΤΝ), τότε για την ποσοτικοποίηση των αποτελεσμάτων, επιλέγεται το μέσον του διαστήματος για την τελική τιμή (πχ. 10 για το τρέχων παράδειγμα).

Η επιλογή αυτή βασίζεται σε έναν διαισθητικό κανόνα που ακολουθείται με συνέπεια στις εκάστοτε κλίμακες αξιολόγησης όλων των κριτηρίων (εκτός της αποτελεσμαστικότητας) με τον ίδιο ακριβώς τρόπο, βάση της βιβλιογραφικής ανασκόπησης. Με αυτόν τον τρόπο, διασφαλίζεται ότι η κάθε τελική τιμή αποτυπώνει την ακριβή θέση του αλγορίθμου εντός του εκάστοτε εύρους.

4.2.3.1 Αποτελέσματα Σύγκρισης για Τεχνικές ΤΝ που εστιάζουν σε Email-Based Επιθέσεις

Ο παρακάτω πίνακας σύγκρισης παράχθηκε για τις τεχνικές ΤΝ με βάση το περιεχόμενο των άρθρων **[34]**, **[53], [54], [55], [56], [57], [58], [59], [60]** και τα κριτήρια που έχουν τεθεί.

Πίνακας 2: Συγκριτική Ανάλυση Τεχνικών ΤΝ κατά των Email-Based Επιθέσεων ΚΜ.

Τεχνική	Τύπος	Αποτελεσματικότητα (Ακρίβεια)	Υπολογιστικό Κόστος	Κλιμάκωση	Προσαρμοστικότητα
SVM	Στατική	Υψηλή, με μικρές αποκλίσεις σε μεγάλα σύνολα δεδομένων (ιδιαίτερα καλός για την	Υψηλό (υπολογιστικά δαπανηρό, λόγω υπολογιστικής πολυπλοκότητας)	Μέτρια, δεν λειτουργεί καλά σε πολύ μεγάλα	Μέτρια (Μπορεί να προσαρμοστεί σε κάποιο βαθμό με περιοδική ρύθμιση, με καλύτερες επιδόσεις σε περιβάλλοντα με ελάχιστες αλλαγές)

		ανίχνευση phishing)		σύνολα δεδομένων	
Random Forest	Στατική	Πολύ υψηλή (Σχεδόν καθόλου ψευδώς θετικά/αρνητικά σε ποικίλες επιθέσεις κοινωνικής μηχανικής)	Μέτριο, λόγω της χρήσης πολλαπλών δέντρων	Υψηλή, ικανότητα επεξεργασίας μεγάλων συνόλων δεδομένων χωρίς μείωση της απόδοσης	Υψηλή (Προσαρμόζεται καλά σε διάφορες συνθήκες, χρήσιμο για την ανίχνευση νέων απειλών σε πολλαπλά πλαίσια λόγω της ευέλικτης δομής του συνόλου)
Decision Tree	Στατική	Μέτρια (επιρρεπής σε υπερπροσαρμογή, κατάλληλη για μικρότερα σύνολα δεδομένων)	Χαμηλό, λόγω της απλότητας του μοντέλου	Χαμηλή, δυσκολεύεται σε πολύπλοκα και μεγάλα δεδομένα	Χαμηλή (Περιορίζεται σε στατικά περιβάλλοντα-δυσκολεύεται με την προσαρμογή σε εξελισσόμενες ή σύνθετες απειλές χωρίς αναδιαμόρφωση)
KNN	Στατική	Υψηλή (μεγάλη ακρίβεια στις επιθέσεις phishing)	Υψηλό (απαιτεί σύγκριση με όλα τα δεδομένα)	Πολύ χαμηλή, δεν μπορεί να αντιμετωπίσει τη πολυπλοκότητα στα μεγάλα σύνολα δεδομένων	Πολύ χαμηλή (Σχεδιασμένη για συγκεκριμένες εργασίες με ελάχιστη προσαρμοστικότητα-αντιμετωπίζει προκλήσεις σε δυναμικά ή υψηλής διάστασης περιβάλλοντα)

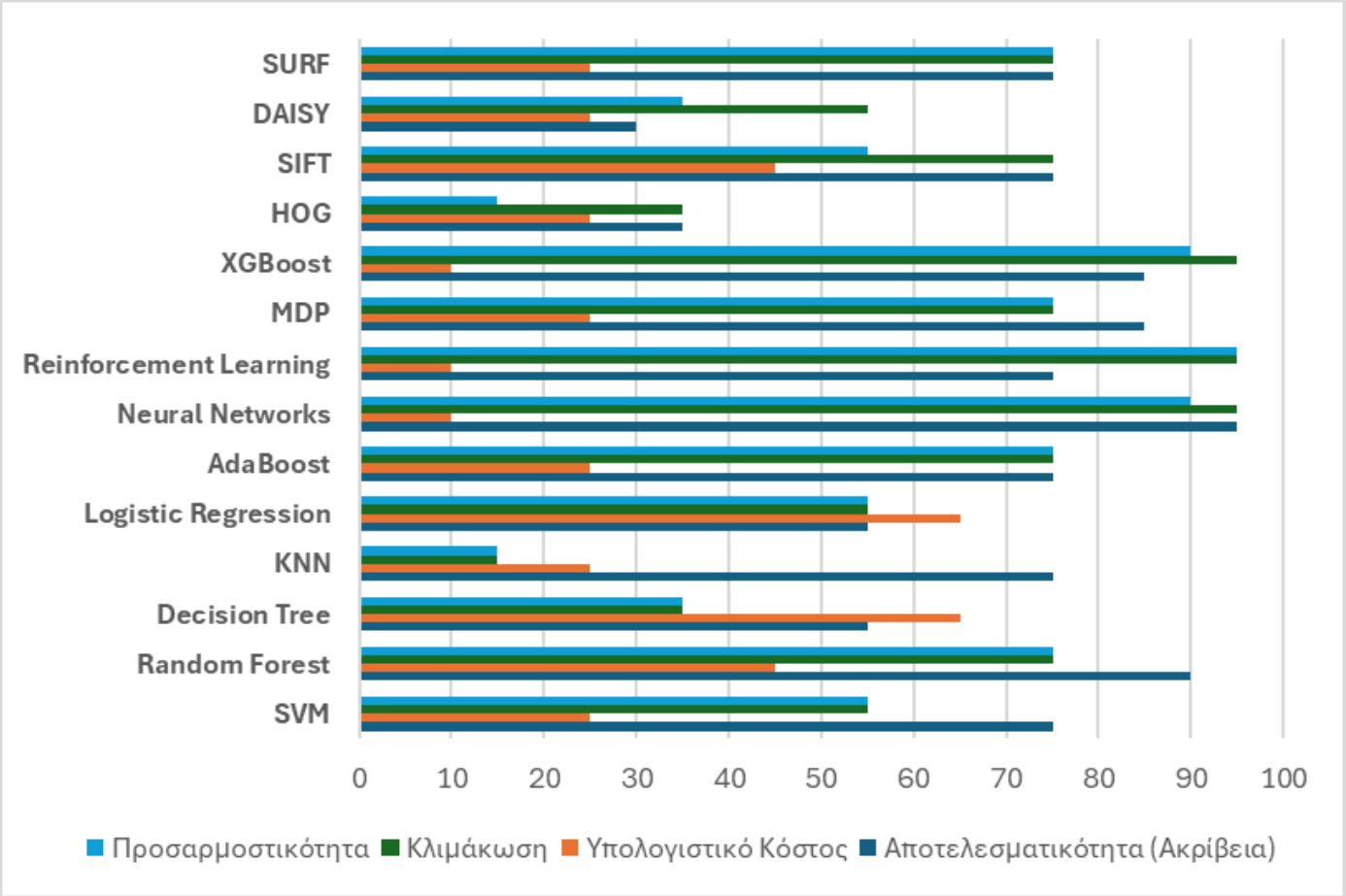
Logistic Regression	Στατική	Μέτρια (καλή γενικά ακρίβεια με ελαφρώς αυξημένα ψευδώς αρνητικά σε μεγαλύτερα σύνολα δεδομένων)	Χαμηλό, λόγω της απλότητας του μοντέλου	Μέτρια, διότι μπορεί να αντιμετωπίσει επαρκώς δεδομένα μέχρι ενός συγκεκριμένου μεγέθους	Μέτρια (Παρέχει κάποια προσαρμοστικότητα, αλλά απαιτεί περιοδική αναπροσαρμογή σε μεταβαλλόμενα περιβάλλοντα)
AdaBoost	Δυναμική	Υψηλή (με μειωμένα ψευδώς θετικά/αρνητικά σε δυναμικά περιβάλλοντα)	Υψηλό, λόγω της ανάγκης διαδοχικής βελτίωσης μοντέλων	Υψηλή, ιδανική για μεγάλα σύνολα δεδομένων	Υψηλή (Αρκετά ευέλικτη για διάφορα περιβάλλοντα, προσαρμόζεται με ευκολία σε μέτριες μεταβολές απειλών)
Neural Networks	Δυναμική	Πολύ Υψηλή (ιδιαίτερα σε ανάλυση μεγάλου όγκου δεδομένων, ανίχνευση προσαρμοσμένων επιθέσεων)	Πολύ υψηλό (απαιτεί σημαντικούς υπολογιστικούς πόρους για την εκπαίδευση και χρήση τους)	Πολύ υψηλή (κατάλληλη για μεγάλα σύνολα δεδομένων και πολύπλοκα σενάρια επιθέσεων)	Πολύ υψηλή (Εξαιρετικά προσαρμοστική, κατάλληλη για δυναμικά, πολύπλοκα περιβάλλοντα, όπου μπορεί να ανιχνεύει εξελισσόμενες απειλές χωρίς επαναδιαμόρφωση)

Reinforcement Learning	Δυναμική	Υψηλή (δυναμική προσαρμογή)	Πολύ Υψηλό, υπολογιστικά δαπανηρό, εξαιτίας της ανάγκης των βρόγχων για ανατροφοδότηση	Πολύ υψηλή, μπορεί να διαχειριστεί αποτελεσματικά αυξανόμενα σύνολα δεδομένων	Πολύ υψηλή (Μαθαίνει και προσαρμόζεται συνεχώς, ιδανική για περιβάλλοντα με μεταβαλλόμενες συνθήκες και εξελισσόμενες απειλές)
MDP	Δυναμική	Πολύ Υψηλή (προσαρμοστική στρατηγική), ικανή να ανιχνεύσει σύνθετες-διαδοχικές επιθέσεις με ακρίβεια	Μέτριο	Υψηλή	Υψηλή (Ικανή να χειρίζεται διαδοχικές αποφάσεις με προσαρμοστικότητα σε νέες απειλές σε δυναμικά περιβάλλοντα)
XGBoost	Δυναμική	Πολύ υψηλή (ανθεκτική σε σύνθετες επιθέσεις, όπως το spear phishing)	Πολύ Υψηλό (απαιτείται ισχυρή υπολογιστική ισχύς λόγω του μεγάλου αριθμού παραμέτρων και της πολυπλοκότητας)	Πολύ Υψηλή (χειρίζεται εξαιρετικά μεγάλα σύνολα δεδομένων και σύνθετες επιθέσεις)	Πολύ υψηλή (Βελτιστοποιημένη για προσαρμοστικότητα, προσαρμόζεται καλά σε πολύπλοκα δεδομένα και μεταβολές απειλών)
HOG	Στατική	Χαμηλή	Υψηλό λόγω επεξεργασίας gradients	Χαμηλή σε μικρά και μεσαία σύνολα	Πολύ χαμηλή (Έχει ελάχιστη ευελιξία, κατάλληλη μόνο για συγκεκριμένες εργασίες με

				δεδομένων, πολύ χαμηλή σε αλλαγές κλίμακας & μεγαλύτερα σύνολα.	περιορισμένες δυνατότητες προσαρμογής)
SIFT	Στατική	Υψηλή (ιδιαίτερα αποτελεσματική για την ανάλυση οπτικών δεδομένων)	Μέτριο, υψηλό σε μεγάλα δεδομένα, λόγω ανάλυσης εικόνας	Υψηλή (ανθεκτική σε συνδυασμό με XGBoost)	Μέτρια (Κάποια προσαρμοστικότητα, αποτελεσματική για ένα εύρος περιβαλλόντων, αλλά απαιτεί ρύθμιση για τον χειρισμό σύνθετων ή εξελισσόμενων απειλών)
DAISY	Στατική	Χαμηλή	Υψηλό λόγω πυκνής δειγματοληψίας & αυξημένης ανάγκης για πόρους επεξεργασίας	Μέτρια	Χαμηλή (Περιορισμένη προσαρμοστικότητα, αποτελεσματική κυρίως σε στατικά περιβάλλοντα με χαμηλή μεταβλητότητα)
SURF	Στατική	Υψηλή (αποτελεσματική για την ανίχνευση phishing σε πραγματικό χρόνο)	Υψηλό, απαιτεί σημαντική υπολογιστική ισχύ λόγω των πολύπλοκων υπολογισμών	Υψηλή, μπορεί να χρησιμοποιηθεί σε μεγαλύτερα σύνολα δεδομένων χωρίς μεγάλη απώλεια απόδοσης.	Υψηλή (Προσαρμόζεται καλά σε πολλαπλά σενάρια, διατηρώντας την απόδοση σε δυναμικά περιβάλλοντα με μεταβαλλόμενα προφίλ απειλών)

Γράφημα Ανάλυσης Άρθρων **[34]**, **[53], [54], [55], [56], [57], [58], [59], [60], [62].**

Μέθοδοι Κατά Των Email-Based Επιθέσεων:



Εικόνα 9: Οπτική αναπαράσταση της απόδοσης των τεχνικών ΤΝ κατά των Email-Based Επιθέσεων ΚΜ.

Ακολουθεί επεξηγηματικό κείμενο σχολιασμού του γραφήματος, με τα αντίστοιχα συμπεράσματα:

1. ***Αποτελεσματικότητα (Ακρίβεια)***: Οι αλγόριθμοι Neural Networks (NNs) και Random Forest κατατάσσονται με πολύ υψηλή ακρίβεια (95% για τα NNs & 90% για τον Random Forest). Τα Νευρωνικά Δίκτυα επιτυγχάνουν υψηλή ακρίβεια με τον αποτελεσματικό χειρισμό πολύπλοκων, μεγάλου όγκου δεδομένων και ο Random Forest ακολουθεί με μεγάλη ακρίβεια λόγω της προσέγγισης ensemble που εφαρμόζει (όπου συνδυάζει πολλαπλά δέντρα αποφάσεων για να βελτιώσει τη συνολική αξιοπιστία ανίχνευσης), η οποία μειώνει τα σφάλματα σε ποικίλες απειλές κοινωνικής μηχανικής. Παράλληλα, οι MDP & XGBoost έρχονται αμέσως μετά με ακρίβεια μόλις 5% χαμηλότερη (85%), καθώς ανταπεξέρχονται εξίσου καλά σε μεγάλα datasets. Αντίθετα, οι HOG και DAISY κατατάσσονται στη χαμηλότερη θέση σε ακρίβεια - ο

DAISY απαιτεί υποστήριξη από άλλους αλγορίθμους (πχ. SVM ή k-NN) για τη μείωση των υψηλών ψευδώς αρνητικών αποτελεσμάτων, ενώ ο HOG δυσκολεύεται με την ορθή ταξινόμηση δεδομένων, οδηγώντας σε συχνές αποκλίσεις στα αποτελέσματα, σε περιβάλλοντα μεγάλης κλίμακας. Παρόλα αυτά, εξακολουθεί να είναι πιο αξιόπιστος από το DAISY (35% σε σύγκριση με 30%) για την ανίχνευση απλών οπτικών ενδείξεων που σχετίζονται με ορισμένους τύπους επιθέσεων κοινωνικής μηχανικής.

2. ***Υπολογιστικό Κόστος***: Παρατηρούμε ότι οι αλγόριθμοι Logistic Regression και Decision Tree είναι οι πιο αποδοτικοί όσον αφορά το Υπολογιστικό Κόστος, ιδανικοί για συστήματα με περιορισμένους πόρους. Οι Neural Networks, Reinforcement Learning & XGBoost έχουν το υψηλότερο κόστος, καθώς απαιτούν μεγάλες υπολογιστικές υποδομές για να αποδώσουν βέλτιστα.
3. ***Κλιμάκωση***: Οι αλγόριθμοι Neural Networks, Reinforcement Learning & XGBoost αντιδρούν γρήγορα και είναι ιδανικοί για σύνθετες επιθέσεις και μεγάλα δεδομένα, καθιστώντας τους την καλύτερη επιλογή για την κλιμάκωση (με ποσοστό 95% και για τους τρεις). Τα Νευρωνικά Δίκτυα και η Ενισχυτική Μάθηση, ειδικότερα, αξιοποιούν πολυεπίπεδες και προσαρμοστικές δομές, οι οποίες τους επιτρέπουν να επεξεργάζονται αποτελεσματικά τεράστιους όγκους δεδομένων, χωρίς αισθητή επιδείνωση των επιδόσεών τους. Ο KNN, από την άλλη, ανταποκρίνεται χειρότερα σε σύγκριση με όλους τους άλλους αλγόριθμους και αποτελεί την λιγότερο κατάλληλη επιλογή για κλιμάκωση (με το χαμηλότερο ποσοστό στο 15%), αφού απαιτεί τη σύγκριση κάθε σημείου δεδομένων για κάθε πρόβλεψη, γεγονός που τον καθιστά ανίκανο να διαχειριστεί μεγάλα σύνολα δεδομένων.
4. ***Προσαρμοστικότητα***: Όπως μπορούμε να διακρίνουμε από το γράφημα μεταξύ των τριών ιδιαίτερα προσαρμόσιμων αλγορίθμων (Νευρωνικά Δίκτυα, Ενισχυτική Μάθηση και XGBoost), η Ενισχυτική Μάθηση (RL) ξεχωρίζει με το υψηλότερο ποσοστό που φτάνει το 95%. Η προσαρμοστικότητα της RL πηγάζει από το πλαίσιο συνεχούς μάθησης, το οποίο προσαρμόζεται σε μεταβαλλόμενα περιβάλλοντα και μπορεί να ενσωματώνει δυναμικά νέα δεδομένα χωρίς να χρειάζεται προκαθορισμένους κανόνες. Αυτό την καθιστά ιδιαίτερα κατάλληλη για περιβάλλοντα που απαιτούν λήψη αποφάσεων σε πραγματικό χρόνο και ταχείες προσαρμογές, καθώς η RL μαθαίνει από τις αλληλεπιδράσεις και προσαρμόζεται ανάλογα, ακόμη και όταν εξελίσσονται τα σενάρια απειλών. Από την άλλη πλευρά, οι KNN και HOG, είναι οι λιγότερο προσαρμόσιμοι αλγόριθμοι (ποσοστό 15%), καθώς είναι κατά κύριο λόγο κατάλληλοι για συγκεκριμένες, στατικές εργασίες, με περιορισμένη ικανότητα προσαρμογής σε νέες ή μεταβαλλόμενες απειλές.

Συνολική Εκτίμηση:

Με βάση την αξιολόγηση όλων των κριτηρίων, τα Νευρωνικά Δίκτυα και η Ενισχυτική Μάθηση καταγράφουν τα καλύτερα αποτελέσματα συνολικά, εξαιτίας της άριστης ακρίβειας που σημειώνουν. Οπότε, μπορεί να θεωρηθούν πως εντάσσονται και τα δύο στην πρώτη θέση (το καθένα "νικάει" σε ένα κριτήριο ενώ είναι ισόπαλα σε δύο). Παρόλα αυτά βάση προσωπικής κρίσης, τα Νευρωνικά Δίκτυα αναδεικνύονται ως ο καλύτερος αλγόριθμος για την ανίχνευση Email-based επιθέσεων επειδή θεωρούμε πως η ακρίβεια/αποτελεσματικότητα έχει υψηλότερη βαρύτητα σε σχέση με τα άλλα κριτήρια (όπως την προσαρμοστικότητα). Τα Νευρωνικά Δίκτυα επιτυγχάνουν εξαιρετικές επιδόσεις στην ακρίβεια, με το υψηλότερο ποσοστό 95%, καθιστώντας τα εξαιρετικά κατάλληλα για τον εντοπισμό σύνθετων μοτίβων phishing και άλλων εξελιγμένων επιθέσεων. Παρά το υψηλό υπολογιστικό κόστος τους - που απαιτεί σημαντική υποδομή - τα νευρωνικά δίκτυα εξακολουθούν να προτιμώνται επειδή η πολυεπίπεδη αρχιτεκτονική τους παρέχει μεγάλες δυνατότητες κλιμάκωσης. Αυτή η κλιμακωσιμότητα επιτρέπει στα Νευρωνικά Δίκτυα να χειρίζονται τεράστια σύνολα δεδομένων και περίπλοκα δεδομένα επιθέσεων χωρίς μείωση της απόδοσης, ένας κρίσιμος παράγοντας σε μεγάλης κλίμακας συστήματα ασφαλείας ηλεκτρονικού ταχυδρομείου. Η Ενισχυτική Μάθηση (RL) ακολουθεί ως ο δεύτερος καλύτερος αλγόριθμος (με βάση την προσωπική μας κρίση), με καλύτερη προσαρμοστικότητα όμως ελαφρώς χαμηλότερη ακρίβεια. Η προσαρμοστικότητα της RL είναι απαράμιλλη, καθώς μαθαίνει συνεχώς και προσαρμόζεται με βάση τις νέες αλληλεπιδράσεις, καθιστώντας την ιδανική για δυναμικά εξελισσόμενα τοπία απειλών, όπου οι προσαρμογές σε πραγματικό χρόνο είναι απαραίτητες. Αν και το υπολογιστικό της κόστος είναι υψηλό - όπως και των νευρωνικών δικτύων - αυτό αντισταθμίζεται από την παροχή γρήγορης προσαρμογής και αποτελεσματικού χειρισμού όλο και πιο πολύπλοκων απειλών και μεγάλων συνόλων δεδομένων. Οι στατικές μέθοδοι από την άλλη πλευρά, όπως HOG, DAISY και KNN, είναι καλύτερες για απλούστερες επιθέσεις phishing, με χαμηλότερο υπολογιστικό κόστος αλλά περιορισμένη κλιμάκωση και ακρίβεια σε σύνθετα δεδομένα. Η μέθοδος HOG κατατάσσεται στη χειρότερη θέση λόγω του κακού συνδυασμού περιορισμένης προσαρμοστικότητας, χαμηλής κλιμάκωσης και υψηλού ποσοστού σφάλματος.

4.2.3.2 Αποτελέσματα Σύγκρισης για Τεχνικές ΤΝ που εστιάζουν σε Social Media-Based Επιθέσεις

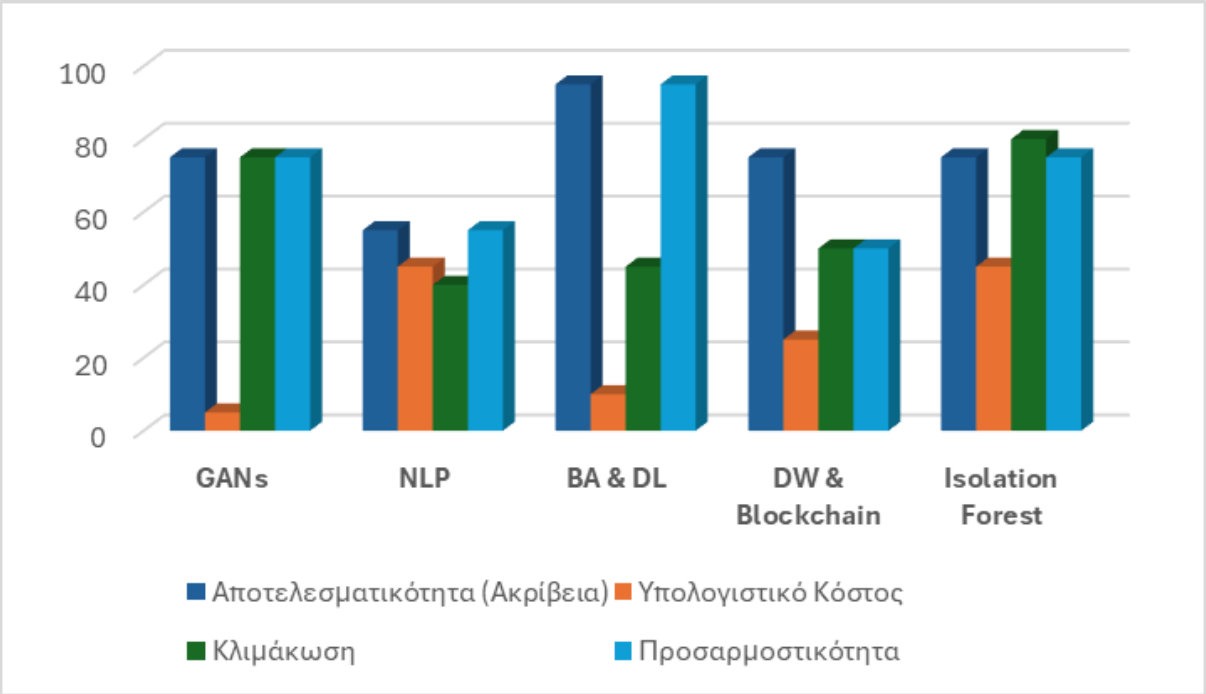
Ο παρακάτω πίνακας σύγκρισης παράχθηκε για τις τεχνικές ΤΝ με βάση το περιεχόμενο των άρθρων **[70], [71], [72], [73], [74], [75], [76]** και τα κριτήρια που έχουν τεθεί.

Πίνακας 3: Συγκριτική Ανάλυση Τεχνικών ΤΝ κατά των Social Media-Based Επιθέσεων ΚΜ.

Τεχνική	Τύπος	Αποτελεσματικότητα (Ακρίβεια)	Υπολογιστικό Κόστος	Κλιμάκωση	Προσαρμοστικότητα
GANs	Δυναμική	Υψηλή (Αποτελεσματική στον εντοπισμό πλαστών λογαριασμών και deepfakes	Πολύ Υψηλό (υψηλή επεξεργαστική ισχύ)	Υψηλή (καλή ταχύτητα επεξεργασίας σε μεγάλα σύνολα δεδομένων λόγω της πολυεπίπεδης δομής της, η πολυπλοκότητα μπορεί να προκαλέσει μικρές αποκλίσεις)	Υψηλή (αποτελεσματική στη δημιουργία προσαρμοστικών δεδομένων για την ανίχνευση συνθετικών ταυτοτήτων και ανωμαλιών)
NLP	Δυναμική	Μέτρια (Συχνές αποκλίσεις στα αποτελέσματα, δυσκολεύεται με την ακρίβεια σε πολύγλωσσα και δυναμικά γλωσσικά μοτίβα)	Μέτριο (λίγες ανάγκες σε πόρους)	Μέτρια (Κλιμακώνεται λογικά, αλλά αντιμετωπίζει μειωμένη ταχύτητα και ακρίβεια με εκτεταμένα ή πολύπλοκα δεδομένα ιδίως σε πολύγλωσσα περιβάλλοντα)	Μέτρια (Παρέχει μέτρια ευελιξία σε όλα τα γλωσσικά περιβάλλοντα, αλλά δυσκολεύεται να προσαρμοστεί σε ταχέως εξελισσόμενα γλωσσικά πρότυπα στα μέσα κοινωνικής δικτύωσης)
Biometric Authentication & Deep Learning (BA & DL)	Δυναμική	Πολύ Υψηλή (Πολύ λίγα έως καθόλου ψευδώς θετικά/αρνητικά, με αξιόπιστη επαλήθευση χρηστών σε πραγματικό χρόνο)	Πολύ Υψηλό (λόγω της εντατικής επεξεργασίας δεδομένων)	Μέτρια (Χειρίζεται σύνολα δεδομένων μεσαίου μεγέθους, αλλά απαιτεί προσαρμογές για πολύ μεγάλους όγκους δεδομένων)	Πολύ Υψηλή (ευπροσάρμοστη σε διάφορα περιβάλλοντα και μπορεί να ανταποκριθεί αποτελεσματικά σε νέους τύπους απειλών)

Digital Watermarking and Blockchain	Στατική	Υψηλή (Εξασφαλίζει την αυθεντικότητα των δεδομένων, αλλά είναι ευάλωτη στην παραποίηση, μειώνοντας ελαφρώς την ακρίβεια)	Υψηλό (σημαντικοί πόροι για την ενσωμάτωση και την επικύρωση υδατογραφήματος εντός blockchain)	Μέτρια (προκλήσεις απόδοσης σε σύνθετα σενάρια επιθέσεων)	Μέτρια (κατάλληλη για επαλήθευση δεδομένων σε ελεγχόμενα περιβάλλοντα, αλλά περιορισμένη σε πιο δυναμικές ή ταχέως μεταβαλλόμενες συνθήκες)
Isolation Forest	Δυναμική	Υψηλή (Αποδίδει καλά στον εντοπισμό ανωμαλιών σε μεγάλα σύνολα δεδομένων, με περιστασιακά ψευδώς θετικά αποτελέσματα)	Μέτριο (μέτριες απαιτήσεις επεξεργασίας δεδομένων)	Υψηλή (Διαχειρίζεται αποτελεσματικά μεγάλο όγκο δεδομένων)	Υψηλή (Προσαρμόζεται καλά σε μεταβαλλόμενα μοτίβα δεδομένων, ιδιαίτερα χρήσιμο για την ανίχνευση ανωμαλιών, αλλά ενδέχεται να χρειάζεται περιοδική προσαρμογή)

Γράφημα Ανάλυσης Άρθρων **[70], [71], [72], [73], [74], [75], [76].**
Μέθοδοι Κατά Των Social Media-Based Επιθέσεων:



Εικόνα 10: Οπτική αναπαράσταση των τεχνικών ΤΝ κατά των Social Media-Based Επιθέσεων ΚΜ.

Ακολουθεί επεξηγηματικό κείμενο σχολιασμού του γραφήματος, με τα αντίστοιχα συμπεράσματα:

1. **Αποτελεσματικότητα (Ακρίβεια):** Η Biometric Authentication & Deep Learning (BA & DL) ξεχωρίζει με την πολύ υψηλή ακρίβεια, καθώς καταγράφει ελάχιστα ψευδώς θετικά ή αρνητικά αποτελέσματα, ιδιαίτερα σε ταυτοποίηση σε πραγματικό χρόνο μέσω βιομετρικών δεδομένων. Η NLP παρουσιάζει τη χαμηλότερη ακρίβεια με αποκλίσεις στα αποτελέσματα, ειδικά σε πολύγλωσσα περιβάλλοντα ή όταν οι επιθέσεις χρησιμοποιούν μη προβλέψιμες γλωσσικές συμπεριφορές.
2. **Υπολογιστικό Κόστος**: Τα GANs και η Biometric Authentication & Deep Learning (BA & DL) έχουν το υψηλότερο υπολογιστικό κόστος, με τα GANs να υπολογίζονται δαπανέστερα κατά 5% απαιτώντας σημαντικούς πόρους. Αυτό οφείλεται κυρίως στο γεγονός ότι τα GANs απαιτούν εκτεταμένους κύκλους εκπαίδευσης που περιλαμβάνουν τόσο το δίκτυο γεννήτριας όσο και το δίκτυο διαχωρισμού, τα οποία χρειάζονται συνεχείς προσαρμογές για την παραγωγή ρεαλιστικών συνθετικών δεδομένων, αυξάνοντας σημαντικά τις απαιτήσεις επεξεργασίας. Αντιθέτως, η Isolation Forest και NLP έχουν το βέλτιστο υπολογιστικό κόστος, με μέτριες απαιτήσεις σε πόρους.
3. **Κλιμάκωση**: Ο Isolation Forest διαχειρίζεται αποτελεσματικά μεγάλα δεδομένα και σύνθετες επιθέσεις, επιδεικνύοντας υψηλή κλιμάκωση χωρίς σημαντική απώλεια απόδοσης, οπότε διακρίνεται με το υψηλότερο ποσοστό (80%). Ο NLP από την άλλη πλευρά, παρουσιάζει την χειρότερη κλιμάκωση, καθώς δυσκολεύεται να διαχειριστεί μεγάλα σύνολα δεδομένων και σύνθετες επιθέσεις, με επιδείνωση της απόδοσής του.
4. ***Προσαρμοστικότητα*:** Ο BA & DL αναδεικνύεται ως ο κορυφαίος αλγόριθμος ως προς την προσαρμοστικότητα (95%), επιτυγχάνοντας υψηλή ευελιξία σε ποικίλα περιβάλλοντα, ιδίως όσον αφορά τη πολύπλοκη επαλήθευση ταυτότητας σε πραγματικό χρόνο. Οι αλγόριθμοι Digital Watermarking & Blockchain και NLP παρουσιάζουν την χείριστη προσαρμοστικότητα, με τον πρώτο να υστερεί λίγο περισσότερο από τον δεύτερο. Αυτό συμβαίνει γιατί ο DW&B είναι ιδιαίτερα εξειδικευμένος για στατικές εργασίες επαλήθευσης δεδομένων, γεγονός που καθιστά δύσκολη την προσαρμογή του σε εξελισσόμενα τοπία απειλών που είναι τυπικά για τα μέσα κοινωνικής δικτύωσης. Σε γενικές γραμμές και οι δύο όμως είναι κατάλληλοι κυρίως για στατικά περιβάλλοντα δεδομένων ή λιγότερο δυναμικά σενάρια βασισμένα στη γλώσσα.

Συνολική Εκτίμηση:

Οι τεχνικές ΤΝ BA & DL και Isolation Forest θεωρούνται πως είναι οι καλύτερες, όπου η κάθε μία υπερτερεί σε διαφορετικά κριτήρια. Ειδικότερα, η τεχνική Isolation Forest επιτυγχάνει βέλτιστα αποτελέσματα και υπερτερεί στους τομείς της κλιμάκωσης και του υπολογιστικού κόστου ενώ η τεχνική BA & DL στην ακρίβεια και στην προσαρμοστικότητα. Αλλά και πάλι με

βάση την προσωπική μας κρίση, κρίνοντας πως η αποτελεσματικότητα είναι πιο σημαντική, θεωρούμε πως η τεχνική BA & DL είναι η καλύτερη. Ειδικότερα, η BA & DL έχει πολύ υψηλή αποδοτικότητα στην αποτελεσματικότητα, καταγράφοντας ελάχιστα ψευδώς θετικά και αρνητικά αποτελέσματα σε σενάρια ταυτοποίησης σε πραγματικό χρόνο, με εξαιρετική προσαρμοστικότητα σε διάφορα περιβάλλοντα, ιδίως σε δυναμικά περιβάλλοντα όπου η ταυτοποίηση σε πραγματικό χρόνο είναι ζωτικής σημασίας.

Παρόμοια, αν κοιτάξουμε τον πάτο της βαθμολογίας, θα παρατηρήσουμε πως σε αυτόν βρίσκονται οι τεχνικές NLP & DW-Blockchain. Η NLP είναι καλύτερη από την DW-Blockchain σε κόστος και προσαρμοστικότητα ενώ η DW-Blockchain είναι καλύτερη από την NLP ως προς την κλιμάκωση και την αποτελεσματικότητα. Οπότε, αν και πάλι θεωρήσουμε πως η αποτελεσματικότητα έχει περισσότερη βαρύτητα, τότε από τις δύο αυτές τεχνικές ξεχωρίζουμε την NLP ώς την χειρότερη. Επομένως, παρόλο που η NLP έχει βέλτιστο υπολογιστικό κόστος σε σύγκριση με άλλες τεχνικές της κατηγορίας, η μειωμένη χρησιμότητά της σε περιβάλλοντα μεγάλης κλίμακας, επηρεάζει σημαντικά την αποτελεσματικότητά της ως ολοκληρωμένης λύσης για την ανίχνευση επιθέσεων με βάση τα μέσα κοινωνικής δικτύωσης.

4.2.3.3 Αποτελέσματα Σύγκρισης για Τεχνικές ΤΝ που εστιάζουν σε Phone-Based Επιθέσεις

Ο παρακάτω πίνακας σύγκρισης παράχθηκε για τις τεχνικές ΤΝ με βάση το περιεχόμενο των άρθρων **[77] & [78],** και τα κριτήρια που έχουν τεθεί.

Πίνακας 4: Συγκριτική Ανάλυση Τεχνικών ΤΝ κατά των Phone-Based Επιθέσεων ΚΜ.

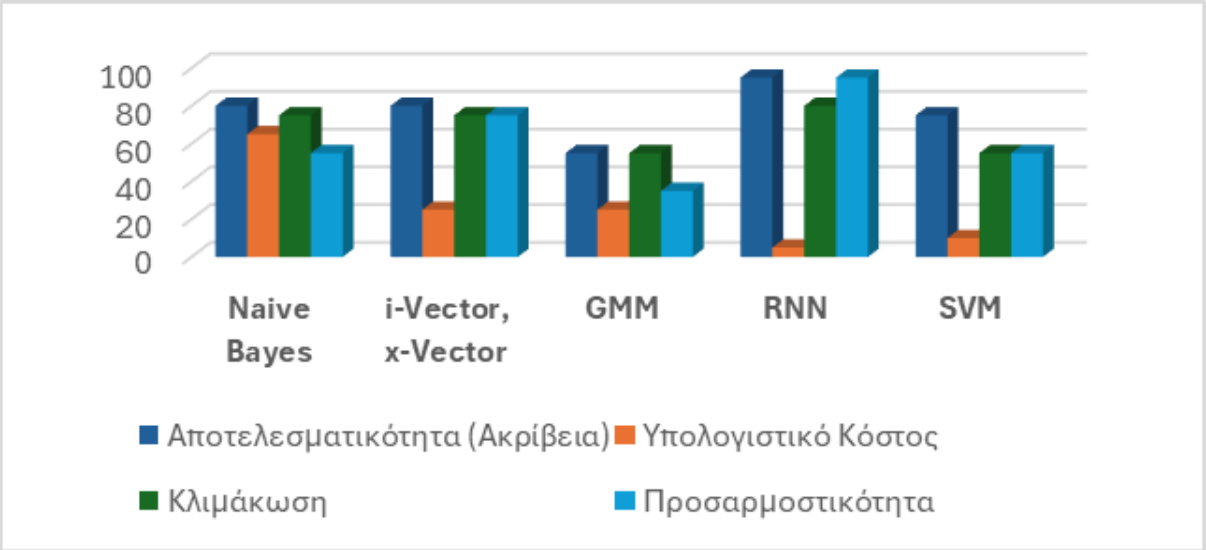
Τεχνική	Τύπος	Αποτελεσματικότητα (Ακρίβεια)	Υπολογιστικό Κόστος	Κλιμάκωση	Προσαρμοστικότητα
Naive Bayes	Στατική	Υψηλή (ανάλογα με τα δεδομένα με μικρές αποκλίσεις στα αποτελέσματα)	Χαμηλό (κατάλληλο για μεγάλους όγκους ανίχνευσης smishing)	Υψηλή (μεγάλα σύνολα δεδομένων, με σταθερή απόδοση)	Μέτρια (περιορισμένη προσαρμοστικότητα σε ποικίλα γλωσσικά πρότυπα)

i-Vector, x-Vector	Δυναμική	Υψηλή (Αξιόπιστη σε φωνητικά δεδομένα)	Υψηλό (επεξεργασία δεδομένων φωνής και βιομετρική ανάλυση με σημαντική σπατάλη πόρων)	Υψηλή (Σταθερή απόδοση σε αυξανόμενα φωνητικά βιομετρικά δεδομένα σε πραγματικό χρόνο)	Υψηλή (Επιδεικνύει ισχυρή προσαρμοστικότητα για βιομετρικά στοιχεία φωνής στην ανίχνευση vishing)
Gaussian Mixture Models (GMM)	Δυναμική	Μέτρια (επιρρεπής σε μειωμένη ακρίβεια με χαμηλής ποιότητας ήχο)	Υψηλό (λεπτομερής ανάλυση των χαρακτηριστικών του ήχου που συνεπάγεται υπολογιστική ισχύ)	Μέτρια	Χαμηλή (Περιορίζεται σε συγκεκριμένα περιβάλλοντα)
Recurrent Neural Networks (RNN)	Δυναμική	Πολύ Υψηλή (Εξαιρετικά ακριβή στην ανίχνευση σύνθετων μοτίβων ομιλίας και στον εντοπισμό διαδοχικών ανωμαλιών, κάτι που τη καθιστά ιδανικό για την ανίχνευση vishing και robocall επιθέσεων)	Πολύ Υψηλό (ιδίως για την ανάλυση διαδοχικών δεδομένων κατά την ανίχνευση μοτίβων που βασίζονται στη φωνή)	Υψηλή (Κλιμακώνεται καλά με μεγάλα σύνολα δεδομένων και σύνθετα μοτίβα δεδομένων)	Πολύ υψηλή (υπερέχει στην προσαρμογή σε διαφορετικά μοτίβα επίθεσης)

Support Vector Machine (SVM)	Δυναμική	Υψηλή (ακριβής στην ανίχνευση σύνθετων μοτίβων, αποτελεσματική για smishing)	Πολύ Υψηλό (απαιτεί εκτενή επεξεργασία)	Μέτρια (αποτελεσματική για δομημένα δεδομένα)	Μέτρια (απαιτεί προσαρμογές για πολύπλοκα, ποικίλα σύνολα δεδομένων)

Γράφημα Ανάλυσης Άρθρων **[77], [78].**

Μέθοδοι Κατά Των Phone-Based Επιθέσεων:



Εικόνα 11: Οπτική αναπαράσταση των τεχνικών ΤΝ κατά των Phone-Based Επιθέσεων ΚΜ.

Ακολουθεί επεξηγηματικό κείμενο σχολιασμού του γραφήματος, με τα αντίστοιχα συμπεράσματα:

1. **Αποτελεσματικότητα (Ακρίβεια):** Τα Recurrent Neural Networks (RNN) κατατάσσονται ως η καλύτερη τεχνική ΤΝ ως προς την ακρίβεια στην ανίχνευση επιθέσεων με βάση το τηλέφωνο, λόγω της ισχυρής ικανότητάς τους να ανιχνεύουν διαδοχικά μοτίβα στα δεδομένα φωνής, που είναι ζωτικής σημασίας για τον εντοπισμό κοινωνικής μηχανικής σε πραγματικό χρόνο. Επίσης, είναι εξαιρετικά αποδοτικά στην αναγνώριση πλαστογραφημένων κλήσεων ή robocalls. Η τεχνική Gaussian Mixture Models (GMM) παρουσιάζει τη χαμηλότερη ακρίβεια, η οποία ποικίλλει ανάλογα με τη

διάρκεια και την ποιότητα του ήχου, καθιστώντας την απόδοσή της μέτρια σε πιο σύνθετες καταστάσεις.

2. **Υπολογιστικό Κόστος**: Η τεχνική Naive Bayes για την ανίχνευση smishing έχει, όπως παρατηρείται, το χαμηλότερο υπολογιστικό κόστος, καθιστώντας την ιδανική για περιβάλλοντα με μεγάλα δεδομένα και περιορισμένους πόρους. Από την άλλη μεριά, τα Recurrent Neural Networks (RNN) παρουσιάζουν το υψηλότερο υπολογιστικό κόστος, καθώς επεξεργάζονται διαδοχικά δεδομένα, τα οποία απαιτούν τη διατήρηση του πλαισίου με την πάροδο του χρόνου, συχνά μέσω πολλαπλών επιπέδων και επαναλαμβανόμενων βρόχων. Κάθε χρονικό βήμα σε ένα RNN απαιτεί υπολογισμούς που εξαρτώνται από το προηγούμενο βήμα, οδηγώντας σε αυξημένες απαιτήσεις επεξεργασίας, ειδικά όταν αναλύονται σύνθετα, χρονικά πρότυπα, όπως τα δεδομένα φωνής. Αμέσως μετά, με διαφορά 5% έρχονται τα SVM, καθώς απαιτούν περισσότερους υπολογισμούς λόγω της διαδικασίας εκπαίδευσης και της χρήσης του kernel για τις προβλέψεις.
3. **Κλιμάκωση**: Τα RNNs έχουν την υψηλότερη κλιμάκωση, καθώς διαχειρίζονται αποτελεσματικά μεγάλα δεδομένα και παρουσιάζουν σταθερή απόδοση με την αύξηση του όγκου δεδομένων. Τα RNN έχουν σχεδιαστεί για να χειρίζονται δεδομένα που εξαρτώνται από το χρόνο, επιτρέποντάς τους να διατηρούν σταθερή απόδοση ακόμη και όταν το σύνολο δεδομένων αυξάνεται σε μέγεθος και πολυπλοκότητα, καθιστώντας τα εξαιρετικά επεκτάσιμα για δυναμικές εφαρμογές σε πραγματικό χρόνο, όπου η γρήγορη προσαρμογή και η συνεχής επεξεργασία δεδομένων είναι απαραίτητες. Από την άλλη, οι τεχνικές GMM και SVM έχουν την χειρότερη κλιμάκωση, παρουσιάζοντας μειωμένη απόδοση καθώς αυξάνεται ο όγκος των δεδομένων ή μειώνεται η ποιότητα του ήχου.
4. ***Προσαρμοστικότητα*:** Τα Recurrent Neural Networks (RNN) κατατάσσονται υψηλότερα στην προσαρμοστικότητα, καθώς η δομή τους επιτρέπει την επεξεργασία χρονικών ακολουθιών και τη διατήρηση ιστορικού πληροφορίας, καθιστώντας τα ιδανικά για ποικίλα και μεταβαλλόμενα τοπία απειλών. Η τεχνική Gaussian Mixture Models (GMM) κατατάσσεται αρκετά χαμηλότερα, καθώς βασίζεται σε στατιστικές κατανομές που είναι αποτελεσματικές σε σταθερά ή ελεγχόμενα περιβάλλοντα αλλά δυσκολεύονται να προσαρμοστούν σε ποικιλία δεδομένων ή ταχέως μεταβαλλόμενες συνθήκες.

Συνολική Εκτίμηση:

Τα Recurrent Neural Networks (RNN) παρουσιάζουν την καλύτερη συνολική απόδοση λόγω της υψηλής ακρίβειας σε μεγάλα δεδομένα και της ισχυρής προσαρμοστικότητάς τους, αν και έχουν υψηλό υπολογιστικό κόστος. Η τεχνική Gaussian Mixture Models (GMM) κατατάσσεται ως η λιγότερο αποδοτική τεχνική ΤΝ, λόγω της μέτριας αποτελεσματικότητας και κλιμάκωσης καθώς και της μειωμένης προσαρμοστικότητας και υψηλού υπολογιστικού κόστους.

4.2.3.4 Αποτελέσματα Σύγκρισης για Τεχνικές ΤΝ που εστιάζουν σε In-Person Επιθέσεις

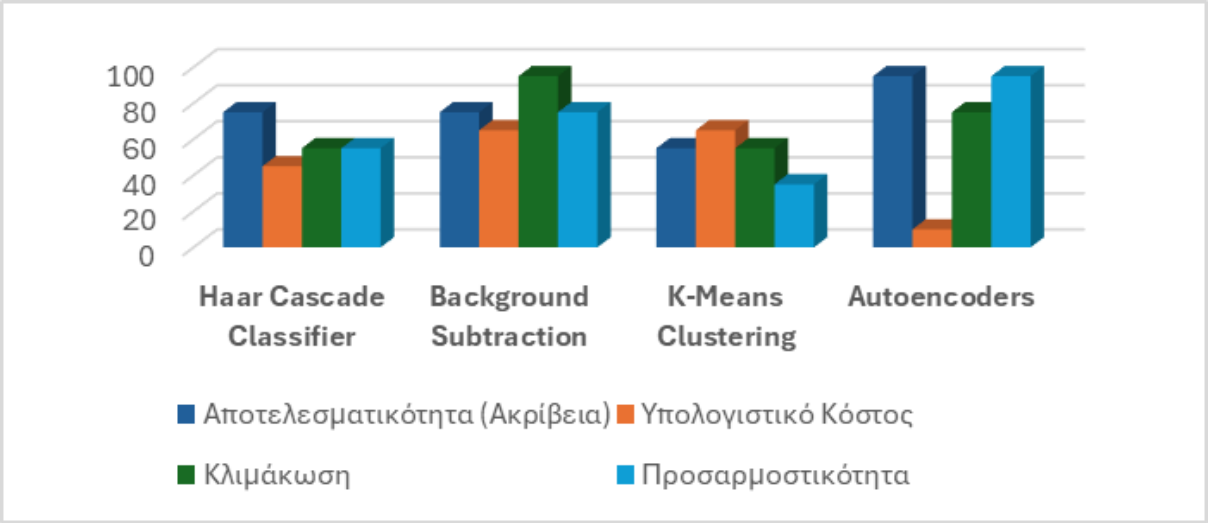
Ο παρακάτω πίνακας σύγκρισης παράχθηκε για τις τεχνικές ΤΝ με βάση το περιεχόμενο των άρθρων **[81] & [82],** και τα κριτήρια που έχουν τεθεί.

Πίνακας 5: Συγκριτική Ανάλυση Τεχνικών ΤΝ κατά των In-Person Επιθέσεων ΚΜ.

Τεχνική	Τύπος	Αποτελεσματικότητα (Ακρίβεια)	Υπολογιστικό Κόστος	Κλιμάκωση	Προσαρμοστικότητα
Haar Cascade Classifier	Στατική	Υψηλή (στην ανίχνευση προσώπων, μερικά ψευδώς θετικά σε κακές συνθήκες φωτισμού)	Μέτριο (Λογική κατανάλωση πόρων)	Μέτρια (Καλή απόδοση, αλλά επηρεάζεται από συνθήκες φωτισμού)	Μέτρια (περιορισμένη απόκριση σε μεταβαλλόμενες φυσικές συνθήκες)
Background Subtraction	Δυναμική	Υψηλή (Εξαιρετική στην ανίχνευση κίνησης)	Χαμηλό (Ελάχιστοι υπολογιστικοί πόροι)	Πολύ υψηλή (Ικανότητα ισχυρής ανίχνευσης σε πραγματικό χρόνο)	Υψηλή ειδικά για τις περιβαλλοντικές αλλαγές
K-Means Clustering	Στατική	Μέτρια (Εξαρτάται από την πολυπλοκότητα των δεδομένων και τις αρχικές συνθήκες)	Χαμηλό (Λογική κατανάλωση πόρων, κατάλληλη για απλές εργασίες ομαδοποίησης)	Μέτρια (Προσαρμόζεται καλά σε μικρότερα περιβάλλοντα, αλλά έχει περιορισμούς σε μεγάλα δεδομένα)	Χαμηλή (αποδίδει καλύτερα σε απλά περιβάλλοντα και απαιτεί αναδιαμόρφωση σε πολύπλοκα σενάρια)

Autoencoders	Δυναμική	Πολύ Υψηλή (Ακριβής ανίχνευση ανωμαλιών και σπάνιων συμπεριφορών)	Πολύ Υψηλό, απαιτεί υψηλή υπολογιστική ισχύ και μεγάλα σύνολα δεδομένων	Υψηλή (Αποδίδει πολύ καλά σε περιβάλλοντα με μεγάλα και περίπλοκα δεδομένα)	Πολύ υψηλή (εξαιρετική προσαρμοστικότητα σε δυναμικά περιβάλλοντα)

Γράφημα Ανάλυσης Άρθρων **[81], [82].**
Μέθοδοι Κατά Των In-Person Επιθέσεων:



Εικόνα 12: Οπτική αναπαράσταση των τεχνικών ΤΝ κατά των In-Person Επιθέσεων ΚΜ.

Ακολουθεί επεξηγηματικό κείμενο σχολιασμού του γραφήματος, με τα αντίστοιχα συμπεράσματα:

1. ***Αποτελεσματικότητα (Ακρίβεια):*** Ο Autoencoders είναι η καλύτερη τεχνική ΤΝ, καθώς προσφέρει πολύ υψηλή ακρίβεια στην ανίχνευση ανωμαλιών και σπάνιων συμπεριφορών, ιδανική για την αντιμετώπιση σύνθετων επιθέσεων, όπως το pretexting. Αντίθετα, η τεχνική K-Means Clustering είναι η λιγότερο αποτελεσματική, καθώς η ακρίβειά της εξαρτάται σε μεγάλο βαθμό από την πολυπλοκότητα των δεδομένων και τις αρχικές συνθήκες, κάτι που μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα.
2. ***Υπολογιστικό Κόστος:*** Οι τεχνικές Background Subtraction Algorithm και K-Means έχουν το χαμηλότερο κόστος, λόγω της απλότητας των υπολογισμών τους, που βασίζονται σε στατικές διαφορές ή επαναληπτικές διαδικασίες αντίστοιχα. Αντίθετα, η

τεχνική Autoencoders έχει το υψηλότερο κόστος, καθώς απαιτεί μεγάλη υπολογιστική ισχύ και μεγάλα σύνολα δεδομένων για να είναι αποδοτική.

3. ***Κλιμάκωση:*** Η τεχνική Background Subtraction φαίνεται να ξεχωρίζει ως η καλύτερη, χειρίζοντας αποτελεσματικά μεγάλους όγκους δεδομένων και πολύπλοκα σενάρια χωρίς προβλήματα απόδοσης. Οι τεχνικές Haar Cascade Classifier και K-Means Clustering έχουν τη χειρότερη κλιμάκωση, καθώς η απόδοσή τους επηρεάζεται από συνθήκες, όπως ο φωτισμός και οι επιθέσεις που χρησιμοποιούν καλύμματα προσώπου, γεγονός που μειώνει την απόδοσή τους σε πιο απαιτητικά περιβάλλοντα.
4. ***Προσαρμοστικότητα:*** Η τεχνική Autoencoders κατέχει την υψηλότερη θέση στην προσαρμοστικότητα, υπερέχοντας σε δυναμικά και πολύπλοκα περιβάλλοντα όπου απαιτείται ευέλικτη ανίχνευση ανωμαλιών. Αντίθετα, η τεχνική K-Means Clustering έχει τη χαμηλότερη προσαρμοστικότητα, περιορίζεται σε πιο στατικά περιβάλλοντα και δυσκολεύεται με πολύπλοκα ή εξελισσόμενα δεδομένα.

Συνολική Εκτίμηση: Οι δυναμικές τεχνικές, όπως Autoencoders, είναι οι καλύτερες συνολικά (βάση και πάλι την προσωπική μας κρίση ως προς την κρισιμότητα των κριτηρίων), υπερέχοντας στην ακρίβεια, την κλιμακωσιμότητα και την προσαρμοστικότητα. Ανιχνεύουν αποτελεσματικά πολύπλοκα μοτίβα και ανωμαλίες με πολύ υψηλή ακρίβεια (95%) και διατηρούν την απόδοση ακόμη και όταν αυξάνεται ο όγκος δεδομένων και η περιβαλλοντική πολυπλοκότητα, αν και έχουν υψηλό υπολογιστικό κόστος. Παράλληλα, ο Background Subtraction θεωρούμε πως έχει υψηλή απόδοση κατά μήκος των κριτηρίων, πράγμα που τον κατατάσσει σχεδόν στο ίδιο επίπεδο με την τεχνική Autoencoders (αν δεν ληφθεί υπόψιν κάποια διαφοροποίηση ως προς τη βαρύτητα των κριτηρίων). Ειδικότερα, παρουσιάζει υψηλή αποτελεσματικότητα που σε συνδυασμό με την άριστη κλιμάκωση και το πολύ καλό υπολογιστικό κόστος θα μπορούσε να θεωρηθεί πως έρχεται σε ισοπαλία με τους Autoencoders (οι οποίοι υπερέχουν σε ακρίβεια και προσαρμοστικότητα). Οι στατικές τεχνικές, όπως η K-Means Clustering, είναι πιο αποδοτικές σε απλούστερα περιβάλλοντα λόγω χαμηλού υπολογιστικού κόστους, αλλά δεν αποδίδουν καλά σε πολύπλοκες καταστάσεις. Συνολικά, οι δυναμικές μέθοδοι είναι καταλληλότερες για σύνθετες επιθέσεις, ενώ οι στατικές μέθοδοι είναι πιο αποδοτικές σε περιβάλλοντα με περιορισμένους πόρους.

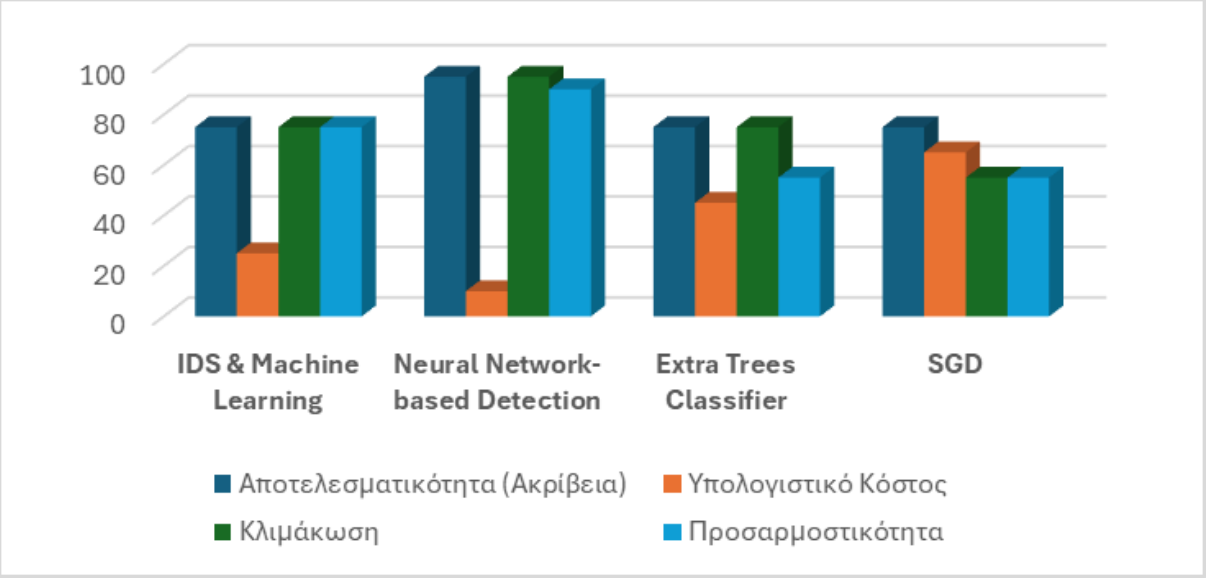
4.2.3.5 Αποτελέσματα Σύγκρισης για Τεχνικές ΤΝ που εστιάζουν σε Web-Based Επιθέσεις

Ο παρακάτω πίνακας σύγκρισης παράχθηκε για τις τεχνικές ΤΝ με βάση το περιεχόμενο των άρθρων **[83] & [84],** και τα κριτήρια που έχουν τεθεί.

Πίνακας 6: Συγκριτική Ανάλυση Τεχνικών ΤΝ κατά των Web-Based Επιθέσεων ΚΜ.

Τεχνική	Τύπος	Αποτελεσματικότητα (Ακρίβεια)	Υπολογιστικό Κόστος	Κλιμάκωση	Προσαρμοστικότητα
Intrusion Detection System (IDS) με Machine Learning	Δυναμική	Υψηλή (Ανίχνευση ανωμαλιών, αλλά μπορεί να έχει ψευδώς θετικά)	Υψηλό (Απαιτεί συνεχή παρακολούθηση και βελτιστοποίηση)	Υψηλή (Ανιχνεύει σε πραγματικό χρόνο μεγάλους όγκους κυκλοφορίας)	Υψηλή (Ικανή να προσαρμόζεται αποτελεσματικά, αλλά απαιτεί συνεχή συντονισμό για πολύπλοκες απειλές)
Neural Network-based Detection	Δυναμική	Πολύ Υψηλή (Ακριβής ανίχνευση λεπτών αποκλίσεων, ελάχιστα ψευδώς θετικά)	Πολύ Υψηλό (Απαιτεί εκτεταμένους υπολογιστικούς πόρους για εκπαίδευση)	Πολύ Υψηλή (Εξαιρετική προσαρμογή σε μεγάλες και περίπλοκες επιθέσεις)	Πολύ υψηλή (προσαρμόζεται καλά σε διάφορες απειλές και δυναμικά περιβάλλοντα αξιοποιώντας αρχιτεκτονικές βαθιάς μάθησης)
Extra Trees Classifier	Στατική	Υψηλή (Καλή απόδοση σε μεγάλους όγκους δεδομένων, με λιγότερο overfitting)	Μέτριο	Υψηλή (Αντιμετωπίζει μεγάλους όγκους δεδομένων χωρίς μεγάλη επιβάρυνση)	Μέτρια (λειτουργεί επαρκώς σε ορισμένα περιβάλλοντα και ταιριάζει καλύτερα σε σταθερές ρυθμίσεις)
Stochastic Gradient Descent (SGD)	Δυναμική	Υψηλή (Ακρίβεια σε μεγάλα πολυδιάστατα δεδομένα)	Χαμηλό (Αποδοτική σε πραγματικό χρόνο με σχετικά μικρό κόστος)	Μέτρια (αντιμετωπίζει μειωμένη απόδοση σε πολύ μεγάλα ή σύνθετα δεδομένα)	Μέτρια (για ποικίλες απειλές που βασίζονται στο διαδίκτυο με περιοδικό συντονισμό, καθιστώντας την αποτελεσματική για σταθερές ανιχνεύσεις σε πραγματικό χρόνο)

Γράφημα Ανάλυσης Άρθρων **[83], [84].**
Μέθοδοι Κατά Των Web-based Επιθέσεων:



Εικόνα 13: Οπτική αναπαράσταση των τεχνικών ΤΝ κατά των Web-Based Επιθέσεων ΚΜ.

Ακολουθεί επεξηγηματικό κείμενο σχολιασμού του γραφήματος, με τα αντίστοιχα συμπεράσματα:

1. ***Αποτελεσματικότητα (Ακρίβεια):*** Ο αλγόριθμος που ξεχωρίζει είναι ο Neural Network-based Detection, καθώς προσφέρει πολύ υψηλή ακρίβεια, ανιχνεύοντας λεπτές αποκλίσεις στη συμπεριφορά του δικτύου και των ιστοτόπων, με ελάχιστα ψευδώς θετικά αποτελέσματα. Αυτό τον καθιστά ιδανικό για την ανίχνευση Watering Hole επιθέσεων. Αντίθετα, οι άλλοι 3 αλγόριθμοι βρίσκονται στην ίδια κλίμακα, σε ένα ικανοποιητικό, όμως χαμηλότερο επίπεδο από τον NN. Έτσι ο IDS με Machine Learning έχει ακρίβεια 75% λόγω της τάσης του να παράγει σχετικά περισσότερα ψευδώς θετικά αποτελέσματα, απαιτώντας συνεχή βελτιστοποίηση και παρακολούθηση. Ενώ οι SGD & Extra Trees Classifier αποδίδουν επίσης καλά (75%), αλλά έχουν εξίσου ευαισθησία σε ψευδώς θετικά αποτελέσματα.
2. ***Υπολογιστικό Κόστος:*** Ο Stochastic Gradient Descent (SGD) είναι ο πιο αποδοτικός αλγόριθμος ΤΝ, καθώς μπορεί να διαχειριστεί μεγάλα σύνολα δεδομένων με χαμηλό κόστος και είναι ιδιαίτερα γρήγορος στην ανίχνευση κακόβουλων διαφημίσεων. Ο αλγόριθμος Neural Network-based Detection, από την άλλη, έχει το υψηλότερο υπολογιστικό κόστος, καθώς απαιτεί σημαντικούς πόρους για εκπαίδευση και ανάλυση μεγάλων δεδομένων, απαιτώντας ισχυρή υποδομή και καθιστώντας τον τελικά λιγότερο αποδοτικό σε περιβάλλοντα με περιορισμένους πόρους.
3. ***Κλιμάκωση:*** Ο Neural Network-based Detection είναι ο καλύτερος αλγόριθμος, καθώς επεξεργάζεται αποτελεσματικά μεγάλα σύνολα δεδομένων και πολύπλοκες απειλές χωρίς συμβιβασμούς στην απόδοση. Είναι ιδιαίτερα αποτελεσματικός σε

σύνθετα περιβάλλοντα, όπως αυτά που απαιτούν ανίχνευση Watering Hole επιθέσεων. Ο Stochastic Gradient Descent (SGD), ενώ είναι αποτελεσματικός για μέτριους όγκους δεδομένων, παρουσιάζει μειωμένη επεκτασιμότητα με μεγαλύτερα σύνολα δεδομένων, καθιστώντας τον λιγότερο κατάλληλο για εκτεταμένη παρακολούθηση απειλών μέσω διαδικτύου.

4. ***Προσαρμοστικότητα:*** Ο Neural Network-based Detection κατέχει την υψηλότερη θέση στην προσαρμοστικότητα, χειρίζοντας ποικίλα και εξελισσόμενα τοπία απειλών λόγω της εξελιγμένης αρχιτεκτονικής μάθησης. Αντίθετα, ο Extra Trees Classifier και ο SGD παρουσιάζουν μέτρια προσαρμοστικότητα, αποδίδοντας καλύτερα σε πιο στατικά ή δυναμικά ελεγχόμενα περιβάλλοντα, όπου οι προσαρμογές είναι λιγότερο συχνές.

Συνολική Εκτίμηση: Ο Neural Network-based Detection είναι ο πιο αποτελεσματικός αλγόριθμος λόγω της υψηλής ακρίβειας, κλιμάκωσης και προσαρμοστικότητάς του, ιδανικός για σύνθετες επιθέσεις, όπως η Watering Hole, αν και έχει υψηλό υπολογιστικό κόστος. Ο IDS αποδίδει καλά, αλλά απαιτεί συνεχή βελτιστοποίηση λόγω ψευδώς θετικών αποτελεσμάτων. Ο Extra Trees Classifier είναι αποδοτικός σε μεγάλα δεδομένα με μέτριο κόστος, ενώ ο SGD προσφέρει χαμηλό κόστος, αλλά απαιτεί προσεκτική παραμετροποίηση εξαιτίας περιορισμένης επεκτασιμότητας και προσαρμοστικότητας. Οι δύο αυτοί αλγόριθμοι θεωρούνται χειρότεροι σε σχέση με τους άλλους 2, με τον SGD να έχει ελαφρώς μικρότερη συνολική απόδοση από τον Extra Trees Classifier και άρα να θεωρείται ο χείριστος αλγόριθμος ΤΝ. Εν ολίγοις, οι αλγόριθμοι με υψηλή ακρίβεια και κλιμάκωση (δυναμικοί), όπως ο Neural Network-based Detection, είναι οι πιο κατάλληλοι για σύνθετες επιθέσεις, ενώ εκείνοι με χαμηλότερο κόστος (στατικοί), όπως ο SGD, προσφέρουν καλή απόδοση για πολύ απλούστερα σενάρια.

4.2.3.6 Καθολικά Αποτελέσματα Ανάλυσης Τεχνικών ΤΝ

Συνοψίζοντας, οι αλγόριθμοι με την καλύτερη απόδοση συνολικά είναι τα Νευρωνικά Δίκτυα (ΝΝ) και τα Επαναληπτικά Νευρωνικά Δίκτυα (RNN), καθώς επιτυγχάνουν υψηλά επίπεδα ακρίβειας, κλιμάκωσης και προσαρμοστικότητας σε σύνθετα περιβάλλοντα επιθέσεων. Ωστόσο, είναι σημαντικό να διευκρινιστεί ότι τα Νευρωνικά Δίκτυα αναλύονται σε διαφορετικές κατηγορίες, με την απόδοσή τους να εξαρτάται από τη συγκεκριμένη εφαρμογή και το είδος των δεδομένων.

Τα ΝΝ, ειδικά τα βαθιά νευρωνικά δίκτυα, έχουν αποδείξει υψηλή ακρίβεια στην ανίχνευση επιθέσεων μέσω email, όπως το phishing, επιτυγχάνοντας ποσοστά ακρίβειας έως και 95%. Αυτό τα καθιστά ιδιαίτερα αποτελεσματικά στον εντοπισμό σύνθετων μοτίβων σε μεγάλα σύνολα δεδομένων. Ωστόσο, απαιτούν σημαντικούς υπολογιστικούς πόρους και ενδέχεται να παρουσιάζουν περιορισμένη προσαρμοστικότητα σε νέες ή εξελισσόμενες απειλές.

Από την άλλη πλευρά, τα RNN είναι σχεδιασμένα για την επεξεργασία διαδοχικών δεδομένων, καθιστώντας τα κατάλληλα για την ανάλυση κειμένου και την επεξεργασία φυσικής γλώσσας. Η ικανότητά τους να διατηρούν πληροφορίες από προηγούμενα βήματα τα καθιστά αποτελεσματικά στην κατανόηση του συμφραζομένου σε ακολουθιακά δεδομένα. Ωστόσο, η εκπαίδευσή τους μπορεί να είναι πιο απαιτητική υπολογιστικά και να παρουσιάζει προκλήσεις.

Συνολικά, η επιλογή μεταξύ NN και RNN εξαρτάται από τη φύση των δεδομένων και τις απαιτήσεις της εφαρμογής. Για την ανίχνευση επιθέσεων μέσω email, όπου η ακρίβεια είναι κρίσιμη, τα NN μπορεί να είναι προτιμητέα. Ωστόσο, για εφαρμογές που απαιτούν κατανόηση της ακολουθίας και του συμφραζομένου, τα RNN προσφέρουν πλεονεκτήματα. Επίσης, για προβλήματα με δομημένα δεδομένα και περιορισμένους πόρους, ο XGBoost μπορεί να είναι πιο αποδοτικός και από τους δυο προαναφερόμενους αλγορίθμους ΤΝ, καθώς είναι γνωστός για την εξαιρετική του απόδοση σε προβλήματα ταξινόμησης - παλινδρόμησης καθώς και είναι ανθεκτικός σε υπερπροσαρμογή. Επίσης, η σύγκριση ανέδειξε πως ο XGBoost υπερέχει των RNN σε κλιμάκωση, ενώ τα RNN νικάνε σε μικρό ποσοστό στην αποτελεσματικότητα. Αντιθέτως, ο χειρότερος αλγόριθμος συνολικά είναι ο Gaussian Mixture Models (GMM), καθώς παρουσιάζει χαμηλή προσαρμοστικότητα και περιορισμένη αποτελεσματικότητα σε δυναμικά ή σύνθετα περιβάλλοντα. Αυτό τον καθιστά κατάλληλο μόνο για συγκεκριμένες, σταθερές εφαρμογές.

Παρόλο που η σύγκριση αναδεικνύει τους καλύτερους και χειρότερους αλγορίθμους, είναι σημαντικό να σημειωθεί ότι κάθε στρατηγική έχει μοναδική δυναμική και μπορεί να υπερέχει σε συγκεκριμένες καταστάσεις. Η αξιολόγηση βοηθά στη σωστή επιλογή ανάλογα με τις απαιτήσεις του περιβάλλοντος εφαρμογής (π.χ., μεγάλα δεδομένα, υψηλή ακρίβεια, χαμηλό κόστος). Τέλος, οι συγκρίσεις αποκαλύπτουν αδυναμίες και περιθώρια βελτίωσης για μελλοντική έρευνα, ειδικά για λύσεις που στοχεύουν σε μεγαλύτερη ευελιξία και αποδοτικότητα σε ένα ευρύ φάσμα επιθέσεων.

4.3 Ανάλυση των Αντίμετρων

Για την υπεράσπιση των συστημάτων από επιθέσεις, τα αντίμετρα παρουσιάζονται ως ολοκληρωμένες τακτικές ή διαδικασίες, που μπορούν να χρησιμοποιούν αλγόριθμους ΤΝ. Επομένως, οι αλγόριθμοι ΤΝ αποτελούν τα επιμέρους εργαλεία που χρησιμοποιούνται για την υλοποίηση των αντίμετρων, καθοδηγώντας τη λήψη αποφάσεων βάσει κατάλληλης ανάλυσης δεδομένων. Τονίζεται, όμως, πως τα αντίμετρα διαφέρουν ως προς τον τρόπο με τον οποίο οι αλγόριθμοι ΤΝ χρησιμοποιούνται και εφαρμόζονται στο πλαίσιο ευρύτερων σχεδίων ασφαλείας.

Στην ενότητα αυτή, η ανάλυση θα επικεντρωθεί στη σύνδεση αυτών των δύο εννοιών που προαναφέρθηκαν και ειδικότερα στη διαφοροποίηση των αντιμέτρων ως προς τον τρόπο χρήσης των αλγορίθμων ΤΝ που ενσωματώνουν. Επίσης, τα αντίμετρα κατά των επιθέσεων κοινωνικής μηχανικής κατηγοριοποιούνται με βάση τη διάσταση του τύπου αντιμετώπισης σε τρεις κύριες προσεγγίσεις: Προληπτικά Αντίμετρα, Αντίμετρα για Ανίχνευση σε Πραγματικό Χρόνο, και Αντιδραστικά Αντίμετρα. Συνεπώς, η διεξοδική ανάλυση των αντιμέτρων με βάση τη διάσταση αυτή διευκολύνει την έρευνα των καλύτερων προσεγγίσεων για κάθε τύπο αντιμετώπισης. Ως αποτέλεσμα, προσφέρει ένα σαφές πλαίσιο και καθιστά δυνατή την αποτελεσματικότερη στόχευση των αντιμέτρων με βάση τις απαιτήσεις για τον κάθε τύπο αντιμετώπισης των επιθέσεων. Επίσης, αναδεικνύει τη συμπληρωματικότητά τους, παρέχοντας καθοδήγηση στον τρόπο επιλογής και σύνθεσής τους.

Ως προς την δομή της παρισταμένης ενότητας, έπειτα από την σύγκριση των τεχνικών / αλγορίθμων ΤΝ που πραγματοποιήθηκε, θα παρουσιαστούν τα ολοκληρωμένα αντίμετρα που τα πλαισιώνουν. Πιο συγκεκριμένα, η δομή της ενότητας 4.3 βασίζεται σε τρεις κύριες υπο-ενότητες. Αρχικά, η υπο-ενότητα 4.3.1 παρουσιάζει τα αντίμετρα που αφορούν την πρόληψη των επιθέσεων ΚΜ, με τους αντίστοιχους αλγόριθμους και τεχνικές που τα πλαισιώνουν. Στη συνέχεια, η υπο-ενότητα 4.3.2 αναλύει τα αντίμετρα κατά των επιθέσεων σε πραγματικό χρόνο. Τέλος, η υπο-ενότητα 4.3.3 περιλαμβάνει την ανάλυση των αντίμετρων που εστιάζουν στην ανάκαμψη από τις επιθέσεις αυτές, εφόσον αυτές έχουν συμβεί.

4.3.1 Προληπτικά Αντίμετρα

Στόχος των αντιμέτρων αυτών είναι η ανίχνευση και αποτροπή επιθέσεων πριν αυτές συμβούν, μέσω ανάλυσης συμπεριφοράς και μοτίβων. Περιλαμβάνουν αλγόριθμους ΤΝ που είναι σχεδιασμένοι να εντοπίζουν ύποπτες δραστηριότητες προτού αυτές εξελιχθούν σε επιθέσεις.

- Οι αλγόριθμοι ΤΝ που ενσωματώνονται είναι οι εξής:
    - **Naive Bayes**: Ο συγκεκριμένος αλγόριθμος αποδίδει εξαιρετικά στην ανάλυση κειμένου, καθιστώντας τον κατάλληλο για την ανίχνευση επιθέσεων phishing ή smishing μέσω της πρόβλεψης μοτίβων στο περιεχόμενο μηνυμάτων. Η προληπτική του φύση έγκειται στο ότι αναλύει το κείμενο πριν ολοκληρωθεί η επίθεση, εντοπίζοντας πιθανά κακόβουλα στοιχεία.
    - **K-Means Clustering**: Αυτός ο αλγόριθμος είναι εξειδικευμένος στην ανίχνευση ανωμαλιών σε σύνολα δεδομένων, κάτι που τον καθιστά κατάλληλο για την αναγνώριση ύποπτων μοτίβων. Η χρήση του είναι προληπτική, καθώς επιτρέπει την ανίχνευση ύποπτων δραστηριοτήτων πριν αυτές εξελιχθούν σε επιθέσεις.

- **SVM (Support Vector Machine)**: Εφαρμόζεται στην ανίχνευση μοτίβων σε δεδομένα και είναι ιδιαίτερα αποδοτικός για την πρόληψη επιθέσεων phishing, καθώς και άλλων επιθέσεων που βασίζονται σε ανάλυση συμπεριφοράς ή κειμένου. Η προληπτική του συνεισφορά έγκειται στη δυνατότητα εντοπισμού πιθανών απειλών μέσω χαρακτηριστικών μοτίβων.
- **Random Forest**: Αυτός ο αλγόριθμος, βασισμένος σε πολλαπλά δέντρα απόφασης, είναι ικανός να ανιχνεύει περίπλοκα μοτίβα και απειλές σε μεγάλα σύνολα δεδομένων. Η προληπτική του προσέγγιση ενισχύεται από την ικανότητά του να εντοπίζει σύνθετα χαρακτηριστικά που υποδηλώνουν επικείμενες επιθέσεις.
- **KNN (K-Nearest Neighbors)**: Κατάλληλος για την ανίχνευση μοτίβων συμπεριφοράς και ανωμαλιών, ο KNN χρησιμοποιείται σε προληπτικά συστήματα λόγω της χαμηλής υπολογιστικής ισχύος που απαιτεί, ενώ παρέχει αξιόπιστα αποτελέσματα στην αναγνώριση ύποπτων δραστηριοτήτων.
- **Logistic Regression**: Με τη δυνατότητα πρόβλεψης δυαδικών καταστάσεων, όπως "επίθεση ή όχι", ο αλγόριθμος Logistic Regression είναι ιδανικός για προληπτικά συστήματα, προσφέροντας λύσεις ανίχνευσης με χαμηλές υπολογιστικές απαιτήσεις και γρήγορη εφαρμογή.

Παρακάτω παρουσιάζονται τα κυριότερα προληπτικά αντίμετρα (ως ολοκληρωμένα εργαλεία) που χρησιμοποιούνται, με την εφαρμογή των οποίων επιτυγχάνεται πιο αποτελεσματική θωράκιση των συστημάτων.

- **Πρωτόκολλο ISO/IEC 27001 [88]:** Το ISO/IEC 27001 είναι ένα παγκοσμίως αναγνωρισμένο πρότυπο ασφάλειας πληροφοριών που περιγράφει μια συστηματική προσέγγιση για τη διαχείριση ευαίσθητων εταιρικών πληροφοριών, διασφαλίζοντας την εμπιστευτικότητα, την ακεραιότητα και τη διαθεσιμότητά τους. Είναι ιδιαίτερα αποτελεσματικό στην πρόληψη των επιθέσεων κοινωνικής μηχανικής με την καθιέρωση ενός ισχυρού συστήματος διαχείρισης ασφάλειας πληροφοριών (ISMS). Το πρωτόκολλο αυτό περιλαμβάνει εκτιμήσεις κινδύνου, προσαρμοσμένους ελέγχους ασφαλείας και εκπαίδευση του προσωπικού, οι οποίοι συλλογικά ελαχιστοποιούν το ανθρώπινο λάθος - έναν κρίσιμο παράγοντα για τις ευπάθειες της κοινωνικής μηχανικής. Το ISO/IEC 27001 επιβάλλει σαφείς πολιτικές για τον έλεγχο πρόσβασης, την ταξινόμηση των δεδομένων και τα κρυπτογραφικά μέτρα, δημιουργώντας μια πολυεπίπεδη άμυνα που αποτρέπει προληπτικά τους επιτιθέμενους. Δίνοντας έμφαση στις τακτικές αξιολογήσεις κινδύνου και στη συνεχή βελτίωση, βοηθά τους οργανισμούς να εντοπίζουν και να αντιμετωπίζουν πιθανά κενά ασφαλείας πριν αυτά γίνουν αντικείμενο εκμετάλλευσης. Το συγκεκριμένο πρότυπο εστιάζει σε ένα ευέλικτο πλαίσιο, το οποίο οι οργανισμοί μπορούν να ενισχύσουν με εργαλεία που λειτουργούν με τεχνητή νοημοσύνη, όπως συστήματα ανίχνευσης ανωμαλιών ή ανάλυση

συμπεριφοράς, εφόσον το επιθυμούν. Αυτό καθιστά το ISO/IEC 27001 ένα αποτελεσματικό και προσαρμόσιμο προληπτικό πρωτόκολλο κατά των επιθέσεων κοινωνικής μηχανικής, ευθυγραμμιζόμενο με τους ευρύτερους στόχους της κυβερνοασφάλειας.

- **Vectra AI [93]:** Η Vectra AI αποτελεί ένα σύγχρονο εργαλείο που χρησιμοποιεί την Τεχνητή Νοημοσύνη (ΤΝ) για την πρόληψη επιθέσεων κοινωνικής μηχανικής, αξιοποιώντας προηγμένες τεχνικές ανάλυσης δεδομένων και ανίχνευσης ανωμαλιών. Βασιζόμενη σε αλγόριθμους μηχανικής μάθησης (που εστιάζουν σε προβλεπτική ανάλυση, δηλαδή προβλέπουν πιθανές απειλές βάσει ιστορικών δεδομένων και τάσεων, επιτρέποντας την προληπτική λήψη μέτρων ασφαλείας), η Vectra AI έχει την ικανότητα να αναλύει τεράστιους όγκους δεδομένων, εντοπίζοντας πρότυπα και ανωμαλίες που υποδεικνύουν πιθανές παραβιάσεις ασφαλείας. Μέσω της ανάλυσης ιστορικών δεδομένων και της προσαρμογής σε νέες απειλές, το εργαλείο παρέχει προληπτική προστασία ενάντια σε κοινωνικές επιθέσεις που στοχεύουν ευάλωτους χρήστες και υποδομές. Το εργαλείο της Vectra AI, πέρα από την προληπτική του λειτουργία, παρέχει και δυνατότητες διαχείρισης περιστατικών ασφαλείας μέσω αυτοματοποιημένων διαδικασιών, μειώνοντας το περιθώριο ανθρώπινου λάθους. Αυτό καθιστά δυνατή τη γρήγορη αντίδραση σε πιθανούς κινδύνους, βελτιώνοντας την ανθεκτικότητα των συστημάτων έναντι επιθέσεων. Παράλληλα, οι αναφορές από τη Vectra AI επιτρέπουν την αποτελεσματική κατηγοριοποίηση και την εστίαση σε κρίσιμες απειλές, διασφαλίζοντας τη συνεχή προστασία έναντι εξελισσόμενων κοινωνικών επιθέσεων σε περιβάλλοντα υπολογιστικού νέφους.

- **Darktrace [94]**: Το αντίμετρο Darktrace μπορεί να παρέχει αποτελεσματική θωράκιση έναντι επιθέσεων κοινωνικής μηχανικής, αξιοποιώντας την προσέγγιση που βασίζεται στην ΤΝ για τον εντοπισμό και την αντίδραση σε μη φυσιολογικές συμπεριφορές εντός ενός δικτύου.Το Darktrace χρησιμοποιεί **μη επιβλεπόμενη μηχανική μάθηση** για να δημιουργήσει μια βασική γραμμή «κανονικής» συμπεριφοράς σε χρήστες, συσκευές και το δίκτυο στο σύνολό του. Αυτή η δυνατότητα είναι ιδιαίτερα χρήσιμη για την ανίχνευση επιθέσεων που περιλαμβάνουν Spear Phishing (επιθέσεις που αξιοποιούν προσωπικά δεδομένα για να στοχεύσουν συγκεκριμένα άτομα ή ομάδες) και Insider Threats (Κακόβουλες ενέργειες ή απροσεξίες εκ μέρους εσωτερικών χρηστών που αποκλίνουν από τη συνηθισμένη συμπεριφορά). Αυτό είναι επίσης ιδιαίτερα αποτελεσματικό για την αναγνώριση νέων ή εξελισσόμενων φορέων επίθεσης που δεν αντιστοιχούν σε προηγουμένως γνωστά μοτίβα, κάτι που αποτελεί κοινή τακτική στην κοινωνική μηχανική. Συγκεκριμένα, η **Antigena Network** του Darktrace αποτελεί βασικό συστατικό του συστήματος, χρησιμοποιώντας μηχανική μάθηση για την απομόνωση ή τον περιορισμό ύποπτης δραστηριότητας σε πραγματικό χρόνο. Εντοπίζει και αντιμετωπίζει απειλές, θέτοντας αποτελεσματικά σε καραντίνα δυνητικά εκτεθειμένους λογαριασμούς χρηστών ή συσκευές. Το συστατικό **Antigena Email** στοχεύει ειδικά την κοινωνική μηχανική που βασίζεται στο ηλεκτρονικό ταχυδρομείο,

μπλοκάροντας ή καθαρίζοντας ύποπτα μηνύματα ηλεκτρονικού ταχυδρομείου, όπως απόπειρες phishing, πριν φτάσουν στον παραλήπτη. Με την ενσωμάτωση τόσο της ασφάλειας δικτύου όσο και της ασφάλειας ηλεκτρονικού ταχυδρομείου, το Darktrace παρέχει μια πολυεπίπεδη άμυνα που ελαχιστοποιεί τις πιθανότητες μιας επιτυχημένης επίθεσης κοινωνικής μηχανικής, ενώ παράλληλα προσαρμόζεται αυτόνομα στις αναδυόμενες απειλές.

- **Cisco Umbrella [89] [95]**: Το αντίμετρο Cisco Umbrella χρησιμεύει ως ένα ισχυρό προληπτικό εργαλείο κατά των επιθέσεων κοινωνικής μηχανικής, αποκλείοντας την πρόσβαση σε κακόβουλους τομείς, διευθύνσεις IP και διευθύνσεις URL μέσω φιλτραρίσματος τομέων, προγνωστικής νοημοσύνης και ελέγχου βάσει IP. Αναλύοντας τα αιτήματα DNS και αξιοποιώντας την παγκόσμια ευφυΐα απειλών της Cisco, το Umbrella σταματά προληπτικά τις απειλές στο επίπεδο DNS, εμποδίζοντας τους χρήστες να φτάσουν σε ιστότοπους phishing ή φιλοξενίας κακόβουλου λογισμικού. Η προγνωστική νοημοσύνη που χρησιμοποιείται βασίζεται σε μοντέλα μηχανικής μάθησης, όπως **Random Forest**, **Logistic Regression**, και **Gradient Boosting**, για την ανάλυση μεγάλων ποσοτήτων δεδομένων απειλών και τον εντοπισμό πιθανών κακόβουλων δραστηριοτήτων πριν εξελιχθούν σε πραγματικές επιθέσεις. Αυτά τα μοντέλα εκπαιδεύονται σε ιστορικά δεδομένα επιθέσεων και είναι ικανά να εντοπίζουν σχέσεις μεταξύ χαρακτηριστικών που μπορεί να μην είναι εμφανείς με παραδοσιακές μεθόδους **[95]**. Επιπλέον, η ολοκληρωμένη ορατότητα του Umbrella σε όλες τις συσκευές και η ικανότητά του να προβλέπει νέες και αναδυόμενες απειλές παρέχουν ισχυρή προληπτική κάλυψη ασφαλείας. Ωστόσο, η ενσωμάτωση του πλαισίου Human-as-a-Security-Sensor (HaaSS) μπορεί να συμπληρώσει το Cisco Umbrella προσθέτοντας ένα ανθρώπινο επίπεδο ανίχνευσης. Το HaaSS δίνει τη δυνατότητα στους χρήστες να ενεργούν ως προληπτικοί αισθητήρες, αναφέροντας παραπλανητικές ενδείξεις που μπορεί να μην ενεργοποιούν την ανίχνευση σε επίπεδο τομέα **[89]**. Με μοντέλα **μηχανικής μάθησης (Random Forest, Logistic Regression, KNN)** που επικυρώνουν αυτές τις αναφορές των χρηστών, το HaaSS ενισχύει την ευφυΐα απειλών της Cisco Umbrella σε πραγματικό χρόνο, ιδίως όσον αφορά την ανίχνευση νέων μοτίβων επιθέσεων. Αυτή η ενσωμάτωση ενισχύει τις προληπτικές δυνατότητες του Cisco Umbrella, ενσωματώνοντας τόσο το αυτοματοποιημένο φιλτράρισμα τομέων όσο και τις πληροφορίες που δημιουργούνται από τους χρήστες, διαμορφώνοντας μια πιο προσαρμοστική, πολυεπίπεδη άμυνα κατά των απειλών κοινωνικής μηχανικής.

Κύρια Συνεισφορά:

Η συνεισφορά των προληπτικών αντίμετρων έγκειται στη δυνατότητά τους να εντοπίζουν πιθανές επιθέσεις προτού αυτές εξελιχθούν σε πραγματική απειλή, μειώνοντας τον

κίνδυνο ζημιάς. Τα εργαλεία όπως το Cisco Umbrella, το Darktrace, το Vectra AI προσφέρουν προηγμένες λύσεις για την παρακολούθηση και ανάλυση μοτίβων σε ψηφιακό επίπεδο, αποτρέποντας πιθανά περιστατικά μέσα από τη φιλτράρισμα και τον αποκλεισμό κακόβουλων τομέων, διευθύνσεων URL και μη φυσιολογικών συμπεριφορών πριν φτάσουν στους τελικούς χρήστες. Αυτές οι τεχνολογίες, αξιοποιώντας τεράστιες βάσεις δεδομένων απειλών και πληροφοριών, επιτρέπουν την πρόβλεψη και την προσαρμογή στα αναδυόμενα μοτίβα επιθέσεων. Ειδικότερα, οι λύσεις, όπως το Darktrace και το Vectra AI, χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για την ανίχνευση ανωμαλιών και την κατηγοριοποίηση κινδύνων, προσφέροντας ένα εξαιρετικά ανθεκτικό και προληπτικό επίπεδο άμυνας που είναι ικανό να μειώσει την έκθεση σε phishing, κακόβουλο λογισμικό και άλλες μορφές κοινωνικής μηχανικής. Παράλληλα, εργαλεία, όπως το Cisco Umbrella, προσφέρουν προληπτική κάλυψη μέσω φιλτραρίσματος DNS και τον αποκλεισμό κακόβουλων τοποθεσιών σε πραγματικό χρόνο. Οι αλγόριθμοι μηχανικής μάθησης ενισχύουν περαιτέρω την άμυνα, επικυρώνοντας και ιεραρχώντας αναφορές απειλών, μειώνοντας τα ψευδή θετικά και διασφαλίζοντας ότι οι πραγματικές απειλές αντιμετωπίζονται αποτελεσματικά. Αυτός ο δυναμικός συνδυασμός εργαλείων και αλγορίθμων προσφέρει μια ολοκληρωμένη, προσαρμοστική άμυνα, ικανή να προβλέπει και να μετριάζει εξελισσόμενες τακτικές επιθέσεων, ενισχύοντας έτσι την ασφάλεια των οργανισμών απέναντι σε μια ολοένα και πιο περίπλοκη απειλή κοινωνικής μηχανικής.

Ωστόσο, κάθε εργαλείο έχει και την δική του αρνητική πτυχή. Το **ISO/IEC 27001**, αν και παρέχει ένα ισχυρό πλαίσιο διαχείρισης της ασφάλειας των πληροφοριών, είναι γενικό και απαιτεί σημαντικούς πόρους για την προσαρμογή του στις ανάγκες κάθε οργανισμού, γεγονός που μπορεί να αποδειχθεί δαπανηρό και χρονοβόρο για μικρές και μεσαίες επιχειρήσεις, ενώ παράλληλα είναι άμεσα εξαρτώμενο από την συμμόρφωση των ατόμων. Το **Vectra AI**, ενώ προσφέρει αποτελεσματική ανίχνευση τεχνικών ανωμαλιών, ενδέχεται να μην εντοπίζει αποτελεσματικά επιθέσεις κοινωνικής μηχανικής που βασίζονται σε ανθρώπινες αδυναμίες. Επιπλέον, υπάρχει η πιθανότητα ψευδώς θετικών ειδοποιήσεων, οι οποίες μπορεί να προκαλέσουν κόπωση στους αναλυτές ασφάλειας. Το **Darktrace**, από την άλλη πλευρά, βασίζεται στην ανίχνευση ανωμαλιών, κάτι που μπορεί να μην είναι αρκετό για επιθέσεις κοινωνικής μηχανικής χωρίς εμφανείς τεχνικές ενδείξεις. Παράλληλα, η πολυπλοκότητα της εγκατάστασης και το κόστος χρήσης του το καθιστούν δύσκολα προσβάσιμο σε μικρότερες επιχειρήσεις. Τέλος, το **Cisco Umbrella**, παρόλο που παρέχει ισχυρή προληπτική κάλυψη μέσω φιλτραρίσματος DNS και αποκλεισμού κακόβουλων τοποθεσιών, περιορίζεται στο επίπεδο DNS, κάτι που μπορεί να μην επαρκεί για πιο περίπλοκες μορφές επιθέσεων κοινωνικής μηχανικής. Επίσης, η αποτελεσματικότητά του εξαρτάται από τη συμβατότητα με τις υπάρχουσες υποδομές.

4.3.2 Αντίμετρα για Ανίχνευση σε Πραγματικό Χρόνο

Τα αντίμετρα αυτά παρακολουθούν τη δραστηριότητα του δικτύου και ανιχνεύουν επιθέσεις την στιγμή της εκτέλεσής τους. Συνεπώς, τα αντίμετρα παρακολουθούν συνεχώς τη ροή δεδομένων και ανιχνεύουν επιθέσεις καθώς αυτές συμβαίνουν.

- Οι αλγόριθμοι ΤΝ που ενσωματώνουν τα αντίμετρα αυτά είναι οι εξής:
    - **IDS & Machine Learning**: Τα συστήματα Intrusion Detection Systems (IDS) σε συνδυασμό με Machine Learning μοντέλα χρησιμοποιούνται για την ανίχνευση εισβολών σε πραγματικό χρόνο μέσω της συνεχούς παρακολούθησης δικτύων ή συστημάτων.
    - **Neural Network-based Detection**: Τα νευρωνικά δίκτυα είναι εξαιρετικά στην αναγνώριση μοτίβων σε δεδομένα πραγματικού χρόνου και συχνά χρησιμοποιούνται σε συστήματα παρακολούθησης και ανίχνευσης επιθέσεων.
    - **RNN**: Τα Recurrent Neural Networks είναι κατάλληλα για την ανίχνευση ακολουθιών δεδομένων, κάτι που τα καθιστά χρήσιμα για την ανίχνευση επιθέσεων, όπως τα robocalls και οι διαδοχικές phishing επιθέσεις, σε πραγματικό χρόνο.
    - **Isolation Forest**: Χρησιμοποιείται για την ανίχνευση ανωμαλιών και επιθέσεων κατά τη διάρκεια της ροής δεδομένων, καθιστώντας τον κατάλληλο για χρήση σε πραγματικό χρόνο.
    - **AdaBoost**: Χρησιμοποιείται για την ενίσχυση της ακρίβειας άλλων αλγορίθμων σε πραγματικό χρόνο, ενσωματώνοντας νέες παρατηρήσεις από τα δεδομένα που επεξεργάζονται για ανίχνευση επιθέσεων. Με αυτό τον τρόπο, το παραγόμενο σύνθετο μοντέλο παρουσιάζει περισσότερη ακρίβεια ανίχνευσης.
    - **Gradient Boosting**: Ένας ισχυρός αλγόριθμος για ανίχνευση ανωμαλιών σε πραγματικό χρόνο, αφού συνεχώς βελτιώνει την ακρίβεια του (σύνθετου) μοντέλου του (τύπου ensemble), αναγνωρίζοντας απειλές σε εισερχόμενα δεδομένα.
    - **MDP (Markov Decision Process)**: Κατάλληλο για διαδοχικές αποφάσεις σε πραγματικό χρόνο, για ανίχνευση και απόκριση επιθέσεων κατά τη διάρκεια που συμβαίνουν. Συνήθως χρησιμοποιείται σε συνδυασμό με το Reinforcement Learning (RL), το οποίο αξιοποιεί τον MDP ως ένα μαθηματικό πλαίσιο για τη μοντελοποίηση των αλληλεπιδράσεων και τη συνεχή βελτίωση της στρατηγικής απόκρισης. Αυτός ο συνδυασμός επιτρέπει στο σύστημα να μαθαίνει δυναμικά και να προσαρμόζεται σε νέες τακτικές επιθέσεων, εξασφαλίζοντας έτσι αποτελεσματική απόκριση.
    - **GMM (Gaussian Mixture Models)**: Ιδανικός για την ανάλυση συνεχών δεδομένων, όπως ήχος ή φωνητικά δεδομένα, οπότε ο αλγόριθμος GMM

μπορεί να χρησιμοποιηθεί για ανίχνευση σε πραγματικό χρόνο επιθέσεων vishing.

Παρακάτω παρουσιάζονται τα κυριότερα αντίμετρα για ανίχνευση σε πραγματικό χρόνο, με την εφαρμογή των οποίων επιτυγχάνεται η άμεση ανίχνευση και απόκριση σε επιθέσεις καθώς αυτές συμβαίνουν:

- **Splunk Enterprise Security [96]**: Το Splunk Enterprise Security (ES) αποτελεί μια ολοκληρωμένη λύση SIEM (Security Information and Event Management) που προσφέρει σε πραγματικό χρόνο παρακολούθηση, ανάλυση και ανίχνευση απειλών, συμπεριλαμβανομένων των επιθέσεων κοινωνικής μηχανικής. Ενσωματώνει προηγμένες αναλυτικές μεθόδους και αλγορίθμους μηχανικής μάθησης για τον εντοπισμό ανωμαλιών και ύποπτων δραστηριοτήτων που σχετίζονται με τέτοιες επιθέσεις. Επιπλέον, το Splunk ES συλλέγει συνεχώς δεδομένα από διάφορες πηγές της υποδομής IT του οργανισμού, παρακολουθώντας τη συμπεριφορά χρηστών και τη δραστηριότητα δικτύου. Με αυτόν τον τρόπο, εντοπίζει μοτίβα που αποκλίνουν από το φυσιολογικό, όπως ασυνήθιστες προσπάθειες σύνδεσης ή πρόσβαση σε ευαίσθητες πληροφορίες. Όταν εντοπιστούν τέτοιες ανωμαλίες, το Splunk ES δημιουργεί ειδοποιήσεις, επιτρέποντας στις ομάδες ασφαλείας να ερευνήσουν και να ανταποκριθούν άμεσα, μειώνοντας τις πιθανότητες να εκμεταλλευτούν οι επιτιθέμενοι ευπάθειες.
  Θα πρέπει να τονιστεί πως το Η Splunk ES ενσωματώνει πληθώρα αλγορίθμων μηχανικής μάθησης, χρησιμοποιώντας πάνω από 250 κανόνες βασισμένους σε μηχανική μάθηση. Υποστηρίζει μοντέλα όπως:
  - *Batch models*, για μακροχρόνια ανάλυση.
  - *Streaming models,* για ανίχνευση απειλών σε πραγματικό χρόνο.

  Οι αλγόριθμοι αυτοί επιτρέπουν στο σύστημα να μαθαίνει από ιστορικά δεδομένα, να αναγνωρίζει αναδυόμενα μοτίβα κακόβουλης συμπεριφοράς και να προσαρμόζεται σε νέες επιθέσεις. Παρόλο που το Splunk ES υποστηρίζει ευελιξία στην επιλογή αλγορίθμων μέσω του Machine Learning Toolkit (MLTK), όπως **Neural Networks, AdaBoost** και **Gaussian Mixture Models (GMM)**, η επιλογή συγκεκριμένου αλγορίθμου εξαρτάται από τις ανάγκες του οργανισμού και τη φύση των απειλών που πρέπει να εντοπιστούν. Τέλος, το Splunk MLTK επιτρέπει την προσαρμογή και την ενσωμάτωση πρόσθετων αλγορίθμων, διευρύνοντας τις δυνατότητες ανίχνευσης και βελτιώνοντας τους μηχανισμούς ασφαλείας. Αυτή η ευελιξία επιτρέπει στους οργανισμούς να διαμορφώσουν τις τεχνικές ανίχνευσης σύμφωνα με τις ιδιαίτερες απαιτήσεις τους και τις εξελισσόμενες απειλές. Έτσι, μέσα από αυτή τη δυναμική προσέγγιση, το Splunk ES παρέχει ισχυρή, σε πραγματικό χρόνο ανίχνευση και απόκριση, αποτελώντας ένα πολύτιμο εργαλείο για την προστασία απέναντι σε επιθέσεις κοινωνικής μηχανικής.

- **IBM QRadar Security Information and Event Management (SIEM) [97] [98]**: Το **IBM QRadar SIEM** αποτελεί μια ολοκληρωμένη λύση διαχείρισης πληροφοριών και συμβάντων ασφαλείας (SIEM) που εξειδικεύεται σε τομείς, όπως η ανάλυση συμπεριφοράς χρηστών (User Behavior Analytics - UBA), η ανίχνευση ανωμαλιών και η νοημοσύνη απειλών (Threat Intelligence). Χρησιμοποιεί προηγμένες τεχνολογίες, όπως το **IBM Watson** που ενσωματώνει δυνατότητες Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) και Μηχανικής Μάθησης (ML), ενισχύοντας τις αναλυτικές του δυνατότητες. Στον τομέα της ανίχνευσης ανωμαλιών, το QRadar βασίζεται σε αλγορίθμους μη επιβλεπόμενης μάθησης, όπως **Isolation Forest** και **Διεργασίες Αποφάσεων Markov (Markov Decision Processes - MDP)**. Ο **Isolation Forest** αποδεικνύεται εξαιρετικά αποτελεσματικός στον εντοπισμό εξαιρετικών τιμών (outliers), μέσω της επαναληπτικής κατάτμησης που απομονώνει τις ανωμαλίες. Οι **MDPs**, από την άλλη, μοντελοποιούν διαδικασίες λήψης αποφάσεων όπου τα αποτελέσματα εξαρτώνται εν μέρει από την τυχαιότητα και εν μέρει από τις ενέργειες των χρηστών, βοηθώντας στην κατανόηση και πρόβλεψη σύνθετων συμπεριφορών. Η ενσωμάτωση του **IBM Watson** επιτρέπει στο QRadar να επεξεργάζεται και να αναλύει τεράστιες ποσότητες μη δομημένων δεδομένων, ενισχύοντας την ικανότητά του να εντοπίζει προηγμένες απειλές μέσω της κατανόησης του πλαισίου και της πρόθεσης, κάτι κρίσιμο για την αναγνώριση επιθέσεων κοινωνικής μηχανικής. Επιπλέον, το **UBA** του QRadar αναλύει τις δραστηριότητες των χρηστών για τη δημιουργία βάσεων φυσιολογικής συμπεριφοράς και την ανίχνευση αποκλίσεων που μπορεί να υποδεικνύουν κακόβουλη πρόθεση. Μέσω παρακολούθησης μοτίβων, όπως οι ώρες σύνδεσης, οι προσβάσεις σε πόρους και οι όγκοι μεταφοράς δεδομένων, εντοπίζονται πιθανές εσωτερικές απειλές ή λογαριασμοί που έχουν παραβιαστεί. Τέλος, η συνεχής ενσωμάτωση ροών απειλητικής νοημοσύνης διασφαλίζει ότι το σύστημα ενημερώνεται για νέες απειλές, συμπεριλαμβανομένων αυτών που χρησιμοποιούνται σε κοινωνικές επιθέσεις. Η συνδυαστική αυτή προσέγγιση, που βασίζεται σε αλγορίθμους όπως το **Isolation Forest** και οι **Διεργασίες Αποφάσεων Markov**, σε συνδυασμό με τις δυνατότητες του IBM Watson, καθιστά το QRadar ένα ισχυρό εργαλείο για την ανίχνευση επιθέσεων κοινωνικής μηχανικής.

- **McAfee Enterprise Security Manager (ESM) [65] [99]:** Το **McAfee** ενισχύει την ανίχνευση σε πραγματικό χρόνο και την αντιμετώπιση επιθέσεων κοινωνικής μηχανικής με την ενσωμάτωση προηγμένων πληροφοριών απειλών και ανάλυσης συμπεριφοράς χρηστών (**UBA**) στις δυνατότητες ενός προϊόντος τύπου Security Information and Event Management (SIEM). Αξιοποιώντας τροφοδοσίες πληροφοριών απειλών που εστιάζουν σε τακτικές κοινωνικής μηχανικής, το ESM εντοπίζει μοτίβα που σχετίζονται με αυτές τις επιθέσεις, επισημαίνοντας πιθανές απειλές με βάση

ύποπτες δραστηριότητες. Επιπλέον, το UBA που περιλαμβάνει αλγόριθμους τεχνητής νοημοσύνης, όπως ο **Random Forest, IDS & Machine Learning, Gradient Boosting, RNN**, επιτρέπει στο McAfee ESM να παρακολουθεί και να καθορίζει βασικές συμπεριφορές χρηστών, ειδοποιώντας τους διαχειριστές όταν εμφανίζονται αποκλίσεις - όπως ασυνήθιστοι χρόνοι σύνδεσης ή άτυπη πρόσβαση σε πόρους - που σηματοδοτούν πιθανές παραβιάσεις κοινωνικής μηχανικής. Η ενσωμάτωση του ESM με λύσεις ασφάλειας ηλεκτρονικού ταχυδρομείου του επιτρέπει να ανιχνεύει και να συσχετίζει προσπάθειες phishing, παρέχοντας μια ολοκληρωμένη, πολυδιάστατη εικόνα των απειλών σε όλο το δίκτυο. Η ειδοποίηση σε πραγματικό χρόνο και η αυτοματοποίηση της αντιμετώπισης περιστατικών στο πλαίσιο του ESM επιτρέπουν την ταχεία ανάληψη δράσης, όπως η απομόνωση των επηρεαζόμενων συστημάτων, για τον περιορισμό των παραβιάσεων κατά την εμφάνισή τους. Αυτή η πολυεπίπεδη προσέγγιση, η οποία συνδυάζει την αυτοματοποιημένη ανίχνευση με τις γνώσεις συμπεριφοράς, προσφέρει μια ισχυρή, προληπτική άμυνα κατά των τακτικών κοινωνικής μηχανικής, υπογραμμίζοντας τη σημασία των συστημάτων SIEM στην αντιμετώπιση σύνθετων απειλών στον κυβερνοχώρο με επίκεντρο τον άνθρωπο.

- **CrowdStrike Falcon [64]:** Το CrowdStrike Falcon αποτελεί μια προηγμένη πλατφόρμα κυβερνοασφάλειας τύπου EDR (Endpoint Detection & Response), που αξιοποιεί την Τεχνητή Νοημοσύνη και τη μηχανική μάθηση για την ανίχνευση και την απόκριση σε πραγματικό χρόνο απέναντι σε επιθέσεις κοινωνικής μηχανικής. Μέσω συνεχούς παρακολούθησης τερματικών και ανάλυσης συμπεριφορικών προτύπων, το Falcon εντοπίζει και μετριάζει κακόβουλες δραστηριότητες κατά τη στιγμή που συμβαίνουν. Η πλατφόρμα χρησιμοποιεί αλγόριθμους μηχανικής μάθησης, όπως **Neural Networks, Decision Trees και Anomaly Detection Models**, οι οποίοι εφαρμόζονται για τη δημιουργία βάσεων φυσιολογικής συμπεριφοράς και την ανίχνευση αποκλίσεων από αυτές. Αυτή η διαδικασία επιτρέπει στο Falcon να εντοπίζει μοτίβα κακόβουλης δραστηριότητας, όπως ανεξήγητες συνδέσεις, ασυνήθιστες μεταφορές δεδομένων, ή επιθέσεις spear phishing. Επιπλέον, το Falcon αξιοποιεί τη συλλογική νοημοσύνη από το δίκτυο πελατών του (CrowdStrike Threat Graph) για την ανάλυση δισεκατομμυρίων συμβάντων καθημερινά, επιτρέποντας την προσαρμογή του σε νέα μοτίβα επιθέσεων. Η τεχνολογία του ενσωματώνει τόσο **unsupervised learning** για την ανίχνευση ανωμαλιών, όσο και **supervised learning** για την ταξινόμηση και πρόβλεψη κινδύνων. Μέσω αυτών των τεχνολογιών, η πλατφόρμα προσφέρει ολοκληρωμένη προστασία σε επίπεδο τελικών σημείων (EDR), καθιστώντας την ιδανική για την αντιμετώπιση επιθέσεων κοινωνικής μηχανικής σε πραγματικό χρόνο.
  Τέλος, το Falcon ενισχύει την ασφάλεια ενός οργανισμού μέσω της ενσωμάτωσής του με άλλα εργαλεία ασφάλειας και συστήματα SIEM, επιτρέποντας την ενοποιημένη παρακολούθηση και ανάλυση των απειλών.

Κύρια Συνεισφορά:

Η βασική συνεισφορά των αντιμέτρων ανίχνευσης σε πραγματικό χρόνο, όπως το Splunk ES, IBM QRadar, McAfee ESM και CrowdStrike Falcon, έγκειται στην ικανότητά τους να ανιχνεύουν επιθέσεις ακριβώς τη στιγμή που εκδηλώνονται, επιτρέποντας την έγκαιρη απόκριση και τον περιορισμό της ζημιάς. Αυτά τα συστήματα προσφέρουν συνεχή παρακολούθηση δραστηριοτήτων, αναλύοντας δεδομένα από πολλαπλές πηγές και παρέχοντας ολοκληρωμένες πληροφορίες σχετικά με ύποπτες συμπεριφορές. Η ενσωμάτωση προηγμένων αλγορίθμων τεχνητής νοημοσύνης, όπως Neural Networks, Isolation Forest, Gradient Boosting, και μοντέλα ανίχνευσης ανωμαλιών, ενισχύει την ακρίβεια της ανίχνευσης, περιορίζοντας ψευδώς θετικά αποτελέσματα και προσαρμόζοντας τα συστήματα σε νέες και εξελισσόμενες τακτικές επιθέσεων. Η ικανότητα αναγνώρισης μοτίβων, ανίχνευσης μη φυσιολογικής συμπεριφοράς και χρήσης προσαρμοστικής μάθησης επιτρέπει στα αντίμετρα να εντοπίζουν ακόμα και τις πιο λεπτές αλλαγές στις τακτικές των επιθέσεων κοινωνικής μηχανικής, που συχνά παραβλέπονται από παραδοσιακές μεθόδους.

Ωστόσο, κάθε εργαλείο έχει και τα αρνητικά του. Το **Splunk ES**, αν και ισχυρό, χαρακτηρίζεται από υψηλό κόστος υλοποίησης και συντήρησης, καθιστώντας το λιγότερο προσιτό για μικρές και μεσαίες επιχειρήσεις, ενώ παράλληλα η απόδοση της πλατφόρμας μπορεί να υποβαθμιστεί σημαντικά υπό μεγάλο φόρτο εργασίας, οδηγώντας σε αργή λειτουργία σε περιβάλλοντα με μεγάλα σύνολα δεδομένων ή εκτεταμένα αρχεία καταγραφής. Το ζήτημα αυτό επιδεινώνεται από τον απαιτητικό χαρακτήρα του, ο οποίος μπορεί να επιβαρύνει τους πόρους του συστήματος, ιδίως σε εταιρείες με περιορισμένη υποδομή. Επιπλέον, η πολυπλοκότητα του Splunk στην εγκατάσταση και τη διαμόρφωση δημιουργεί προκλήσεις για τους άπειρους χρήστες, απαιτώντας σημαντική τεχνογνωσία για την αποτελεσματική εγκατάσταση και διαχείριση **[45]**.

Το **IBM QRadar** παρουσιάζει επίσης αρκετά μειονεκτήματα, παρά τις ισχυρές του δυνατότητες. Ένας σημαντικός περιορισμός είναι ο ιδιόκτητος χαρακτήρας του, που απαιτεί ένα δαπανηρό μοντέλο αδειοδότησης, το οποίο μπορεί να το καταστήσει απρόσιτο για οργανισμούς με περιορισμένο προϋπολογισμό. Επιπλέον, τα προηγμένα χαρακτηριστικά του συχνά συνοδεύονται από μια απότομη καμπύλη εκμάθησης, η οποία απαιτεί ουσιαστική εκπαίδευση και εξοικείωση με το ολοκληρωμένο περιβάλλον εργασίας του για αποτελεσματική χρήση. Επιπλέον, ζητήματα, όπως οι καθυστερημένες ενημερώσεις συμβατότητας με συστήματα τρίτων και οι νέες εκδόσεις λογισμικού, ενδέχεται να εμποδίζουν την προσαρμοστικότητα σε πραγματικό χρόνο. **[46]**.

Το **McAfee ESM** αντιμετωπίζει προκλήσεις που σχετίζονται με την εξάρτησή του από εκτεταμένα αρχεία καταγραφής δεδομένων για την απόκτηση επίγνωσης της κατάστασης, γεγονός που μπορεί να οδηγήσει σε υπερφόρτωση δεδομένων, δυσχεραίνοντας τη διαχείριση από τις ομάδες πληροφορικής και καθυστερώντας την άμεση εξαγωγή χρήσιμων πληροφοριών **[47]**.

Τέλος, με βάση το άρθρο **[48]**, το **CrowdStrike Falcon** παρουσιάζει ορισμένα μειονεκτήματα παρά τις ολοκληρωμένες δυνατότητές του. Ένας από τους βασικούς περιορισμούς του είναι η εξάρτησή του από τη συνεχή συνδεσιμότητα στο διαδίκτυο, καθώς πρόκειται κυρίως για λύση που βασίζεται στο υπολογιστικό νέφος. Αυτή η εξάρτηση μπορεί να είναι προβληματική για οργανισμούς που λειτουργούν σε περιβάλλοντα με αναξιόπιστη ή περιορισμένη πρόσβαση στο διαδίκτυο. Επίσης, ενώ η πλατφόρμα είναι εξαιρετικά αποτελεσματική στον εντοπισμό απειλών, το κόστος αδειοδότησης μπορεί να είναι απαγορευτικά ακριβό για τους μικρότερους οργανισμούς ή εκείνους με περιορισμένο προϋπολογισμό. Επιπλέον, η πολυπλοκότητα των προηγμένων χαρακτηριστικών του μπορεί να απαιτεί ειδική εμπειρογνωμοσύνη για τη σωστή διαμόρφωση και λειτουργία, καθιστώντας το CrowdStrike Falcon ως την λιγότερο προσιτή λύση σε οργανισμούς που δεν διαθέτουν εξειδικευμένη ομάδα ασφάλειας πληροφορικής.

Παρά τα αρνητικά του κάθε εργαλείου, είναι σημαντικό να ειπωθεί πως η πολυεπίπεδη άμυνα που παρέχεται μέσω αυτών των τεχνολογιών συνδυάζει παρακολούθηση σε πραγματικό χρόνο με έξυπνη και προσαρμοστική ανάλυση, καθιστώντας τα ικανά να διαχειρίζονται σύνθετες, ανθρωποκεντρικές απειλές στον κυβερνοχώρο. Με αυτόν τον τρόπο, τα εργαλεία αυτά ενισχύουν τη συνολική ανθεκτικότητα των συστημάτων ασφαλείας, παρέχοντας στους οργανισμούς τη δυνατότητα να ανταποκριθούν αποτελεσματικά στις εξελίξεις του σύγχρονου κυβερνοεγκλήματος.

4.3.3 Αντιδραστικά Αντίμετρα

Τα αντίμετρα αυτού του είδους ενεργοποιούνται μετά την αναγνώριση μιας επίθεσης με σκοπό να την αντιμετωπίσουν ή να μετριάσουν τις επιπτώσεις της.

- Ενσωματώνονται οι εξής αλγόριθμοι στα αντιδραστικά αντίμετρα:
    - **GANs**: Χρησιμοποιούνται αντιδραστικά για την αναγνώριση ψεύτικων δεδομένων ή ταυτοτήτων, ειδικά μετά από μια επίθεση. Μπορούν να χρησιμοποιηθούν για να δημιουργήσουν αντιδραστικά αντίμετρα σε επιθέσεις παραποίησης.
    - **MSPAD-SGM (Multistage Signaling Game Model) [66]**: Βασίζεται στη θεωρία παιγνίων και χρησιμοποιείται για τη δυναμική αντίδραση σε επιθέσεις αφού ανιχνευθούν, βοηθώντας στη διαμόρφωση στρατηγικών κατά επιθέσεων spear-phishing ή άλλων επιθέσεων κοινωνικής μηχανικής.
    - **Reinforcement Learning**: Μπορεί να χρησιμοποιηθεί και σε αντιδραστικά συστήματα που προσαρμόζουν τη στρατηγική τους με βάση τις ενέργειες του επιτιθέμενου, ιδιαίτερα σε περιβάλλοντα που απαιτούν συνεχή μάθηση για τον εντοπισμό των καταλληλότερων αντιδράσεων του συστήματος, σε αντίστοιχες ενεργειες του επιτιθέμενου.

- **Isolation Forest**: Χρησιμοποιείται για την ανίχνευση ανωμαλιών και επιθέσεων κατά τη διάρκεια της ροής δεδομένων, καθιστώντας τον κατάλληλο για χρήση σε πραγματικό χρόνο. Προφανώς, χρησιμοποιείται μόνο για την ανίχνευση σε αυτή την περίπτωση. Συνεπώς, η αντίδραση θα πρέπει να υλοποιηθεί ξεχωριστά.
- **SVM (Support Vector Machine)**: Χρησιμοποιείται για την ανίχνευση μοτίβων σε δεδομένα, κατάλληλο για την πρόληψη επιθέσεων phishing και άλλων επιθέσεων που βασίζονται σε δεδομένα συμπεριφοράς ή κειμένου. Και σε αυτή την περίπτωση, η αντίδραση πρέπει να υλοποιηθεί ξεχωριστά.

Παρακάτω παρουσιάζονται τα κυριότερα αντιδραστικά αντίμετρα, τα οποία χρησιμοποιούνται για την αντιμετώπιση των επιθέσεων μόλις αυτές ανιχνευτούν, προς περιορισμό της προκληθείσας ζημιάς ή αποτροπή περαιτέρω μελλοντικής εμφάνισης της επίθεσης.

- **Cortex XDR (Palo Alto Networks) [18]:** Το Cortex XDR είναι ένα εργαλείο EDR (Endpoint Detection and Response) που αξιοποιεί την Τεχνητή Νοημοσύνη για την ανίχνευση, ανάλυση και αποκατάσταση επιθέσεων. Εστιάζει κυρίως στη συμπεριφορική ανάλυση χρηστών και συστημάτων, προσφέροντας λύσεις για την αποκατάσταση από ransomware, παραβιάσεις δεδομένων και άλλες επιθέσεις. Το συγκεκριμένο εργαλείο, χρησιμοποιεί μηχανική μάθηση για την ανίχνευση ανωμαλιών, **unsupervised learning** για την προσαρμογή σε νέα μοτίβα επιθέσεων, καθώς και supervised learning για την κατηγοριοποίηση και πρόβλεψη κινδύνων. Ως προς τις λειτουργίες του, παρέχει **root cause analysis**, αποκλεισμό κακόβουλου λογισμικού και πλήρη αποκατάσταση μετά από επιθέσεις. Η συλλογική νοημοσύνη του επιτρέπει την αναγνώριση μοτίβων κακόβουλης δραστηριότητας μέσω διαρκούς ανάλυσης δεδομένων. Συνδυάζει προηγμένους αλγορίθμους ΤΝ με ολοκληρωμένη παρακολούθηση τελικών σημείων, καθιστώντας το εξαιρετικό για την αποκατάσταση κοινωνικών και άλλων επιθέσεων.

- **IBM Resilient Incident Response Platform [19]**: Το IBM Resilient αποτελεί μια ολοκληρωμένη πλατφόρμα για τη διαχείριση περιστατικών, εστιάζοντας στην αυτοματοποίηση της απόκρισης και της αποκατάστασης. Είναι ιδανικό για την αντιμετώπιση επιθέσεων κοινωνικής μηχανικής, όπως phishing και credential harvesting, και χρησιμοποιείται ευρέως σε οργανισμούς με ανάγκες υψηλού επιπέδου ασφάλειας. Ως προς τους αλγορίθμους ΤΝ, ενσωματώνει το **Watson**, που συνδυάζει **NLP** και **ML**, για την κατανόηση και ανάλυση σύνθετων περιστατικών. Επίσης, χρησιμοποιεί **unsupervised learning** για την ανίχνευση ανωμαλιών και **supervised**

learning για την ταξινόμηση περιστατικών. Όσον αφορά τις λειτουργίες του, το IBM Resilient παρέχει αυτοματοποιημένες προτάσεις αποκατάστασης, βασισμένες στην ανάλυση των περιστατικών, και εντοπίζει διαρροές δεδομένων, βοηθώντας στην άμεση αποκατάσταση. Παράλληλα, υποστηρίζει τη δημιουργία σχεδίων ανάκαμψης και τη συνεχή παρακολούθηση κινδύνων. Τέλος, η χρήση του Watson παρέχει ένα ισχυρό πλεονέκτημα στην αναγνώριση και ανάλυση κοινωνικών επιθέσεων, καθιστώντας το IBM Resilient εξαιρετικά αποτελεσματικό για οργανισμούς που χρειάζονται άμεση απόκριση.

- **Automation and Orchestration Platforms (SOAR) [99]**: Οι πλατφόρμες αυτοματοποίησης και ενορχήστρωσης (Security Orchestration, Automation, and Response - SOAR) παρέχουν σημαντική υποστήριξη στην ανάκαμψη ενός οργανισμού μετά από μια επίθεση κοινωνικής μηχανικής. Αυτές οι πλατφόρμες αυτοματοποιούν τις διαδικασίες απόκρισης και αποκατάστασης, επιτρέποντας την ταχύτερη απομόνωση, ανάλυση και ανάκαμψη των συστημάτων. Χρησιμοποιώντας προηγμένα εργαλεία τεχνητής νοημοσύνης και μηχανικής μάθησης, μπορούν να ανιχνεύσουν ανωμαλίες, να προβλέψουν τον κίνδυνο επανεμφάνισης επιθέσεων και να βελτιστοποιήσουν στρατηγικές ανάκαμψης μέσω reinforcement learning. Με τον συντονισμό δεδομένων από διάφορες πηγές (π.χ. SIEM, IDS/IPS) και την εκτέλεση προκαθορισμένων playbooks, διασφαλίζουν την ταχεία αποκατάσταση από παραβιάσεις, την εξάλειψη κακόβουλων παραμέτρων και την επαναφορά των συστημάτων σε πλήρη λειτουργικότητα. Επίσης, ενσωματώνονται με συστήματα αντιγράφων ασφαλείας (backups) για την ταχεία αποκατάσταση των προσβεβλημένων συστημάτων. Οι πλατφόρμες SOAR αξιοποιούν αλγορίθμους μηχανικής μάθησης για την ενίσχυση της ακρίβειας των ανιχνεύσεων και της πρόβλεψης επιθέσεων. Αλγόριθμοι όπως Isolation Forest βοηθούν στον εντοπισμό ανωμαλιών, ενώ οι Random Forest και Support Vector Machine (SVM) συμβάλλουν στην ταξινόμηση των απειλών και την πρόβλεψη του κινδύνου επανεμφάνισης. Μέσω Reinforcement Learning, αυτές οι πλατφόρμες βελτιστοποιούν τις στρατηγικές απόκρισης, προσαρμόζοντας δυναμικά τις διαδικασίες για να ενισχύσουν την ανθεκτικότητα των οργανισμών απέναντι σε μελλοντικές επιθέσεις.

- **Disaster Recovery as a Service (DRaaS) [99]**: Το DRaaS προσφέρει υπηρεσίες αντιγράφων ασφαλείας και ανάκτησης στο cloud, επιτρέποντας στους οργανισμούς να αποκαθιστούν γρήγορα δεδομένα και συστήματα που έχουν παραβιαστεί. Με την τακτική αναπαραγωγή κρίσιμων δεδομένων και διαμορφώσεων συστημάτων σε ασφαλή τοποθεσία εκτός του κύριου δικτύου, το DRaaS διασφαλίζει την ακεραιότητα των δεδομένων ακόμη και μετά από επίθεση. Αυτό το εργαλείο επιτρέπει την απρόσκοπτη ανάκτηση εφαρμογών, αρχείων και ρυθμίσεων δικτύου, μειώνοντας τον

χρόνο εκτός λειτουργίας και αποτρέποντας την απώλεια δεδομένων. Επιπλέον, οι πλατφόρμες DRaaS συχνά υποστηρίζουν αυτόματη μεταφορά στο εφεδρικό περιβάλλον, διασφαλίζοντας τη συνέχεια των επιχειρησιακών λειτουργιών και ελαχιστοποιώντας την επιχειρησιακή διακοπή μετά από μια επίθεση κοινωνικής μηχανικής. Παράλληλα, με τη χρήση **GANs**, οι πλατφόρμες ανάκαμψης μπορούν να δημιουργήσουν ρεαλιστικά σενάρια επίθεσης για εκπαιδευτικούς σκοπούς, δοκιμάζοντας την ικανότητα των συστημάτων να εντοπίζουν και να αποκρίνονται σε εξελιγμένες τακτικές κοινωνικής μηχανικής. Αυτό ενισχύει την ανθεκτικότητα του συστήματος, καθώς οι οργανισμοί μπορούν να προετοιμαστούν καλύτερα για άγνωστες ή σύνθετες επιθέσεις μέσω ρεαλιστικών προσομοιώσεων. Επίσης, με την χρήση του **MSPAD-SGM**, τα συστήματα ανάκαμψης μπορούν να βελτιστοποιήσουν τη στρατηγική τους στην ανίχνευση και απόκριση με βάση τη δυναμική των επιτιθέμενων και την ανταπόκριση των αμυντικών μηχανισμών. Αυτό το μοντέλο, που βασίζεται σε παιχνίδι πολλαπλών σταδίων, αναλύει τη συμπεριφορά των επιτιθέμενων και επιλέγει κατάλληλες τακτικές για να αποτρέψει περαιτέρω επιθέσεις ή να περιορίσει τη ζημιά σε εξέλιξη. Έτσι, βελτιώνεται συνολικά η ανθεκτικότητα των οργανισμών, καθώς αποκτούν τη δυνατότητα, όχι μόνο να ανακάμπτουν από επιθέσεις, αλλά και να προσαρμόζουν τις μεθόδους τους σύμφωνα με τις εξελισσόμενες τακτικές των επιτιθέμενων.

Κύρια Συνεισφορά:

Η κύρια συνεισφορά των αντιδραστικών εργαλείων ανάκαμψης για έναν οργανισμό έγκειται στην ικανότητά τους να περιορίζουν τη ζημιά, να εξασφαλίζουν τη συνέχεια των επιχειρησιακών λειτουργιών και να διασφαλίζουν την ακεραιότητα των δεδομένων μετά από μια επίθεση. Μέσω άμεσης αποκατάστασης συστημάτων, απομόνωσης απειλών και δυνατότητας γρήγορης επιστροφής σε κανονική λειτουργία, τα εργαλεία αυτά ελαχιστοποιούν τις οικονομικές απώλειες και το χρόνο διακοπής λειτουργίας που μπορεί να προκαλέσει μια επίθεση. Η ενσωμάτωση αλγορίθμων, όπως Reinforcement Learning και GANs, παρέχει τη δυνατότητα δημιουργίας ρεαλιστικών σεναρίων για την εκπαίδευση συστημάτων και την προετοιμασία έναντι άγνωστων επιθέσεων. Παράλληλα, εργαλεία, όπως το Cortex XDR, αξιοποιούν τεχνολογίες ανίχνευσης ανωμαλιών και κατηγοριοποίησης κινδύνων για την αναγνώριση και την απομόνωση κακόβουλων δραστηριοτήτων σε πραγματικό χρόνο. Επιπλέον, η παρακολούθηση της ακεραιότητας των δεδομένων μέσω μηχανικής μάθησης επιτρέπει την αποκατάσταση μόνο όταν είναι απαραίτητο, αποφεύγοντας περιττές διαδικασίες. Για παράδειγμα, σε περιπτώσεις διαρροής δεδομένων χωρίς τροποποίηση ή διαγραφή, η αποκατάσταση δεν είναι απαραίτητη, γεγονός που μειώνει το λειτουργικό κόστος και τον χρόνο απόκρισης. Το IBM Resilient και το DRaaS ενσωματώνουν προχωρημένες δυνατότητες ανάλυσης περιστατικών και δημιουργίας σχεδίων ανάκαμψης, εστιάζοντας στη συνεχή παρακολούθηση και στη βελτιστοποίηση των στρατηγικών αποκατάστασης. Συνολικά, τα

αντίμετρα αυτά παρέχουν ένα πολυεπίπεδο σύστημα ανάκαμψης που ενισχύει την ανθεκτικότητα των οργανισμών, διασφαλίζοντας όχι μόνο την αποκατάσταση από επιθέσεις αλλά και τη συνεχή βελτίωση των διαδικασιών για την αντιμετώπιση εξελισσόμενων απειλών. Με την ενσωμάτωση μηχανισμών παρακολούθησης της ακεραιότητας και τη χρήση έξυπνων συστημάτων ανάκαμψης, οι οργανισμοί αποκτούν τη δυνατότητα να αντιδρούν στοχευμένα και αποδοτικά σε επιθέσεις κοινωνικής μηχανικής.

Προς την ολοκληρωμένη διατύπωση της συνεισφοράς κάθε εργαλείου, σημαντικό είναι να αναφερθούν και τα αρνητικά τους. Έτσι αρχικά, βάση του άρθρου **[99]**, η υπηρεσία Disaster-Recovery-as-a-Service (**DRaaS**) σημειώνει και κάποια μειονεκτήματα παρά τα πλεονεκτήματά της. Ένας σημαντικός περιορισμός είναι το κόστος που συνδέεται με την υλοποίησή της, το οποίο μπορεί να είναι απαγορευτικά υψηλό για τις μικρές και μεσαίες επιχειρήσεις. Αυτό οφείλεται κυρίως σε περιττές δαπάνες που μπορεί να προκύψουν από λανθασμένο σχεδιασμό και πρόβλεψη απαιτήσεων του συστήματος, όπως η επιλογή υπηρεσιών που υπερβαίνουν τις πραγματικές ανάγκες της επιχείρησης. Από την άλλη πλευρά, οι λύσεις που δεν είναι επαρκώς μελετημένες ενέχουν τον κίνδυνο απώλειας δεδομένων ή μη τήρησης των συμφωνιών επιπέδου υπηρεσιών (SLA). Το DRaaS απαιτεί επίσης συνεχείς ενημερώσεις και αναδιαμόρφωση για την ευθυγράμμιση με τις εξελισσόμενες επιχειρηματικές ανάγκες, τις τεχνολογίες και τον αυξανόμενο όγκο δεδομένων. Επιπλέον, ενώ το μοντέλο βασίζεται σε μεγάλο βαθμό σε υποδομές υπολογιστικού νέφους, οποιαδήποτε διακοπή στις υπηρεσίες νέφους ή στη συνδεσιμότητα μπορεί να επηρεάσει σοβαρά την αποτελεσματικότητά του.

Ένας από τους βασικούς περιορισμούς του **Cortex XDR**, βάση του άρθρου **[49]**, είναι η εξάρτησή του από μια αρχιτεκτονική βασισμένη στο νέφος, η οποία μπορεί να δημιουργήσει προκλήσεις για οργανισμούς με περιορισμένη πρόσβαση στο διαδίκτυο. Επιπλέον, οι πολύπλοκες απαιτήσεις διαμόρφωσης και ενσωμάτωσης του συστήματος μπορεί να απαιτούν σημαντική τεχνογνωσία, καθιστώντας το λιγότερο κατάλληλο για οργανισμούς χωρίς εξειδικευμένες ομάδες ασφάλειας πληροφορικής. Αυτός ο παράγοντας, σε συνδυασμό με την υψηλή τιμολόγησή του, ενδέχεται να περιορίζουν την προσβασιμότητα για τις μικρότερες επιχειρήσεις ή εκείνες με περιορισμένο προϋπολογισμό.

Όσον αφορά το εργαλείο **IBM Resilient Incident Response Platform,** ο χρόνος απόκρισης υποστήριξης αποτελεί ένα από τα ρίσκα στην εφαρμογή του. Συγκεκριμένα έχουν εντοπιστεί δυσκολίες σχετικά με την ανταπόκριση της ομάδας υποστήριξης της IBM, υποδεικνύοντας ότι απαιτούνται βελτιώσεις στις υπηρεσίες υποστήριξης για την άμεση αντιμετώπιση των προβλημάτων των οργανισμών. Εκτός αυτού έχει παρατηρηθεί δυσχέρεια ενσωμάτωσης του IBM Resilient με εφαρμογές και εργαλεία τρίτων κατασκευαστών. Η πλατφόρμα φαίνεται πολλές φορές να απαιτεί πρόσθετες δέσμες ενεργειών ή ενδιάμεσους διακομιστές για τη διευκόλυνση αυτών των ενσωματώσεων, γεγονός που μπορεί να περιπλέξει την ανάπτυξη και τη συντήρησή του.

Τέλος, οι πλατφόρμες **SOAR [101]** αντίστοιχα παρουσιάζουν ορισμένα μειονεκτήματα και κινδύνους. Ένα βασικό μειονέκτημα είναι το υψηλό κόστος υλοποίησης, καθώς απαιτείται σημαντική επένδυση σε υλικοτεχνική υποδομή, λογισμικό και συνεχή συντήρηση. Επιπλέον, οι πλατφόρμες αυτές απαιτούν εκτεταμένη προσαρμογή ώστε να ανταποκρίνονται στις ανάγκες και τις διαδικασίες ενός οργανισμού, γεγονός που μπορεί να αυξήσει την πολυπλοκότητα της διαχείρισής τους. Ένας άλλος κίνδυνος είναι η εξάρτηση από αυτοματοποιημένες αποφάσεις, οι οποίες, αν είναι εσφαλμένες, μπορεί να οδηγήσουν σε λανθασμένες αποκρίσεις, όπως η αχρείαστη απομόνωση συστημάτων ή ο αποκλεισμός κρίσιμων διαδικασιών. Συνολικά, ενώ οι πλατφόρμες SOAR προσφέρουν σημαντικά οφέλη, τα μειονεκτήματά τους πρέπει να αξιολογούνται προσεκτικά πριν την υιοθέτησή τους, κάτι που παρατηρείται φυσικά στην χρήση όλων των εργαλείων ανάκαμψης.

4.4 Σύγκριση των Αντίμετρων

Στην Ενότητα 4.2, εξετάσαμε την απόδοση των τεχνικών ΤΝ σε συγκεκριμένα πλαίσια και περιπτώσεις. Σε αυτήν την ενότητα, θα αναλύσουμε συγκριτικά τα αντίμετρα που χρησιμοποιούν αυτές τις τεχνικές (ΤΝ) ώστε να προκύψει μια πιο ολοκληρωμένη εικόνα, λαμβάνοντας υπόψη όλες τις (δυνατές) παραμέτρους. Η σύγκριση των αντίμετρων επικεντρώνεται σε ένα μεγαλύτερο πλαίσιο, όπου ο ρόλος της ΤΝ είναι μόνο ένα μέρος της συνολικής στρατηγικής αντιμετώπισης των επιθέσεων. Ουσιαστικά, ένα αντίμετρο είναι μια ολιστική προσέγγιση, η οποία μπορεί να περιλαμβάνει διάφορα στάδια, όπως την προεπεξεργασία των δεδομένων και το φιλτράρισμά τους, ώστε η χρησιμοποιούμενη τεχνική ΤΝ να λειτουργεί βέλτιστα καθώς και αρμονικά σε σχέση με άλλες τεχνολογίες και μεθόδους προστασίας που γίνονται εκμεταλλεύσιμες από το αντίμετρο αυτό.

Στους παρακάτω πίνακες συγκρίνονται τα παραπάνω αντίμετρα, βάσει των ακόλουθων κριτηρίων:

- **Οι Αλγόριθμοι ΤΝ που Περιλαμβάνονται:** Επεξηγεί ποιοι αλγόριθμοι Τεχνητής Νοημοσύνης χρησιμοποιούνται μέσα στο αντίμετρο για την ανίχνευση, πρόληψη ή αντιμετώπιση των επιθέσεων. Προφανώς, δεν μας ενδιαφέρει πάντοτε η ποσότητα των αλγορίθμων αλλά η ποιότητα, δηλ. η χρήση των καλύτερων αλγορίθμων εκεί που πραγματικά χρειάζονται. Όπως φάνηκε στην Ενότητα 4.2, σε κάθε είδος επιθέσεων, συγκεκριμένοι αλγόριθμοι ξεχωρίζουν, οπότε θα είναι θετικό για ένα αντίμετρο να τους υιοθετεί προς αυτό τον σκοπό.

- **Οι Επιθέσεις που Αντιμετωπίζονται:** Εστιάζει στις επιθέσεις κοινωνικής μηχανικής που το αντίμετρο είναι σχεδιασμένο να εντοπίσει ή να αποτρέψει.
- **Τα Θετικά του Αντίμετρου:** Παρουσιάζει τα πλεονεκτήματα που προσφέρει το εκάστοτε αντίμετρο.
- **Τα Αρνητικά του Αντίμετρου:** Περιγράφει τους περιορισμούς ή τα προβλήματα που μπορεί να προκύψουν από την εφαρμογή του αντίμετρου.
- **Πολυπλοκότητα Αντιμέτρου:** Επικεντρώνεται στο επίπεδο πολυπλοκότητας των αλγορίθμων και μεθόδων που χρησιμοποιεί το αντίμετρο, ως προς το επίπεδο δυσκολίας εφαρμογής και διαχείρισής τους.
- **Επιπλέον Τεχνικές/Μηχανισμοί Ασφάλειας:** Αναφέρεται σε πρόσθετες τεχνικές (όπως η ανίχνευση ανωμαλιών, η κρυπτογράφηση, και οι μηχανισμοί πολιτικής πρόσβασης) που ενσωματώνονται για να ενισχύσουν το προσφερόμενο επίπεδο ασφάλειας του αντιμέτρου.

Όσον αφορά την δομή της συγκεκριμένης ενότητας, θα αναδειχθούν τρεις ξεχωριστοί συγκριτικοι πίνακες, ανά κατηγορία των αντιμέτρων που αναλύθηκαν, χρησιμοποιώντας το προαναφερόμενο σύνολο κριτηρίων. Έτσι, η ενότητα ξεκινάει με την υπο-ενότητα 4.4.1 στην οποία συγκρίνονται τα αντίστοιχα αντίμετρα που αφορούν την πρόληψη των επιθέσεων ΚΜ. Έπειτα, προχωράμε στην υπο-ενότητα 4.4.2 που συγκρίνονται τα αντίμετρα που εστιάζουν στην αντιμετώπιση επιθέσεων ΚΜ σε πραγματικό χρόνο. Παρομοίως, στην υπο-ενότητα 4.4.3, παρουσιάζονται σε μορφή συγκριτικού πίνακα τα αντίμετρα ανάκαμψης. Τέλος, η υπο-ενότητα 4.4.4 παρέχει τα καθολικά αποτελέσματα που εξήχθησαν, ανεξάρτητα από την εκάστοτε κατηγορία των αντιμέτρων. Έτσι, συνολικά από την ανάλυση αυτή, προκύπτουν παρατηρήσεις και συμπεράσματα που ισχύουν τόσο σε επίπεδο τύπου αντιμετώπισης των επιθέσεων όσο και σε καθολικό επίπεδο.

4.4.1 Συγκριτική Ανάλυση Προληπτικών Αντίμετρων

Πίνακας 7: Αποτελέσματα Σύγκρισης Προληπτικών Αντιμέτρων.

Αντίμετρο	Αλγόριθμοι ΤΝ	Επιθέσεις που Αντιμετωπίζονται	Θετικά Αντίμετρου	Αρνητικά Αντίμετρου	Πολυπλοκότητα Αντιμέτρου	Επιπλέον Τεχνικές/Μηχανισμοί Ασφάλειας

ISO/IEC 27001 Protocol	Δεν περιλαμβάνει αλγορίθμους ΤΝ, οι οργανισμοί που το εφαρμόζουν όμως μπορούν να τους ενσωματώσουν.	Δεν εστιάζει σε συγκεκριμένες επιθέσεις ΚΜ, αλλά προάγει γενικές πρακτικές ασφάλειας	- Διεθνής αναγνώριση - Παραμετροποιήσιμο πλαίσιο - Ευκολία ενσωμάτωσης - Διαχείριση κινδύνων -Εκπαίδευση προσωπικού	- Σημαντικές επενδύσεις σε χρόνο & χρήματα, που ενδέχεται να μην αποφέρουν άμεση προστασία - Εξαρτώμενο από τη συμμόρφωση εργαζομένων	- Κλιμακούμενο για διαφορετικά μεγέθη - Σαφής τεκμηρίωση που αυξάνει την πολυπλοκότητα εξαιτίας της ανάγκης για εξειδικευμένη γνώση	- Αρχιτεκτονική Μηδενικής Εμπιστοσύνης (Zero Trust) - Ανάλυση Συμπεριφοράς - Μηχανισμοί Ελέγχου Πρόσβασης - Κρυπτογράφηση Δεδομένων
Vectra AI	K-Means Clustering, Logistic Regression	- Phishing - Spear Phishing - Credential Harvesting	- Αυτόματη ανάλυση απειλών - Οπτικοποίηση και αναφορές απειλών - Προσαρμογή σε νέα μοτίβα επιθέσεων	- Υψηλό κόστος υλοποίησης & συντήρησης - Εξάρτηση από ιστορικά δεδομένα - Πιθανότητα ψευδώς θετικών ειδοποιήσεων	- Βασίζεται σε εξειδικευμένα μοντέλα ΜΜ	- Αυτόματη Κατηγοριοποίηση Απειλών - Διαχείριση Συμβάντων
Darktrace	K-Means Clustering, Random Forest, SVM	Phishing, Spear Phishing, Watering Hole,	- Δημιουργία βασικών συμπεριφορών για χρήστες και δίκτυα	- Περιορισμένη διαφάνεια λειτουργίας - Περιορισμένες	- Υβριδική Αρχιτεκτονική Μοντέλου (συνδυάζει μη επιβλεπόμενη και	- Διαχείριση Πρόσβασης και Τμηματοποίηση Δικτύου - Παρακολούθηση και Οπτικοποίηση

		Credential Harvesting	- Αυτόνομη ανίχνευση και απόκριση σε απειλές	δυνατότητες προσαρμογής - Πολυπλοκότητα εγκατάστασης - Υψηλό κόστος	εποπτευόμενη μάθηση)	Δεδομένων (Threat Visualizer)
Cisco Umbrella	Logistic Regression, KNN, Random Forest	Phishing, Credential Harvesting	- Αποκλείει κακόβουλους τομείς και διευθύνσεις URL πριν φτάσουν στους χρήστες - Εύκολη ενσωμάτωση σε υπάρχουσες υποδομές	- Περιορισμένη προστασία εκτός σύνδεσης - Μείωση ταχύτητας διαδικτύου - Εστίαση μόνο στο επίπεδο DNS - Αποτελεσματικότητα εξαρτώμενη από τη συμβατότητα με υπάρχουσες υποδομές	- Απλότητα για τον Τελικό Χρήστη - Κλιμακούμενη Αρχιτεκτονική Cloud (Βασίζεται σε παγκόσμια υποδομή cloud, με λειτουργίες όπως DNS-layer security, secure web gateway, cloud-delivered firewall	- Πολυπαραγοντική αυθεντικοποίηση (MFA) - Φιλτράρισμα τομέων

Συμπεράσματα:

Ο πίνακας προληπτικών αντιμέτρων περιλαμβάνει εργαλεία που επικεντρώνονται στην αποτροπή επιθέσεων πριν αυτές εκδηλωθούν πλήρως. Συγκρίνοντας τα τέσσερα προληπτικά εργαλεία (**ISO/IEC 27001 Protocol, Vectra AI, Darktrace, Cisco Umbrella**) προκύπτουν σημαντικές διαφορές και ομοιότητες με βάση τα κριτήρια των ειδών επιθέσεων, της πολυπλοκότητας του αντιμέτρου, των θετικών και αρνητικών τους χαρακτηριστικών, καθώς και των επιπλέον τεχνικών/μηχανισμών ασφάλειας. Το **ISO/IEC 27001 Protocol** αποτελεί εξαιρετική επιλογή για οργανισμούς που επιδιώκουν μια γενική στρατηγική πρόληψης και ασφάλειας, εστιάζοντας στη δομή πολιτικών και διαδικασιών, χωρίς να ενσωματώνει άμεσα αλγορίθμους ΤΝ. Το **Vectra AI** και το **Darktrace** ξεχωρίζουν για την αντιμετώπιση εξελιγμένων απειλών, όπως το **phishing**, το **spear phishing** και το **credential harvesting**, με το Darktrace να υπερέχει στην κάλυψη, αλλά με μεγαλύτερη πολυπλοκότητα και κόστος. Το **Cisco Umbrella**, από την άλλη, είναι ιδανικό για μικρομεσαίες επιχειρήσεις που χρειάζονται άμεση εφαρμογή και προστασία από βασικές απειλές, όπως το phishing, ενώ η αρχιτεκτονική του στο cloud μειώνει την πολυπλοκότητα. Συνολικά, η επιλογή εξαρτάται από τις ανάγκες του οργανισμού, με το **ISO/IEC 27001** να εστιάζει στη μακροπρόθεσμη δομή, το **Vectra AI** και το **Darktrace** να προσφέρουν εξελιγμένες δυνατότητες, και το **Cisco Umbrella** να αποτελεί μια πιο προσιτή και ευέλικτη λύση.

4.4.2 Συγκριτική Ανάλυση Αντίμετρων Ανίχνευσης Σε Πραγματικό Χρόνο

Πίνακας 8: Αποτελέσματα Σύγκρισης Αντιμέτρων Ανίχνευσης σε Πραγματικό Χρόνο.

Αντίμετρο	Αλγόριθμοι ΤΝ	Επιθέσεις προς Αντιμετώπιση	Θετικά Αντίμετρου	Αρνητικά Αντίμετρου	Πολυπλοκότητα Αντιμέτρου	Επιπλέον Τεχνικές/Μηχανισμοί Ασφάλειας
Splunk Enterprise Security	Neural Networks, AdaBoost, GMM	Phishing, Spear Phishing	- Ευελιξία στην επιλογή αλγορίθμων - Προσαρμοστικότητα σε νέες απειλές	- Υψηλή πολυπλοκότητα εγκατάστασης και διαχείρισης - Υψηλό κόστος	- Ενσωματώνει πολλαπλούς αλγορίθμους ML (MLTK εργαλειοθήκη) - Ενσωμάτωση ποικίλων πηγών	- Ειδοποιήσεις σε πραγματικό χρόνο

			- Ολοκληρωμένη παρακολούθηση απειλών	αδειοδότησης - Μειωμένο επίπεδο κλιμάκωσης	δεδομένων (εκτεταμένη συσχέτιση ετερογενών δεδομένων)	- Οπτικοποίηση δεδομένων
IBM QRadar SIEM	Isolation Forest, MDP,	Phishing, Spear Phishing, Anomalous Behavior	- Ισχυρές δυνατότητες ανάλυσης συμπεριφοράς - Ενσωμάτωση Watson NLP (που παρέχει προηγμένες αναλύσεις)	- Υψηλό κόστος υλοποίησης και συντήρησης - Πολύπλοκη διεπαφή και λειτουργίες που ενδέχεται να απαιτούν εκπαίδευση - καθυστερημένες ενημερώσεις συμβατότητας	- Ενσωματώνει Watson NLP και σύνθετα ML μοντέλα - Η προηγμένη επεξεργασία φυσικής γλώσσας και η μοντελοποίηση δεδομένων αυξάνουν την πολυπλοκότητα του συστήματος	- Χρήση User Behavior Analytics (UBA) - Συσχέτιση περιστατικών - Ιεράρχηση βάσει κινδύνου (Ταξινομεί αυτόματα τις απειλές με βάση τον αντίκτυπο και την πιθανότητά τους)
McAfee Enterprise Security Manager (ESM)	Random Forest, Gradient Boosting, RNN	Phishing, Credential , Harvesting, Anomalous Behavior	- Εξαιρετική προσαρμοστικότητα σε διαφορετικά σενάρια	- Υψηλός χρόνος εγκατάστασης	- Συνδυασμός σχετικά σύνθετων επιμέρους αλγορίθμων, που αυξάνουν	- Χρήση User Behavior Analytics (UBA)

			(Δυνατότητες ενσωμάτωσης) - Προηγμένη Πληροφορία Απειλών	- Μεγάλη χρήση πόρων υποδομής - Υπερφόρτωση δεδομένων	την συνολική πολυπλοκότητα	- Αυτοματοποίηση περιστατικών(Επιταχύνει την απόκριση με προκαθορισμένες ροές εργασίας)
CrowdStrike Falcon	Neural Networks, Decision Trees, RNN	Spear Phishing, Anomalous Behavior	- Ισχυρή προστασία τελικών σημείων (EDR) - Χρήση συλλογικής νοημοσύνης - Αρχιτεκτονική βασισμένη στο νέφος	- Εστίαση μόνο στα τελικά σημεία (endpoint-focused) - Εξαρτάται από τη συνδεσιμότητα στο Διαδίκτυο - Υψηλό κόστος για προηγμένα χαρακτηριστικά - Απαίτηση για ειδική εμπειρογνωμοσύνη	- Γράφημα απειλών που παρέχεται στο νέφος (Αναλύει δισεκατομμύρια καθημερινά συμβάντα, απαιτώντας σημαντική υπολογιστική ισχύ)	- Προσφέρεται προληπτική ικανότητα, λόγω ανίχνευσης ανώμαλης συμπεριφοράς - Threat Graph Intelligence (Μοιράζεται πληροφορίες σε όλους τους πελάτες για συλλογική άμυνα)

Συμπεράσματα:

Η σύγκριση των εργαλείων ανίχνευσης σε πραγματικό χρόνο αναδεικνύει διαφορετικές προσεγγίσεις και δυνατότητες, καθιστώντας την επιλογή κατάλληλου εργαλείου εξαρτώμενη από τις ανάγκες του οργανισμού. Το **Splunk Enterprise Security** ξεχωρίζει για την ευελιξία και την προσαρμοστικότητά του, αξιοποιώντας αλγορίθμους όπως Neural Networks, AdaBoost, και GMM, αλλά η πολυπλοκότητα εγκατάστασης και το υψηλό κόστος αποτελούν σημαντικά εμπόδια. Παράλληλα, ενσωματώνει τεχνικές όπως ειδοποιήσεις σε πραγματικό χρόνο και οπτικοποίηση δεδομένων, προσφέροντας ολοκληρωμένη παρακολούθηση απειλών.

Το **IBM QRadar SIEM** εστιάζει στην ανάλυση συμπεριφοράς μέσω αλγορίθμων, όπως Isolation Forest και MDP, ενώ η ενσωμάτωση του Watson NLP προσδίδει προηγμένες δυνατότητες ανάλυσης. Ωστόσο, το υψηλό κόστος και η απαιτούμενη εκπαίδευση για τη χρήση του αποτελούν προκλήσεις. Παρέχει, ωστόσο, μοναδικές λειτουργίες, όπως η συσχέτιση περιστατικών και η ταξινόμηση απειλών βάσει κινδύνου.

Το **McAfee Enterprise Security Manager (ESM)** ξεχωρίζει για την προσαρμοστικότητά του σε διαφορετικά σενάρια, χρησιμοποιώντας αλγόριθμους όπως Random Forest, Gradient Boosting και RNN. Παρά τα πλεονεκτήματα της αυτοματοποίησης απόκρισης σε περιστατικά και της χρήσης πληροφοριών απειλών, ο υψηλός χρόνος εγκατάστασης και η μεγάλη κατανάλωση πόρων αποτελούν μειονεκτήματα.

Το **CrowdStrike Falcon** επικεντρώνεται στην προστασία τελικών σημείων μέσω Neural Networks, Decision Trees και RNN, με υψηλή αποτελεσματικότητα στην ανίχνευση ανώμαλης συμπεριφοράς. Παρά την ισχυρή αρχιτεκτονική νέφους και τη χρήση συλλογικής νοημοσύνης, εξαρτάται από τη συνδεσιμότητα στο Διαδίκτυο και είναι δαπανηρό για προηγμένα χαρακτηριστικά.

Συνολικά, τα εργαλεία αυτά καλύπτουν διαφορετικές ανάγκες: το Splunk ES παρέχει ευελιξία και προσαρμογή, το IBM QRadar SIEM ισχυρή ανάλυση συμπεριφοράς, το McAfee ESM προσαρμοστικότητα και αυτοματοποίηση, και το CrowdStrike Falcon εξειδίκευση στα τελικά σημεία. Η επιλογή εξαρτάται από την κλίμακα του οργανισμού, τον προϋπολογισμό και τις τεχνικές απαιτήσεις.

4.4.3 Συγκριτική Ανάλυση Αντιδραστικών Αντίμετρων

Πίνακας 9: Αποτελέσματα Σύγκρισης Αντιδραστικών Αντιμέτρων.

Αντίμετρο	Αλγόριθμοι ΤΝ	Επιθέσεις που Αντιμετωπίζονται	Θετικά Αντίμετρου	Αρνητικά Αντίμετρου	Πολυπλοκότητα Αντιμέτρου	Επιπλέον Τεχνικές/Μηχανισμοί Ασφάλειας
Cortex XDR (Palo Alto Networks)	Unsupervised Learning, Supervised Learning	Ransomware, Credential Harvesting	- Παρέχει root cause analysis, αποκλεισμό κακόβουλου λογισμικού, πλήρη αποκατάσταση - Εξαιρετική αναγνώριση κακόβουλων μοτίβων μέσω συλλογικής νοημοσύνης	- Η ενσωμάτωσή του σε υπάρχουσες υποδομές μπορεί να απαιτεί σημαντικό κόστος και χρόνο - Περιορισμοί συνδεσιμότητας - Απαίτηση για σημαντική τεχνογνωσία	- Ενσωματώνει συνδυασμό supervised και unsupervised learning για την προσαρμογή σε νέα μοτίβα	- Διαχείριση endpoint, συνεργασία με SIEM για αναλυτική αξιολόγηση κινδύνων - Κρυπτογράφηση δεδομένων - Μηχανισμοί ελέγχου πρόσβασης
IBM Resilient Incident Response Platform	NLP + ML (Watson), Unsupervised Learning, Supervised Learning	Phishing, credential harvesting	- Αυτοματοποιημένες προτάσεις αποκατάστασης	- Απαίτηση εξειδικευμένου προσωπικού για την πλήρη αξιοποίηση των	- Το Watson ενσωματώνει NLP και ML, αυξάνοντας τη σύνθετη	- Αυτοματοποιημένες προτάσεις αποκατάστασης & σχέδια αντιμετώπισης περιστατικών (Εντοπίζει και αξιολογεί τρωτά

			- Δημιουργία σχεδίων ανάκαμψης - Ικανότητα κατανόησης και ανάλυσης σύνθετων περιστατικών μέσω του Watson.	δυνατοτήτων του - Μη κατάλληλος χρόνος απόκρισης υποστήριξης - Δυσχέρεια ενσωμάτωσης με εργαλεία τρίτων	λειτουργικότητα. - Περίπλοκη ανάπτυξη & συντήρηση	σημεία στα συστήματα) - Συσχέτιση δεδομένων
Automation and Orchestration Platforms (SOAR)	Isolation Forest, Random Forest, SVM, Reinforcement Learning	Phishing, Vishing	- Αυτόματη απομόνωση επηρεασμένων συσκευών. - Συνεργασία με συστήματα αντιγράφων ασφαλείας για ταχεία αποκατάσταση - Ενσωμάτωση με ML για την ταξινόμηση και ανάλυση απειλών.	- Περιορισμένες δυνατότητες εκτός των τελικών σημείων - Μπορεί να απαιτεί επιπλέον εργαλεία για την ολοκληρωμένη κάλυψη - Απαιτεί συνεχή συντήρηση - Υψηλό κόστος υλοποίησης - Εξάρτηση από αυτοματοποι	- Σύνθετες ροές εργασίας (απαιτεί τον σχεδιασμό και τη διαχείριση πολύπλοκων διαδικασιών)	- Συνεργασία με backup για ταχεία αποκατάσταση, απομόνωση επηρεασμένων συσκευών - Ανάλυση απειλών (παρέχει πληροφορίες για την κατανόηση και αντιμετώπιση των απειλών)

				ημένες αποφάσεις		
Disaster Recovery as a Service (DRaaS)	GANs (για σενάρια επίθεσης), MSPAD-SGM (πολυ-σταδιακό παιχνίδι ανάλυσης επιθέσεων)	Ransomware, Phishing	- Απρόσκοπτη ανάκτηση δεδομένων και συστημάτων μέσω cloud - Δυνατότητα προσομοίωσης επιθέσεων για εκπαιδευτικούς σκοπούς - Βελτιστοποίηση στρατηγικής ανίχνευσης και απόκρισης μέσω MSPAD-SGM.	- Εξαρτάται από το υπολ. νέφος, το οποίο μπορεί να προκαλέσει ανησυχίες σχετικά με την ιδιωτικότητα και την ασφάλεια δεδομένων - Θέματα καθυστέρησης (η απόσταση από το κέντρο δεδομένων του παρόχου μπορεί να επηρεάσει τους χρόνους αποκατάστασης) - Υψηλό κόστος υλοποίησης - Απαιτήσεις για συνεχείς ενημερώσεις	- Ενσωματώνει GANs για προσομοίωση επιθέσεων και MSPAD-SGM για στρατηγική ανάλυση, παρόλα αυτά η ακριβής πολυπλοκότητα και των δύο τεχνικών εξαρτάται από τον σχεδιασμό και την εφαρμογή τους.	- Υπηρεσίες ανάκτησης δεδομένων - Προσομοιώσεις επιθέσεων για εκπαιδευτικούς σκοπούς - Βελτιστοποίηση στρατηγικής ανίχνευσης και απόκρισης - Γεωγραφική αναπαραγωγή δεδομένων (αποθήκευση αντιγράφων ασφαλείας σε διαφορετικές γεωγραφικές τοποθεσίες για αυξημένη ανθεκτικότητα)

				& αναδιαμόρφωση		

Συμπεράσματα:

Τα αντιδραστικά μέτρα επικεντρώνονται στην ανάκαμψη μετά από μια επίθεση. Η ανάλυση των αντιμέτρων αποκατάστασης καταδεικνύει τη σημασία της ενσωμάτωσης τεχνητής νοημοσύνης και προηγμένων τεχνικών για την αποκατάσταση. Όταν συγκρίνουμε τις τεχνολογίες **Cortex XDR, IBM Resilient Incident Response Platform**, **SOAR** και **DRaaS**, μπορούμε να παρατηρήσουμε ότι η καθεμία έχει τα δικά της πλεονεκτήματα και δυσκολίες. Το **Cortex XDR** αντιμετωπίζει επιθέσεις όπως το ransomware και η συλλογή διαπιστευτηρίων συνδυάζοντας επιβλεπόμενη και μη επιβλεπόμενη μάθηση, καθώς και προσφέροντας ανάλυση της αιτίας και προστασία από κακόβουλο λογισμικό. Ωστόσο, η ενσωμάτωσή του μπορεί να συνεπάγεται σημαντικό κόστος και χρόνο. Η πλατφόρμα **IBM Resilient Incident Response Platform** χρησιμοποιεί το Watson για την ανάλυση φυσικής γλώσσας και τη μηχανική μάθηση, παρέχοντας αυτοματοποιημένες συστάσεις αποκατάστασης και δημιουργία σχεδίου αποκατάστασης, αλλά οι πλήρεις δυνατότητές της απαιτούν εξειδικευμένο προσωπικό. Οι πλατφόρμες **SOAR** καταπολεμούν τις επιθέσεις phishing με αλγορίθμους, όπως οι Isolation Forest και Random Forest, οι οποίοι παρέχουν αυτόματη απομόνωση των επηρεαζόμενων συσκευών, ωστόσο ενδέχεται να απαιτούνται πρόσθετα εργαλεία για συνολική κάλυψη. Τέλος, το **DRaaS** εστιάζει στην απρόσκοπτη ανάκτηση δεδομένων μέσω cloud, αντιμετωπίζοντας επιθέσεις ransomware, αλλά εξαρτάται από το υπολογιστικό νέφος, με πιθανά θέματα καθυστέρησης. Συνολικά, η επιλογή του κατάλληλου εργαλείου εξαρτάται από τις συγκεκριμένες ανάγκες και τις δυνατότητες του οργανισμού, λαμβάνοντας υπόψη την πολυπλοκότητα των μοντέλων και τις επιπλέον τεχνικές ασφάλειας που προσφέρει το κάθε εργαλείο.

4.4.4 Καθολικά Αποτελέσματα Ανάλυσης Αντίμετρων

Σε συνολικό επίπεδο, τα τρία είδη αντιμέτρων καλύπτουν ένα ευρύ φάσμα επιθέσεων κοινωνικής μηχανικής, με τα προληπτικά μέτρα να επικεντρώνονται στην αποτροπή, τα μέτρα πραγματικού χρόνου στην άμεση ανίχνευση και τα αντιδραστικά στην ανάκαμψη. Τα προληπτικά, τα αντίμετρα σε πραγματικό χρόνο και τα αντιδραστικά αντίμετρα παρουσιάζουν διαφορετικά επίπεδα κάλυψης απέναντι στις επιθέσεις κοινωνικής μηχανικής, με κάθε κατηγορία να επικεντρώνεται σε διαφορετικές φάσεις των απειλών. Τα προληπτικά μέτρα προσφέρουν ισχυρή κάλυψη σε επιθέσεις, όπως το phishing και το credential harvesting, με την έμφαση στην πρόληψη να μειώνει τη συχνότητα εμφάνισης επιθέσεων. Ωστόσο, το μειονέκτημά τους έγκειται στη συχνά περιορισμένη αποτελεσματικότητα απέναντι σε εξελιγμένες και άγνωστες τακτικές, οι οποίες παρακάμπτουν τα αρχικά φίλτρα. Τα μέτρα σε πραγματικό χρόνο υπερέχουν στην ανίχνευση ενεργών επιθέσεων, όπως spear phishing και DDoS, ενσωματώνοντας προηγμένους αλγορίθμους ΤΝ, όπως Neural Networks και Isolation Forest, για τη δυναμική ανίχνευση ανωμαλιών. Παρά τα υψηλά επίπεδα ακρίβειας, τα αντίμετρα αυτά είναι συχνά πιο πολύπλοκα και απαιτούν αυξημένη υποδομή. Στον αντίποδα, τα αντιδραστικά μέτρα διακρίνονται για την αποτελεσματική ανάκαμψη από επιθέσεις, όπως ransomware ή διαρροές δεδομένων, μέσω της χρήσης στρατηγικών αποκατάστασης που βασίζονται σε Reinforcement Learning και GANs. Τέλος, εργαλεία όπως το Darktrace μπορούν να ανήκουν σε περισσότερες από μία κατηγορίες, καλύπτοντας τόσο προληπτικές όσο και ανιχνευτικές λειτουργίες κάτι που μπορεί να θεωρηθεί μεγάλο πλεονέκτημα. Παρόλα αυτά, και πάλι η αποτελεσματικότητά τους εξαρτάται από την προσαρμογή τους σε συγκεκριμένα σενάρια.

Αξιολογώντας τα αντίμετρα σφαιρικά, τα μέτρα σε πραγματικό χρόνο φαίνεται να υπερέχουν σε κάλυψη επιθέσεων, ενώ τα προληπτικά μέτρα είναι πιο ευέλικτα. Τα αντιδραστικά μέτρα, αν και περιορισμένα στην προληπτική τους ικανότητα, είναι αναντικατάστατα για τη διαχείριση κρίσεων. Η συμπληρωματικότητα μεταξύ των κατηγοριών είναι ουσιαστική, καθώς η ολοκληρωμένη ενσωμάτωση και των τριών παρέχει την απαραίτητη πολυεπίπεδη προσέγγιση. Συγκεκριμένα, η χρήση προληπτικών μέτρων μειώνει την έκθεση σε απειλές, τα μέτρα σε πραγματικό χρόνο ανιχνεύουν και περιορίζουν τις επιθέσεις, ενώ τα αντιδραστικά διασφαλίζουν την ανθεκτικότητα και την αποκατάσταση του οργανισμού. Έτσι, δεν μπορούμε να ξεχωρίσουμε ένα και μόνο αντίμετρο ως το πιο αποτελεσματικό συνολικά, καθώς κάθε εργαλείο ανταποκρίνεται καλύτερα σε συγκεκριμένες περιπτώσεις, ανάλογα με τις ανάγκες του οργανισμού, τον τύπο επιθέσεων που αντιμετωπίζει και τις διαθέσιμες υποδομές. Η συνδυαστική χρήση εργαλείων από διαφορετικές κατηγορίες προσφέρει την απαραίτητη πολυεπίπεδη στρατηγική για την ολοκληρωμένη προστασία και ανθεκτικότητα έναντι των εξελισσόμενων απειλών, θωρακίζοντας τα συστήματα.

Κεφάλαιο 5: Προκλήσεις & Μελλοντική Έρευνα

Η ασφάλεια των πληροφοριών και των συστημάτων έχει επεκταθεί σε νέους ορίζοντες με τη χρήση της τεχνητής νοημοσύνης (ΤΝ) όσον αφορά τον εντοπισμό και την αντιμετώπιση των επιθέσεων κοινωνικής μηχανικής. Αν και η ΤΝ έχει αποδειχθεί επιτυχής σε πολλές περιπτώσεις ως προς αυτό τον σκοπό, εξακολουθούν να υπάρχουν ορισμένα εμπόδια και περιορισμοί που δυσχεραίνουν την ικανότητά της να εντοπίζει και αναχαιτίζει αυτές τις απειλές. Αυτά τα εμπόδια θα αναλυθούν στη συνέχεια σε ξεχωριστές ενότητες του κεφαλαίου. Για κάθε τέτοιο εμπόδιο ή πρόκληση, θα παρέχονται επίσης ενδιαφέρουσες ερευνητικές κατευθύνσεις για την αντιμετώπισή του.

5.1 Η Εξάρτηση από την Ανθρώπινη Ψυχολογία

Μία από τις κύριες προκλήσεις για την ανίχνευση και αποτροπή των επιθέσεων κοινωνικής μηχανικής μέσω τεχνητής νοημοσύνης (ΤΝ) είναι η εξάρτηση από την ανθρώπινη ψυχολογία **[58]**. Οι επιθέσεις αυτές συχνά εκμεταλλεύονται τη συναισθηματική χειραγώγηση των θυμάτων, όπως είναι η δημιουργία αίσθησης επείγοντος ή εμπιστοσύνης, καθιστώντας δύσκολη την ακριβή ανίχνευσή τους. Παρόλο που οι σύγχρονες προσεγγίσεις που χρησιμοποιούν εξελιγμένες τεχνικές επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) έχουν σημειώσει κάποια επιτυχία στην κατανόηση και αναγνώριση της γλώσσας συναισθηματικής χειραγώγησης, η κατανόηση των πιο λεπτών ψυχολογικών ενδείξεων παραμένει δύσκολη για τους αλγορίθμους ΤΝ **[59]**.

Μελλοντική Έρευνα:

Η περαιτέρω έρευνα μπορεί να εστιάσει στην ανάπτυξη υβριδικών μοντέλων που συνδυάζουν τεχνικές μηχανικής μάθησης με ψυχολογικά προφίλ των θυμάτων. Αυτά τα μοντέλα θα μπορούσαν να ανιχνεύουν πότε το θύμα ενδέχεται να ενδώσει στην επίθεση, με αποτέλεσμα την αποτροπή της επίθεσης μέσω έγκαιρης ενημέρωσης του θύματος ή ακόμα και της διακοπής της σύνδεσης **[60]**. Μια επιπλέον κατεύθυνση αφορά τη δημιουργία πιο προηγμένων συστημάτων NLP, τα οποία θα μπορούν να αναλύουν τις διακυμάνσεις της συναισθηματικής αντίδρασης του θύματος, σε πραγματικό χρόνο. Αυτά τα συστήματα, σε συνδυασμό με ανατροφοδότηση από τη συμπεριφορά του χρήστη, θα μπορούσαν να προσαρμόζουν δυναμικά τα φίλτρα ασφαλείας, αυξάνοντας το επίπεδο προστασίας κατά τη διάρκεια μιας επίθεσης κοινωνικής μηχανικής. Έτσι, οι αλγόριθμοι ΤΝ θα μπορούσαν να παρέχουν εξατομικευμένη άμυνα, λαμβάνοντας υπόψη την κατάσταση του θύματος και τις συνθήκες υπό

τις οποίες μπορεί να ενδώσει. Η ικανότητα της ΤΝ να κατανοεί όχι μόνο τις λέξεις, αλλά και το συναίσθημα και τη συμπεριφορά πίσω από τα μηνύματα ενώ εξελίσσονται, αποτελεί σημαντικό βήμα προς τη βελτίωση της ανίχνευσης τέτοιων επιθέσεων **[61]**.

5.2 Η Αντιμετώπιση των Μη Ισορροπημένων Δεδομένων & Δεδομένων Χαμηλής Ποιότητας

Μία ακόμα σημαντική πρόκληση αφορά την έλλειψη ισορροπίας στα δεδομένα εκπαίδευσης. Οι επιθέσεις κοινωνικής μηχανικής, όπως το phishing, αποτελούν μικρό ποσοστό του συνολικού όγκου επικοινωνιών (πχ. email), καθιστώντας δύσκολη την εκπαίδευση αλγορίθμων ΤΝ με επαρκή δείγματα. Παρότι το φιλτράρισμα δεδομένων γίνεται ευρέως, η επιλογή και διατήρηση των χαρακτηριστικών που πραγματικά αντιστοιχούν σε επιθέσεις παραμένει δύσκολη. Παράλληλα, δεδομένα χαμηλής ποιότητας, όπως δεδομένα με θόρυβο, ελλείψεις ή μη αντιπροσωπευτικά δείγματα, αυξάνουν την πολυπλοκότητα της εκπαίδευσης και της ανίχνευσης. Τα δεδομένα χαμηλής ποιότητας μπορούν να οδηγήσουν σε μοντέλα ΜΜ με μειωμένη ακρίβεια, ενώ ταυτόχρονα δυσκολεύουν τη διαδικασία εξαγωγής των κρίσιμων χαρακτηριστικών που διακρίνουν τις επιθέσεις από τις φυσιολογικές επικοινωνίες. Μια σημαντική πρόκληση, λοιπόν, είναι η βελτιστοποίηση των τεχνικών φιλτραρίσματος και προεπεξεργασίας, με στόχο την εξαγωγή, την ενίσχυση και την καθαρότητα/ποιότητα των χαρακτηριστικών που διακρίνουν τις επιθέσεις κοινωνικής μηχανικής.

<u>Μελλοντική Έρευνα:</u>

Για την αντιμετώπιση αυτών των προκλήσεων, οι ερευνητές μπορούν να επικεντρωθούν σε τεχνικές ανίχνευσης ανωμαλιών (anomaly detection), όπως Isolation Forest και One-Class SVM, που λειτουργούν καλύτερα με μη ισορροπημένα δεδομένα. Επίσης, η εφαρμογή εξελιγμένων μεθόδων επιλογής χαρακτηριστικών, όπως Recursive Feature Elimination (RFE) **[1]**, μπορεί να βελτιώσει την ακρίβεια και την απόδοση των αλγορίθμων, εστιάζοντας σε καθαρά και αντιπροσωπευτικά στοιχεία των επιθέσεων. Η RFE είναι μια τεχνική που αφαιρεί σταδιακά τα λιγότερο σημαντικά χαρακτηριστικά, διατηρώντας μόνο εκείνα που συμβάλλουν ουσιαστικά στην απόδοση του εκάστοτε αντιμέτρου **[1]**.

Παράλληλα, η χρήση τεχνικών ενίσχυσης δεδομένων (data augmentation), όπως η Random Oversampling/Undersampling, που αυξάνει την παρουσία της μειοψηφικής κλάσης ή μειώνει την παρουσία της κυρίαρχης κλάσης αντίστοιχα **[2]**, μπορεί να προσφέρει μια απλή αλλά αποτελεσματική λύση. Μια πιο εξελιγμένη τεχνική είναι η ADASYN (Adaptive Synthetic Sampling) **[3],** η οποία δημιουργεί συνθετικά δείγματα στις περιοχές όπου η μειοψηφική κλάση δεν εκπροσωπείται επαρκώς. Επιπλέον, τεχνικές, όπως η δημιουργία συνθετικών δειγμάτων μέσω Generative Adversarial Networks (GANs) **[4]** ή Synthetic Minority Over-sampling

Technique (SMOTE) **[2]**, μπορούν να συμβάλουν στην κάλυψη της έλλειψης επιθετικών δεδομένων και στη βελτίωση της ποιότητας των δειγμάτων.

Για την πρόκληση των δεδομένων χαμηλής ποιότητας, είναι απαραίτητο να διερευνηθούν τεχνικές προεπεξεργασίας, όπως η αντιμετώπιση θορύβου μέσω φιλτραρίσματος δεδομένων (noise filtering) **[5]**, η χρήση αλγορίθμων καθαρισμού δεδομένων (data cleaning) **[6]**, και η κανονικοποίηση δεδομένων (data normalization) **[7]**. Επιπλέον, η αντικατάσταση ελλιπών δεδομένων με στατιστικές μεθόδους (imputation) **[8]** μπορεί να ενισχύσει την ποιότητα των δειγμάτων που χρησιμοποιούνται για την εκπαίδευση. Η μελλοντική έρευνα μπορεί να επικεντρωθεί στην ανάπτυξη εξειδικευμένων τεχνικών που να συνδυάζουν ανίχνευση ανωμαλιών και βελτίωση ποιότητας δεδομένων, ώστε να διασφαλιστεί η αποτελεσματική εκπαίδευση αλγορίθμων απέναντι στις κοινωνικές επιθέσεις.

5.3 Το Υψηλό Υπολογιστικό Κόστος

Ένα εξίσου σημαντικό εμπόδιο στην εφαρμογή της ΤΝ είναι το υψηλό υπολογιστικό κόστος. Πολλοί αλγόριθμοι, όπως η βαθιά μάθηση (Deep Learning) και τα Support Vector Machines (SVM), απαιτούν μεγάλη υπολογιστική ισχύ, ιδιαίτερα όταν εφαρμόζονται σε μεγάλα και πολύπλοκα περιβάλλοντα, όπως εταιρικά συστήματα κυβερνοασφάλειας. Σε περιβάλλοντα με περιορισμένους πόρους, ωστόσο, θα μπορούσαν να χρησιμοποιηθούν πιο ελαφριά μοντέλα που απαιτούν λιγότερους πόρους, αν και παραμένει η πρόκληση να διατηρηθεί η ακρίβεια των αποτελεσμάτων σε ικανοποιητικό επίπεδο **[61]**.

Μελλοντική Έρευνα**:**

Η μελλοντική έρευνα μπορεί να επικεντρωθεί στην ανάπτυξη ελαφρύτερων μοντέλων και αλγορίθμων που προσφέρουν ισορροπία μεταξύ απόδοσης και ακρίβειας, ανταποκρινόμενοι στις ανάγκες συστημάτων με περιορισμένους πόρους. Η ελαστικότητα των αλγορίθμων, δηλαδή η ικανότητά τους να προσαρμόζουν την απόδοση ανάλογα με τη διαθεσιμότητα των πόρων, είναι κρίσιμη, καθώς η απλή αύξηση των πόρων δεν διασφαλίζει πάντοτε την ανάλογη βελτίωση των αποτελεσμάτων. Επιπλέον, τεχνικές βελτιστοποίησης πόρων, όπως η κατανεμημένη επεξεργασία και οι κλιμακώσιμες λύσεις βασισμένες στο υπολογιστικό νέφος (cloud), μπορούν να ενισχύσουν την αποδοτικότητα της ΤΝ, διασφαλίζοντας την αξιόπιστη ανίχνευση και αντιμετώπιση των απειλών ακόμη και σε περιορισμένα (ως προς πόρους) περιβάλλοντα, μειώνοντας παράλληλα το κόστος (διότι δεν χρειάζεται (επιπλέον) επένδυση σε επιτόπια υποδομή).

5.4 Τα Ηθικά Ζητήματα

Η ΤΝ βασίζεται στην ανάλυση μεγάλων όγκων δεδομένων, γεγονός που εγείρει ερωτήματα σχετικά με την προστασία της ιδιωτικότητας και την ηθική χρήση αυτής της τεχνολογίας **[62].** Η συλλογή και η επεξεργασία δεδομένων, ιδιαίτερα σε περιπτώσεις που αφορούν προσωπικά δεδομένα, μπορεί να οδηγήσει σε καταχρήσεις και παραβιάσεις της ιδιωτικότητας των ατόμων. Παράλληλα, οι αλγόριθμοι ΤΝ συχνά αντιμετωπίζουν την πρόκληση των προκαταλήψεων που ενδέχεται να είναι ενσωματωμένες στα δεδομένα εκπαίδευσης, αυξάνοντας τον κίνδυνο για ανισότητες ή άδικες αποφάσεις, ειδικά σε εφαρμογές που σχετίζονται με την ασφάλεια **[41]**. Αυτές οι προκαταλήψεις μπορούν να διαδοθούν μέσω των διαδικασιών λήψης αποφάσεων, με συνέπεια να δημιουργούνται ανισότητες ή άδικα αποτελέσματα. Για παράδειγμα, οι αλγοριθμικές προκαταλήψεις σε εφαρμογές που σχετίζονται με την ασφάλεια μπορεί να στοχεύουν δυσανάλογα συγκεκριμένες ομάδες ή να τις αποκλείουν από τη δίκαιη μεταχείριση. Η πρόκληση αυτή είναι κρίσιμη, καθώς τα μεροληπτικά συστήματα τεχνητής νοημοσύνης μπορούν να υπονομεύσουν την εμπιστοσύνη και να διαιωνίσουν τις κοινωνικές ανισότητες **[40]**.

Μελλοντική Έρευνα:

Για την αντιμετώπιση αυτών των προκλήσεων, η μελλοντική έρευνα πρέπει να δώσει προτεραιότητα στην ανάπτυξη ισχυρών πολιτικών και ρυθμιστικών πλαισίων που επιβάλλουν την ηθική χρήση των τεχνολογιών τεχνητής νοημοσύνης. Τα πλαίσια αυτά θα πρέπει να διασφαλίζουν ότι τα συστήματα ΤΝ σχεδιάζονται και εφαρμόζονται με γνώμονα τη δικαιοσύνη, τη διαφάνεια και τη λογοδοσία **[40]**.

Μια πολλά υποσχόμενη ερευνητική κατεύθυνση περιλαμβάνει τον σχεδιασμό αλγορίθμων που ενσωματώνουν ηθικές αρχές της ΤΝ και σέβονται την ιδιωτική ζωή. Για παράδειγμα, τεχνικές όπως η Διαφορική Ιδιωτικότητα -η οποία προσθέτει τυχαίο θόρυβο στα δεδομένα για να αποκρύψει τις ατομικές συνεισφορές, διατηρώντας παράλληλα τη χρησιμότητα του συνόλου δεδομένων- μπορούν να προστατεύσουν την ιδιωτικότητα χωρίς να θυσιάσουν την ακρίβεια. Ομοίως, οι τεχνικές ανωνυμοποίησης, όπως η απο-ταυτοποίηση ή η ψευδωνυμοποίηση, μπορούν να συμβάλλουν στη μείωση των κινδύνων κατάχρησης δεδομένων **[41]**.

Επιπλέον, η επίτευξη ισορροπίας μεταξύ της ασφάλειας και του σεβασμού των προσωπικών δεδομένων απαιτεί καινοτόμες προσεγγίσεις που ευθυγραμμίζουν τις τεχνικές δυνατότητες με τα ηθικά πρότυπα. Η ενσωμάτωση μεθόδων διατήρησης της ιδιωτικής ζωής, όπως η ομοσπονδιακή μάθηση, επιτρέπει την εκπαίδευση μοντέλων ΤΝ σε αποκεντρωμένες πηγές δεδομένων, μειώνοντας τους κινδύνους που συνδέονται με την κεντρική συλλογή δεδομένων και διατηρώντας παράλληλα την ακρίβεια του μοντέλου **[41]**. Σε συνδυασμό με ισχυρές δομές διακυβέρνησης και συνεχή παρακολούθηση, οι προσεγγίσεις αυτές μπορούν να ανοίξουν το δρόμο για υπεύθυνη και δίκαιη τεχνητή νοημοσύνη.

Τέλος, ένας άλλος κρίσιμος τομέας έρευνας περιλαμβάνει την αξιολόγηση και τον μετριασμό των προκαταλήψεων στα συστήματα τεχνητής νοημοσύνης. Προς αυτό τον σκοπό, αναπτύσσονται τεχνικές, όπως η μάθηση με επίγνωση της δικαιοσύνης και ο έλεγχος των προκαταλήψεων, για τον συστηματικό εντοπισμό και τη διόρθωση των προκαταλήψεων κατά την εκπαίδευση και την ανάπτυξη των μοντέλων ΤΝ. Επιπλέον, η διεπιστημονική συνεργασία μεταξύ των ηθικολόγων, των επιστημόνων δεδομένων και των υπεύθυνων χάραξης πολιτικής μπορεί να προωθήσει τη δημιουργία κατευθυντήριων γραμμών για συγκεκριμένο πλαίσιο προς αντιμετώπιση τόσο των τεχνικών όσο και των κοινωνικών επιπτώσεων της ΤΝ **[40]**.

5.5 Επιθέσεις σε Συστήματα ΤΝ & ΜΜ

Η αυξανόμενη ενσωμάτωση της ΤΝ και της ΜΜ σε συστήματα ασφαλείας έχει οδηγήσει στην εμφάνιση νέων τύπων επιθέσεων που στοχεύουν ειδικά αυτά τα συστήματα. Οι επιθέσεις αυτές εκμεταλλεύονται τις ιδιαιτερότητες των αλγορίθμων ΤΝ/ΜΜ, με σκοπό να παρακάμψουν ή να διαβρώσουν τα αντίστοιχα μέτρα ασφαλείας. Οι κύριες κατηγορίες αυτών των επιθέσεων περιλαμβάνουν:

1. **Model/Data Poisoning [10]**: Η δηλητηρίαση μοντέλων και δεδομένων είναι μια κρίσιμη πρόκληση στα συστήματα τεχνητής νοημοσύνης, όπου οι αντίπαλοι χειραγωγούν σύνολα δεδομένων εκπαίδευσης για να υποβαθμίσουν την απόδοση ενός μοντέλου ή να ενσωματώσουν ευπάθειες τύπου backdoor. Αυτή η μορφή επίθεσης μπορεί να θέσει σε κίνδυνο την ακεραιότητα των μοντέλων με την εισαγωγή λανθασμένα επισημασμένων ή εχθρικά διαμορφωμένων δεδομένων που διαστρεβλώνουν τη διαδικασία μάθησης. Μια ιδιαίτερα ανησυχητική πτυχή που τονίζεται στο **[10]** είναι ο κίνδυνος για τα θεμελιώδη μοντέλα, τα οποία χρησιμεύουν ως βάση για τη λεπτομερή ρύθμιση σε πολλαπλές εφαρμογές. Η δηλητηρίαση σε αυτό το θεμελιώδες επίπεδο μπορεί να διαδώσει ευπάθειες σε πολλά συστήματα που το ενσωματώνουν, ενισχύοντας τον αντίκτυπο μιας μεμονωμένης επίθεσης. Για παράδειγμα, οι επιτιθέμενοι μπορούν να εισάγουν στα δεδομένα εκπαίδευσης ενεργοποιητές που ενεργοποιούνται μόνο υπό συγκεκριμένες συνθήκες, επιτρέποντας την εκμετάλλευση σε πραγματικές εφαρμογές/συνθήκες, ενώ παραμένουν απαρατήρητοι κατά τις τυπικές φάσεις δοκιμών.
2. **Evasion Attacks [11]**: Οι επιθέσεις αποφυγής αποτελούν σημαντική απειλή για τους ανιχνευτές (detectors) που βασίζονται στην ενισχυτική μάθηση, ιδίως σε κρίσιμες υποδομές, όπως τα έξυπνα δίκτυα. Αυτές οι επιθέσεις χειραγωγούν τα δεδομένα εισόδου με λεπτούς τρόπους για να κάνουν την κακόβουλη δραστηριότητα να φαίνεται καλοήθης, παρακάμπτοντας έτσι τους μηχανισμούς ανίχνευσης. Η μελέτη των El-Toukhy και συν **[11]** διερευνά την ευπάθεια τέτοιων συστημάτων σε αντίπαλες τεχνικές χρησιμοποιώντας ένα Double Deep Q-Network (DDQN) για τη δημιουργία αντίπαλων δειγμάτων αποφυγής. Τα ευρήματα αποκαλύπτουν ότι οι συμβατικοί ανιχνευτές

δυσκολεύονται να αντιμετωπίσουν αυτές τις επιθέσεις, υπογραμμίζοντας την ανάγκη για προηγμένες άμυνες προσαρμοσμένες στις αντίπαλες εισροές. Αυτό αναδεικνύει την αυξανόμενη πολυπλοκότητα των επιθέσεων αποφυγής και τη δυνατότητά τους να θέσουν σε κίνδυνο βασικά πλαίσια κυβερνοασφάλειας.

3. **Model Stealing & Reverse Engineering [12]**: Η κλοπή μοντέλων και η αντίστροφη μηχανική αποτελούν σημαντικές προκλήσεις για την ασφάλεια των ιδιόκτητων συστημάτων μηχανικής μάθησης, ιδίως σε περιβάλλοντα μηχανικής μάθησης ως υπηρεσία (MLaaS) που βασίζονται στο υπολογιστικό νέφος. Η εργασία στο **[12]** εξετάζει τον τρόπο με τον οποίο οι αντίπαλοι μπορούν να εκμεταλλευτούν τόσο τις μεθοδολογίες επίθεσης βάσει ερωτημάτων όσο και τις μεθοδολογίες επίθεσης πλευρικού καναλιού για να αποσπάσουν ευαίσθητες πληροφορίες από ένα μοντέλο. Οι μέθοδοι που βασίζονται σε ερωτήματα περιλαμβάνουν την αποστολή συνθετικών ή αντίπαλων εισόδων στο σύστημα και την ανάλυση των εξόδων για την ανακατασκευή της αρχιτεκτονικής ή των παραμέτρων του μοντέλου. Αυτές οι επιθέσεις βασίζονται σε μεγάλο βαθμό στην ελαχιστοποίηση του προϋπολογισμού των ερωτημάτων για την αποφυγή της ανίχνευσης, μεγιστοποιώντας παράλληλα την αποδοτικότητα της εξαγωγής. Επιπλέον, οι επιθέσεις πλευρικού καναλιού χρησιμοποιούν μη λειτουργικά δεδομένα, όπως ο χρόνος εξαγωγής συμπερασμάτων ή τα ειδικά για το υλικό πρότυπα, για να συμπεράνουν τις αρχιτεκτονικές του μοντέλου και τις εσωτερικές λειτουργίες χωρίς άμεση πρόσβαση στα βάρη του μοντέλου. Αυτή η προσέγγιση διπλής μεθόδου καταδεικνύει πώς οι αντίπαλες εικόνες (δηλαδή δεδομένα μορφής εικόνας που δημιουργούνται συνθετικά ή σχεδιάζονται ώστε να προκαλέσουν συγκεκριμένες αποκρίσεις από το μοντέλο), σε συνδυασμό με τις πληροφορίες που βασίζονται στον χρόνο, μπορούν να αποτυπώσουν αποτελεσματικά και να αντιστρέψουν τα νευρωνικά δίκτυα συνελικτικής ανάλυσης (CNN) με υψηλή ακρίβεια.

Μελλοντική Έρευνα:

Για την αντιμετώπιση αυτών των προκλήσεων, η έρευνα πρέπει να επικεντρωθεί στις ακόλουθες κατευθύνσεις αριθμημένες σε αντιστοίχιση των προαναφερόμενων προκλήσεων:

1. Για την αντιμετώπιση αυτών των κινδύνων, προτείνονται διάφορες μελλοντικές στρατηγικές. Ενισχυμένοι μηχανισμοί ανίχνευσης, όπως τεχνικές ανίχνευσης ανωμαλιών ή αλγόριθμοι μάθησης χωρίς επίβλεψη, μπορούν να βοηθήσουν στον εντοπισμό των δηλητηριασμένων δεδομένων κατά τη διαδικασία εκπαίδευσης. Επιπλέον, η χρησιμοποίηση ανθεκτικών μεθοδολογιών εκπαίδευσης, συμπεριλαμβανομένων των ανθεκτικών σε αντιπάλους πλαισίων και των τεχνικών διαφορικής ιδιωτικότητας, μπορεί να αντιμετωπίσει προληπτικά τις απόπειρες δηλητηρίασης. Η διασφάλιση της ανθεκτικότητας των διαδικασιών μάθησης μεταφοράς είναι επίσης ζωτικής σημασίας για την αποτροπή των αλυσιδωτών επιπτώσεων της δηλητηρίασης από τα θεμελιώδη στα επόμενα μοντέλα. Επιπλέον, η ενσωμάτωση συστημάτων παρακολούθησης σε πραγματικό χρόνο μπορεί να ανιχνεύσει και να

μετριάσει τις απόπειρες δηλητηρίασης κατά τη διάρκεια ενημερώσεων μοντέλων ή σε περιβάλλοντα συνεχούς μάθησης. Αυτά τα μέτρα στοχεύουν συλλογικά στην ενίσχυση του οικοσυστήματος ΤΝ έναντι των εξελισσόμενων απειλών που θέτουν τα δεδομένα και η δηλητηρίαση μοντέλων **[10]**.

2. Για την αντιμετώπιση των επιθέσεων αποφυγής, η μελέτη στο **[11]** προτείνει διάφορες στρατηγικές για την ενίσχυση της ευρωστίας των συστημάτων ανίχνευσης. Τονίζεται η εκπαίδευση με αντίπαλο τρόπο ως μια κρίσιμη προσέγγιση, η οποία περιλαμβάνει την επανεκπαίδευση των ανιχνευτών με δείγματα διαφυγής για τη βελτίωση της ανθεκτικότητάς τους απέναντι σε χειραγωγημένες εισόδους. Επιπλέον, οι συγγραφείς υποστηρίζουν την ανάπτυξη διασταυρούμενων αμυντικών μοντέλων ικανών να γενικεύουν τις άμυνές τους ενάντια σε διάφορους τύπους αντιπαλίνδρομων επιθέσεων (Adversarial attacks). Η αντιμετώπιση της υπολογιστικής αναποτελεσματικότητας της αντιπαλίνδρομης εκπαίδευσης (Adversarial training) προσδιορίζεται ως βασικός τομέας για μελλοντική έρευνα, με στόχο να καταστούν αυτοί οι μηχανισμοί άμυνας κλιμακούμενοι για εφαρμογές στον πραγματικό κόσμο. Επιπλέον, προτείνονται υβριδικές στρατηγικές άμυνας που συνδυάζουν στατιστικές μεθόδους με μοντέλα μηχανικής μάθησης για την αξιοποίηση των πλεονεκτημάτων και των δύο προσεγγίσεων, δημιουργώντας ένα πιο ολοκληρωμένο και ισχυρό πλαίσιο για την ανίχνευση εξελιγμένων επιθέσεων αποφυγής σε κρίσιμα συστήματα, όπως τα έξυπνα δίκτυα **[11]**.
3. Η άμυνα κατά της κλοπής μοντέλων απαιτεί καινοτόμες στρατηγικές για τη διατήρηση της πνευματικής ιδιοκτησίας και τη διασφάλιση της λειτουργικής ακεραιότητας. Οι συγγραφείς στο **[12]** προτείνουν την ενσωμάτωση του θορύβου στις εξόδους των μοντέλων και την επιβολή αυστηρών μέτρων περιορισμού του ρυθμού (χρήσης) ως άμεσες άμυνες. Ωστόσο, αυτά μπορούν να εμποδίσουν τη νόμιμη χρήση, οδηγώντας στην ανάγκη για πιο προηγμένους μηχανισμούς προστασίας. Αναδυόμενες λύσεις, όπως η υδατοσήμανση μοντέλων ή η αποτύπωση δακτυλικών αποτυπωμάτων, οι οποίες ενσωματώνουν μοναδικά αναγνωριστικά στο μοντέλο κατά τη διάρκεια της εκπαίδευσης, παρέχουν πολλά υποσχόμενες κατευθύνσεις για την αυθεντικοποίηση και τη διασφάλιση των συστημάτων ML. Επιπλέον, η μείωση της δυνατότητας μεταφοράς αντίπαλων παραδειγμάτων και η ενίσχυση της ανθεκτικότητας του μοντέλου έναντι συμπερασμάτων που βασίζονται στον χρόνο είναι απαραίτητα για τη μελλοντική έρευνα. Αυτές οι στρατηγικές πρέπει να συμπληρωθούν από ρυθμιστικά πλαίσια που θα αντιμετωπίζουν τις ηθικές και νομικές διαστάσεις της ασφάλειας των μοντέλων σε ανταγωνιστικά οικοσυστήματα τεχνητής νοημοσύνης **[12]**.

5.6 Προσαρμοστικότητα σε Νέες Επιθέσεις και Διαχείριση Model Drifting

Η προσαρμογή σε νέες επιθέσεις και η διαχείριση της μετατόπισης του μοντέλου (model drifting) αποτελούν σημαντικές προκλήσεις λόγω της δυναμικής φύσης των δεδομένων. Η

μετατόπιση εννοιών (concept drift) αναφέρεται σε αλλαγές στην υποκείμενη κατανομή των δεδομένων με την πάροδο του χρόνου, επηρεάζοντας την ακρίβεια των μοντέλων μηχανικής μάθησης. Αυτές οι αλλαγές μπορεί να είναι ξαφνικές ή σταδιακές, καθιστώντας δύσκολη την ανίχνευση και την προσαρμογή των μοντέλων σε νέα μοτίβα επιθέσεων. Οι αλγόριθμοι προσαρμοστικής μάθησης διαδραματίζουν κρίσιμο ρόλο στην αντιμετώπιση της μετατόπισης εννοιών, ενσωματώνοντας μηχανισμούς "λησμόνησης" για την απόρριψη παλαιότερων, μη σχετικών δεδομένων και την ενημέρωση των μοντέλων με νέες πληροφορίες. Τεχνικές όπως τα ολισθαίνοντα παράθυρα (sliding windows) και τα παράθυρα ορόσημο (landmark windows) χρησιμοποιούνται για τη διαχείριση των πιο πρόσφατων δεδομένων, επιτρέποντας στα συστήματα να αντιδρούν γρήγορα σε απότομες αλλαγές στα πρότυπα επιθέσεων, διατηρώντας παράλληλα τη σταθερότητα κατά τη διάρκεια σταδιακών αλλαγών. Επιπλέον, οι προσεγγίσεις μάθησης συνόλου (ensemble learning) ενισχύουν την προσαρμοστικότητα διατηρώντας μια ομάδα μοντέλων, καθένα από τα οποία έχει εκπαιδευτεί για την αντιμετώπιση συγκεκριμένων πτυχών των εξελισσόμενων δεδομένων. Αυτή η ποικιλομορφία εξασφαλίζει ανθεκτικότητα απέναντι σε διάφορους τύπους μετατόπισης, συμπεριλαμβανομένων νέων ή λεπτών μοτίβων επιθέσεων **[13].**

Μελλοντική Έρευνα:

Οι μελλοντικές ερευνητικές κατευθύνσεις επικεντρώνονται στην ενίσχυση της προβλεψιμότητας, της προσαρμοστικότητας, και της ανθεκτικότητας των μοντέλων μάθησης απέναντι στη μετατόπιση εννοιών και τη δυναμική φύση των δεδομένων. Ένας σημαντικός στόχος είναι η ενσωμάτωση εξελιγμένων μηχανισμών ανίχνευσης ολίσθησης (concept drift detection) απευθείας στα εκπαιδευόμενα μοντέλα μηχανικής μάθησης, επιτρέποντας την έγκαιρη προληπτική προσαρμογή πριν η μετατόπιση επηρεάσει σημαντικά την απόδοσή τους. Επιπλέον, η ανάπτυξη υβριδικών προσεγγίσεων που συνδυάζουν εποπτευόμενες, μη εποπτευόμενες, και ενισχυτικές μεθόδους μάθησης μπορεί να προσφέρει μεγαλύτερη ευελιξία και ακρίβεια στον εντοπισμό λεπτών ή σπάνιων μοτίβων επιθέσεων. Ειδικότερα, η ενίσχυση μηχανισμών "λησμόνησης" για την έξυπνη διαχείριση δεδομένων και η χρήση ολισθαίνοντων παραθύρων ή παραθύρων ορόσημο θα μπορούσαν να βελτιώσουν τη διαχείριση δεδομένων που μεταβάλλονται δυναμικά. Η αποτελεσματική αντιμετώπιση της ανισορροπίας δεδομένων και η ανάπτυξη μεθόδων που επιτρέπουν στα μοντέλα να μαθαίνουν από μικρές ποσότητες νέων δεδομένων είναι επίσης καίριοι τομείς έρευνας. Τέλος, η δημιουργία κλιμακούμενων και υπολογιστικά αποδοτικών συστημάτων που μπορούν να εφαρμοστούν σε μεγάλης κλίμακας ροές δεδομένων είναι απαραίτητη για την πρακτική υλοποίηση αυτών των λύσεων σε περιβάλλοντα πραγματικού κόσμου **[13]**.

5.7 Εξηγησιμότητα και Διαφάνεια

Πολλά μοντέλα ΤΝ λειτουργούν ως «μαύρα κουτιά», με αποτέλεσμα να είναι δύσκολο να κατανοηθούν ή να αξιολογηθούν οι αποφάσεις τους. Αυτό καθιστά δύσκολο τον εντοπισμό και

τη διόρθωση προβλημάτων, όπως η ανακρίβεια ή οι λανθασμένες προβλέψεις **[40]**. Το άρθρο του Herzog **[14]** εξετάζει τη λεπτή διάκριση μεταξύ ερμηνευσιμότητας και εξηγησιμότητας στην τεχνητή νοημοσύνη (AI). Ενώ η ερμηνευσιμότητα επικεντρώνεται στο να καταστήσει κατανοητή την εσωτερική λειτουργία των μοντέλων, η εξηγησιμότητα επεκτείνει την έννοια αυτή ώστε να συμπεριλάβει τη διαφάνεια, τη λογοδοσία και την επικοινωνία με γνώμονα τα ενδιαφερόμενα μέρη. Ο Herzog **[14]** υποστηρίζει ότι η εξηγησιμότητα είναι ζωτικής σημασίας για την υπεύθυνη τεχνητή νοημοσύνη, καθώς δίνει τη δυνατότητα όχι μόνο στους προγραμματιστές αλλά και στους χρήστες και στους ενδιαφερόμενους φορείς να εξετάζουν και να αμφισβητούν τα αποτελέσματα των συστημάτων τεχνητής νοημοσύνης. Η αρχή της εξηγηματικότητας απαιτεί οι εξηγήσεις να προσαρμόζονται σε συγκεκριμένα πλαίσια και ακροατήρια, διασφαλίζοντας ότι οι διαφορετικοί χρήστες - από τεχνικούς εμπειρογνώμονες έως απλούς χρήστες - μπορούν να λαμβάνουν τεκμηριωμένες αποφάσεις με βάση τα αποτελέσματα της ΤΝ. Ο συγγραφέας τονίζει ότι η στήριξη αποκλειστικά στην ερμηνευσιμότητα μπορεί να αποτύχει να αντιμετωπίσει τις ευρύτερες επιχειρησιακές επιπτώσεις της ανάπτυξης της ΤΝ σε πραγματικές συνθήκες **[14].**

Μελλοντική Έρευνα:

Ο Herzog **[14]** περιγράφει διάφορες στρατηγικές για την ενίσχυση της επεξηγηματικότητας στα συστήματα τεχνητής νοημοσύνης. Μια προσέγγιση περιλαμβάνει τον σχεδιασμό διαδραστικών και χρηστοκεντρικών επεξηγηματικών διεπαφών που μεταφράζουν πολύπλοκες αλγοριθμικές διαδικασίες σε κατανοητές αφηγήσεις για τους διάφορους ενδιαφερόμενους. Επιπλέον, ζητείται η ενσωμάτωση της εξηγησιμότητας στον κύκλο ζωής της ανάπτυξης των συστημάτων ΤΝ, ώστε να διασφαλιστεί η λογοδοσία σε όλα τα στάδια, από τον σχεδιασμό έως την ανάπτυξη. Ο συγγραφέας υπογραμμίζει επίσης τη σημασία της ενσωμάτωσης ηθικών και κοινωνικών εκτιμήσεων στα πλαίσια εξηγησιμότητας για την εξισορρόπηση της τεχνικής διαφάνειας με τη λειτουργική σκοπιμότητα. Η μελλοντική έρευνα θα πρέπει να θέσει ως προτεραιότητα την ανάπτυξη εργαλείων και μεθόδων που γεφυρώνουν το χάσμα μεταξύ ερμηνευσιμότητας και εξηγησιμότητας, εστιάζοντας σε συγκεκριμένες εφαρμογές που αφορούν το πλαίσιο και δίνουν τη δυνατότητα σε όλους τους χρήστες να αναλάβουν την ευθύνη για τις αποφάσεις που λαμβάνονται με βάση την ΤΝ. Τέλος, η ανάπτυξη εξηγήσιμων μοντέλων ΤΝ (Explainable AI) μπορεί να διευκολύνει την κατανόηση των αποφάσεων ενός συστήματος, παρέχοντας σαφείς αιτιολογήσεις, ενώ παράλληλα, εργαλεία, όπως το SHAP (Shapley Additive Explanations) **[68]** και το LIME (Local Interpretable Model-agnostic Explanations) **[69]**, μπορούν να χρησιμοποιηθούν για την αναγνώριση των σημαντικών χαρακτηριστικών στα δεδομένα που επηρεάζουν τις αποφάσεις του μοντέλου **[14]**.

Κεφάλαιο 6: Συμπεράσματα

Σε αυτό το κεφάλαιο θα αναφερθούν τα συμπεράσματα και η συνεισφορά της παρούσας μελέτης, καθώς και η προσωπική εμπειρία που αποκομίσθηκε. Επιπλέον, θα δοθεί ιδιαίτερη έμφαση στα πλεονεκτήματα που προκύπτουν από την ανάλυση που πραγματοποιήθηκε. Στη συνέχεια, θα παρουσιαστούν τα πιο βασικά συμπεράσματα που παρήχθησαν, αναδεικνύοντας τις πιο κρίσιμες προκλήσεις που εντοπίστηκαν κατά τη διάρκεια της έρευνας. Τέλος, θα προταθούν οι μελλοντικές κατευθύνσεις που είναι απαραίτητες για την υιοθέτηση άμεσων μέτρων, με στόχο την περαιτέρω ανάπτυξη και βελτίωση στον τομέα της ανίχνευσης και αντιμετώπισης επιθέσεων κοινωνικής μηχανικής.

6.1 Συνεισφορά & Πλεονεκτήματα

Η σημασία της τεχνητής νοημοσύνης (ΤΝ) στον εντοπισμό και την αποτροπή των επιθέσεων κοινωνικής μηχανικής αποτελεί το κυριότερο τμήμα σε αυτή τη διατριβή. Μέσα από τη μεθοδική εξέταση και σύγκριση πολυάριθμων αντιμέτρων για κοινωνικές επιθέσεις, αναδείχθηκε η δυνατότητα που έχουν αυτά τα αντίμετρα να ενισχύουν την ασφάλεια στον κυβερνοχώρο, όταν συνδυάζονται με τις σωστές τεχνικές ΤΝ. Τα αντίμετρα, κατηγοριοποιημένα ανάλογα με το είδος της επίθεσης και τον τρόπο αντιμετώπισης που προσφέρουν (προληπτικά, ανίχνευση σε πραγματικό χρόνο και αντιδραστικά), παρέχουν μια ολιστική άμυνα ενάντια σε απειλές, προσαρμοσμένη στις συγκεκριμένες ανάγκες κάθε τύπου επίθεσης. Η συμβολή της παρούσας μελέτης έγκειται στην ανάλυση της αποδοτικότητας αυτών των αντιμέτρων, καθώς και στον εντοπισμό ιδιαίτερων χαρακτηριστικών τους που μπορούν να οδηγήσουν στη βελτιστοποίηση των μέτρων ασφάλειας.

Παράλληλα, η έρευνα εστίασε στη συγκριτική μελέτη των αλγορίθμων τεχνητής νοημοσύνης που χρησιμοποιούνται από τα αντίμετρα για την ανίχνευση και τον μετριασμό των επιθέσεων κοινωνικής μηχανικής. Με αυτό τον τρόπο, αναδεικνύει εκείνες τις τεχνικές ΤΝ που ξεχωρίζουν ανάλογα και με τις προτιμήσεις ενός οργανισμού για διάφορες κατηγορίες επιθέσεων, δίνοντας στον αναγνώστη έναν οδικό χάρτη για την σωστή επιλογή της κάθε ΤΝ τεχνικής ανά περίσταση. Οι προσεγγίσεις μηχανικής μάθησης και βαθιάς μάθησης, όπως τα Νευρωνικά Δίκτυα και τα Recurrent Neural Networks, αποδείχθηκαν ιδιαίτερα αποτελεσματικές στην ανίχνευση απειλών και την προσαρμογή σε νέα μοτίβα επίθεσης, όπως στο spear phishing και στο vishing. Η ικανότητα των τεχνικών αυτών να ανιχνεύουν με ακρίβεια

τις εξελιγμένες απειλές, διακρίνοντας μεταξύ φυσιολογικών και κακόβουλων επικοινωνιών, υπογραμμίζει τη συμβολή τους στη βελτίωση της κυβερνοασφάλειας. Η μελέτη αναδεικνύει έτσι τον τρόπο με τον οποίο οι αλγόριθμοι ΤΝ μπορούν να αυξήσουν την αποτελεσματικότητα των αντιμέτρων, προσφέροντας μια δυναμική άμυνα απέναντι στις κοινωνικές επιθέσεις.

6.2 Κρίσιμες Προκλήσεις & Κατευθύνσεις

Η ανάλυση αποκάλυψε διάφορα σημαντικά ευρήματα, αναδεικνύοντας τόσο τις δυνατότητες όσο και τις προκλήσεις της ΤΝ στην καταπολέμηση των επιθέσεων κοινωνικής μηχανικής. Η τεχνητή νοημοσύνη υπόσχεται πολλά ως ισχυρό εργαλείο για τον εντοπισμό και τον μετριασμό τέτοιων επιθέσεων, παρέχοντας προσαρμοστικότητα στις εξελισσόμενες απειλές σε πραγματικό χρόνο. Ωστόσο, εξακολουθούν να υπάρχουν σημαντικά εμπόδια, όπως το υπολογιστικό κόστος, η πολυπλοκότητα της ενσωμάτωσης της ΤΝ σε υφιστάμενα συστήματα ασφαλείας και η ποιότητα των δεδομένων εκπαίδευσης. Τα ζητήματα αυτά επιδεινώνονται περαιτέρω από τη δυναμική φύση των στρατηγικών επίθεσης, την επικράτηση ανισόρροπων συνόλων δεδομένων και τους κινδύνους που σχετίζονται με την παρέκκλιση του μοντέλου και την παρέκκλιση των εννοιών.

Εκτός από αυτά τα τεχνικά εμπόδια, ξεχωρίζουν προκλήσεις όπως η έλλειψη εξηγήσιμης και διαφανούς λειτουργίας των μοντέλων τεχνητής νοημοσύνης, ιδίως λόγω της εξάρτησης από μεθοδολογίες «μαύρου κουτιού». Αυτή η έλλειψη διαφάνειας εμποδίζει την εμπιστοσύνη, τη λογοδοσία και την ικανότητα εντοπισμού και αντιμετώπισης ανακριβειών σε εφαρμογές στον πραγματικό κόσμο. Επιπλέον, οι ηθικές και νομικές επιπτώσεις της χρήσης της ΤΝ σε περιβάλλοντα ασφαλείας παραμένουν κρίσιμες. Πρέπει να αντιμετωπιστούν οι πιθανές παραβιάσεις της ιδιωτικής ζωής, οι αλγοριθμικές προκαταλήψεις και η μη συμμόρφωση με τις κανονιστικές διατάξεις, ώστε να διασφαλιστεί η υπεύθυνη υιοθέτηση. Επιθέσεις, όπως η δηλητηρίαση μοντέλων/δεδομένων, οι τακτικές αποφυγής και η αντίστροφη μηχανική, ενισχύουν αυτές τις προκλήσεις στοχεύοντας άμεσα τα συστήματα ΤΝ, γεγονός που απαιτεί ισχυρούς μηχανισμούς άμυνας.

Ενώ όλες οι προκλήσεις που εντοπίστηκαν απαιτούν προσοχή, ορισμένα ζητήματα απαιτούν άμεση ιεράρχηση στο πλαίσιο της κοινωνικής μηχανικής. Η ενσωμάτωση της εξηγησιμότητας και της διαφάνειας στα μοντέλα ΤΝ είναι υψίστης σημασίας, καθώς επηρεάζει άμεσα την εμπιστοσύνη και τη λειτουργική αποτελεσματικότητα των συστημάτων ανίχνευσης. Επιπλέον, η αντιμετώπιση των προκλήσεων που σχετίζονται με τα δεδομένα, συμπεριλαμβανομένης της ανισορροπίας, των συνόλων δεδομένων χαμηλής ποιότητας και της παρέκκλισης του μοντέλου, είναι κρίσιμη για να διασφαλιστεί ότι τα συστήματα ΤΝ παραμένουν ακριβή και συναφή με την πάροδο του χρόνου. Η ανάπτυξη τεχνικών για τη διαχείριση αυτών των προκλήσεων, όπως η ανίχνευση ανωμαλιών, η δημιουργία συνθετικών δεδομένων και οι ενημερώσεις μάθησης σε πραγματικό χρόνο, είναι απαραίτητη για την οικοδόμηση ανθεκτικότητας έναντι των εξελισσόμενων απειλών.

Για την περαιτέρω ανάπτυξη και βελτίωση στον τομέα της ανίχνευσης και αντιμετώπισης επιθέσεων κοινωνικής μηχανικής, είναι απαραίτητη η υιοθέτηση άμεσων μέτρων που εστιάζουν σε κρίσιμες κατευθύνσεις έρευνας και τεχνολογικής ανάπτυξης. Καταρχάς, η ανάπτυξη εξελιγμένων αλγορίθμων που συνδυάζουν ψυχολογικά και τεχνικά μοντέλα αποτελεί βασικό βήμα για την ενίσχυση της ανίχνευσης, καθώς οι επιθέσεις κοινωνικής μηχανικής βασίζονται στη συναισθηματική χειραγώγηση. Οι αλγόριθμοι τεχνητής νοημοσύνης πρέπει να αποκτήσουν τη δυνατότητα αναγνώρισης σύνθετων ψυχολογικών ενδείξεων, ώστε να εντοπίζουν τις σχετικές επιθέσεις αρκετά νωρίς και να τις αναχαιτίζουν εγκαίρως, ενισχύοντας τη συνολική προστασία των συστημάτων.

Παράλληλα, η βελτίωση των τεχνικών διαχείρισης μη ισορροπημένων δεδομένων είναι αναγκαία, δεδομένου ότι οι επιθέσεις κοινωνικής μηχανικής αποτελούν μικρό ποσοστό των συνολικών δεδομένων που αναλύονται. Η εφαρμογή τεχνικών ανίχνευσης ανωμαλιών και η δημιουργία τεχνητών δεδομένων θα επιτρέψει τη δημιουργία πιο ακριβών και αποτελεσματικών μοντέλων ανίχνευσης. Επιπλέον, η μείωση του υπολογιστικού κόστους είναι καθοριστική για τη δυνατότητα ευρείας υιοθέτησης των τεχνολογιών ΤΝ σε πραγματικά περιβάλλοντα. Η μελλοντική έρευνα πρέπει να εστιάσει στην ανάπτυξη πιο αποδοτικών και ελαφρών μοντέλων, καθώς και στη χρήση λύσεων κατανεμημένης επεξεργασίας και υπολογιστικού νέφους, ώστε να διευκολύνει την εφαρμογή τους σε μεγαλύτερη κλίμακα.

Τέλος, οι δεοντολογικές και νομικές ανησυχίες πρέπει να διαχειρίζονται προληπτικά μέσω της θέσπισης σαφών κατευθυντηρίων γραμμών και ισχυρών κανονιστικών πλαισίων. Η διασφάλιση της ιδιωτικότητας των δεδομένων, της διαφάνειας και της δικαιοσύνης όχι μόνο θα ενισχύσει την αξιοπιστία των λύσεων που βασίζονται στην τεχνητή νοημοσύνη αλλά και θα διασφαλίσει την εμπιστοσύνη των χρηστών. Με την ιεράρχηση αυτών των τομέων, η μελλοντική έρευνα και ανάπτυξη μπορεί να επικεντρωθεί στη δημιουργία συστημάτων τεχνητής νοημοσύνης που δεν είναι μόνο τεχνικά αποτελεσματικά αλλά και ηθικά ορθά και λειτουργικά βιώσιμα για την καταπολέμηση της εξελιγμένης και δυναμικής φύσης των επιθέσεων κοινωνικής μηχανικής. Έτσι, συνολικά από όλη την διατριβή, υπογραμμίζεται η διττή φύση της ΤΝ, που μπορεί να χρησιμοποιηθεί τόσο ως όπλο, όσο και ως εργαλείο άμυνας. Για αυτόν τον λόγο, προτείνονται μέτρα για την αποτελεσματική και ηθική χρήση της.

6.3 Εμπειρία

Κατά την εκπόνηση της διατριβής μου, αποκόμισα χρήσιμες πρακτικές στους τομείς της τεχνητής νοημοσύνης και της ασφάλειας στον κυβερνοχώρο. Απέκτησα γνώσεις σχετικά με την διαδικασία σύγκρισης αλγορίθμων και αξιολόγησης της καταλληλότητας του καθενός για ένα συγκεκριμένο σενάριο επίθεσης, χρησιμοποιώντας τεχνικές που βασίζονται στη βιβλιογραφία. Επιπλέον, συνειδητοποίησα τη σημασία της συνεχούς προσθήκης νέων στρατηγικών και τακτικών στα συστήματα ασφαλείας μέσω της ανάλυσης και της επεξεργασίας δεδομένων. Η σφαιρική κατανόησή μου για τις δυνατότητες της τεχνητής νοημοσύνης και τις δυσκολίες στην πρόοδο των συστημάτων ασφάλειας στον κυβερνοχώρο βελτιώθηκε χάρη σε

αυτή τη διπλωματική εργασία. Η εμβάθυνση στις τεχνικές ΤΝ και η ανάλυση των μεθόδων αυτών, με έφεραν αντιμέτωπη με τις πρακτικές προκλήσεις της κυβερνοασφάλειας και ενίσχυσε την ικανότητά μου στην αξιολόγηση και εφαρμογή λύσεων σε σύγχρονα ζητήματα ασφαλείας. Από όλη την έρευνα, μπόρεσα να συνειδητοποιήσω την ουσία της περίφημης ιδέας, που έχει διατυπωθεί από πολλούς ειδικούς στην κυβερνοασφάλεια και αποτέλεσε μια γενική αρχή της, πως:

« Η ασφάλεια είναι μια διαδικασία και όχι ένα τελικό προϊόν, έτσι η απόλυτη ασφάλεια είναι ανέφικτη και στόχο αποτελεί η μέγιστη δυνατή προστασία απέναντι στις συνεχώς εξελισσόμενες απειλές! » **[52]**.

Βιβλιογραφία

[1] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2022). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389-422.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2022). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[3] He, H., Bai, Y., Garcia, E. A., & Li, S. (2018). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 1322-1328.

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[5] García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer.

[6] Rahm, E., & Do, H. H. (2014). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

[7] Jain, A. K., Duin, R. P. W., & Mao, J. (2014). Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 4-37.

[8] Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

[9] Karwan Mst. (2014). Introduction to Computer and Network Security. SlideShare. https://www.slideshare.net/karwanmst/introduction-to-computer-and-network-security-41870289

[10] Wang, G., Zhou, C., Wang, Y., Chen, B., Guo, H., & Yan, Q. (2023). Beyond Boundaries: A Comprehensive Survey of Transferable Attacks on AI Systems. Retrieved from https://arxiv.org/abs/2311.11796

[11] El-Toukhy, A. T., Mahmoud, M. M. E. A., Bondok, A. H., Fouda, M. M., & Alsabaan, M. (2023). Countering evasion attacks for smart grid reinforcement learning-based detectors. IEEE Access, 11, 97373–97388. https://ieeexplore.ieee.org/abstract/document/10242112/

[12] Shukla, S., Alam, M., Mitra, P., & Mukhopadhyay, D. (2024). Stealing the invisible: Unveiling pre-trained CNN models through adversarial examples and timing side-

channels. arXiv preprint arXiv:2402.11953. Retrieved from https://arxiv.org/abs/2402.11953

[13] Khamassi, I., Sayed-Mouchaweh, M., Hammami, M., & Ghédira, K. (2018). Discussion and review on evolving data streams and concept drift adapting. Evolving Systems, 9(1), 1–23. https://link.springer.com/article/10.1007/s12530-016-9168-2

[14] Herzog, C. (2022). On the risk of confusing interpretability with explicability. AI and Ethics, 2(3), 219–225. https://link.springer.com/article/10.1007/s43681-021-00121-9

[15] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 308-318.

[16] Mitnick, K. (2017). The art of invisibility: The world's most famous hacker teaches you how to be safe in the age of big brother and big data. Little, Brown.

[17] Oltramari, A., & Kott, A. (2018). Towards a reconceptualisation of cyber risk: An empirical and ontological study. Journal of Information Warfare, 17(1). arXiv preprintarXiv:1806.08349

[18] Palo Alto Networks. (n.d.). Cortex XDR: Extended detection and response for advanced threat protection. Retrieved from https://www.paloaltonetworks.com/cortex/cortex-xdr

[19] IBM. (n.d.). Resilient Incident Response Platform: Accelerate response to cyber threats with IBM Resilient. Retrieved from https://www.ibm.com/security/resilient

[20] Rajasekharaiah, K. M., Dule, C. S., & Sudarshan, E. (2020, December). Cyber security challenges and its emerging trends on latest technologies. In IOP Conference Series: Materials Science and Engineering (Vol. 981, No. 2, p. 022062). IOP Publishing.

[21] Ramirez, R., & Choucri, N. (2016). Improving interdisciplinary communication with standardized cyber security terminology: A literature review. IEEE Access, 4,2216–2243.

[22] Ross, R., McEvilley, M., and Oren, J. (2016). Systems security engineering: Considerations for a multidisciplinary approach in the engineering of trustworthy secure systems. Technical report, National Institute of Standards and Technology.

[23] Rouse, Margaret. (2019). "Social Engineering (security)." TechTarget. [Online]. Available: https://searchsecurity.techtarget.com/definition/social-engineering.

[24] Salahdine, Fatima (2019). "Social Engineering Attacks: A Survey". School of Electrical Engineering and Computer Science, University of North Dakota. 11 (4): 89.

[25] Schwartau, W. (2018). Information warfare: Chaos on the electronic superhighway. Avalon Publishing Group.

[26] Shackleford, Dave. (2020). "Cybersecurity Threat Intelligence: What You Need to Know." O'Reilly Media.

[27] Timón López, C., Alamillo Domingo, I., & Valero Torrijos, J. (2021, August). Approaching the Data Protection Impact Assessment as a legal methodology to evaluate the degree of privacy by design achieved in technological proposals. A special reference to Identity Management systems. In Proceedings of the 16th International Conference on Availability, Reliability and Security (pp. 1-9).

[28] Wang, Z., Zhu, H., & Sun, L. (2021). Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods. IEEE Access, 9, 11895-11910.

[29] Cybersecurity and Infrastructure Security Agency. (2021, February 1). Avoiding social engineering and phishing attacks. U.S. Department of Homeland Security. https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks

[30] Whitman, Michael E., and Herbert J. Mattord. (2021). "Principles of Information Security." Cengage Learning.

[31] Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. In Telecommunication Systems (Vol. 76, Issue 1, pp. 139–154). Springer.

[32] Heartfield, R., & Loukas, G. (2015). A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. In ACM Computing Surveys (Vol. 48, Issue 3). Association for Computing Machinery.

[33] Chakraborty, A., Biswas, A., & Khan, A. K. (n.d.). Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation.

[34] Rastogi, M., Chhetri, A., Singh, D. K., & Rajan V, G. (2021). Survey on Detection and Prevention of Phishing Websites using Machine Learning. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021, 78–82.

[35] Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions. In Electronics (Switzerland) (Vol. 12, Issue 6).

[36] Hasan, M. M., Islam, M. U., & Uddin, J. (2023). Advanced Persistent Threat Identification with Boosting and Explainable AI. SN Computer Science, 4(3).

[37] Salahdine, F., & Kaabouch, N. (2019). Social engineering attacks: A survey. Future Internet, 11(4), 89

[38] Mitnick, K. D. (2011). The art of deception: Controlling the human element of security. Wiley.

[39] Salahdine, F., & Kaabouch, N. (2019). Social engineering attacks: A survey. Future Internet, 11(4), 89.

[40] Ray, P. P., & Das, P. K. (2023). Charting the terrain of artificial intelligence: A multidimensional exploration of ethics, agency, and future directions. Philosophy & Technology, 36(40). https://link.springer.com/article/10.1007/s13347-023-00643-6

[41] El Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). Machine learning for practical decision making: A multidisciplinary perspective with applications from healthcare, engineering and business analytics (1st ed.). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-031-16990-8_16

[42] Connolly, L. Y., Wall, D. S., & Lang, M. (2020). An empirical study of ransomware attacks on organizations: An assessment of severity and salient factors affecting vulnerability. Journal of Cybersecurity, 6(1), 1-18.

[43] Rahman, M. R., Basak, S. K., Hezaveh, R. M., & Williams, L. (2021). Attackers reveal their arsenal: An investigation of adversarial techniques in CTI reports. ACM Transactions on Software Engineering and Methodology, 30(1), 1-31.

[44] Wright, R. T., Jensen, M. L., Thatcher, J. B., & Dinger, M. (2011). Influence techniques in phishing attacks: An examination of vulnerability and resistance. IEEE

[45] Pan, X. Y. (2024). Independent study of Splunk. Open Access Library Journal, 11(e11496). http://eprints.ditdo.in/id/eprint/2192/

[46] Šuškalo, D., Morić, Z., Redžepagić, J., & Regvart, D. (2023). Comparative analysis of IBM QRadar and Wazuh for security information and event management. Proceedings of the 34th DAAAM International Symposium on Intelligent Manufacturing and Automation.

https://www.daaam.info/Downloads/Pdfs/proceedings/proceedings_2023/working_papers/dpn34056_a_3_Moric.pdf

[47] Intel Corporation. (2022). Increasing situational awareness and multi-zone protection of industrial and utility infrastructure. https://www.intel.cn/content/dam/www/public/us/en/documents/solution-briefs/energy-intel-mcafee-utility-security-brief.pdf

[48] Cappello, M. (2024). A Comprehensive Analysis of EDR (Endpoint Detection & Response), EPP (Endpoint Protection Platform), and Antivirus Security Technologies [Master's thesis]. University of Piraeus, Department of Informatics. https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/16751/Cappello_mpked21016.pdf?sequence=5&isAllowed=y

[49] Topala, R. R. (2022). Cybersecurity system for enterprise telecommunications resources [Bachelor's thesis, National Aviation University]. Faculty of Aeronavigation, Electronics and Telecommunications. https://dspace.nau.edu.ua/handle/NAU/58500

[50] Jakobi, T., Nolte, A., & Tijou, A. (2021). The risks of remote work and the need for digital sovereignty: Social engineering during the COVID-19 pandemic. Journal of Cybersecurity, 7(3), 14-30. https://doi.org/10.1016/j.cyber.2021.103001

[51] Hall, M., & Smith, L. (2020). Targeting remote workers: Increased social engineering risks in the age of telecommuting. Security and Privacy, 18(2), 54-72. https://doi.org/10.1109/SP.2020.00007

[52] National Institute of Standards and Technology. (2018). *Framework for Improving Critical Infrastructure Cybersecurity* (Version 1.1). https://doi.org/10.6028/NIST.CSWP.04162018

[53] Subburaj, T., & Suthendran, K. (2018). Optimal Defense Strategy Selection for Spear-Phishing Attack Based on a Multistage Signaling Game. Journal of Cyber Security, 7(1),1-12. https://ieeexplore.ieee.org/document/8635453#:~:text=Motivated%20by%20this%20consideration,%20we%20construct%20the%20multistage%20spear-phishing%20attack-defense

[54] Huang, L., Wang, X., & Li, C. (2020). Machine learning-based phishing detection using dynamic and static methods. Journal of Information Security, 9(3), 201-216.https://arxiv.org/abs/2009.11116#:~:text=One%20of%20the%20most%20successful%20methods%20for%20detecting%20these%20malicious

[55] Chaudhuri, A. (2023). Clone phishing: Attacks and defenses. International Journal of Scientific and Research Publications, 13(4), 180-184. https://www.ijsrp.org/research-paper-0423.php?rp=P13612817

[56] Liaqat, M. S., Mumtaz, G., Rasheed, N., & Mubeen, Z. (2024). Exploring phishing attacks in the AI age: A comprehensive literature review. Journal of Computing & Biomedical Informatics, 7(2). https://www.jcbi.org/index.php/Main/article/view/567/534

[57] Bozkır, A. S., & Aydos, M. (2019). Local image descriptor-based phishing web page recognition as an open-set problem. European Journal of Science and Technology, (Special Issue), 444-451. https://dergipark.org.tr/en/pub/ejosat/issue/49069/638404

[58] Shafiq, M., Gu, Z., Bilal, M., & Sher, A. (2021). Comparative performance analysis of machine learning algorithms for efficient detection of phishing emails. Indian Journal of Computer Science and Engineering, 12(3), 298-308. https://www.ijcse.com/docs/INDJCSE21-12-03-298.pdf

[59] Hussain, F., Gupta, N., & Jalali, R. (2022). CCiDD: Cyber-crimes intrusion detection dataset for cyber security applications using machine learning and deep learning techniques. Computers & Electrical Engineering, 100, 107915. https://www.sciencedirect.com/science/article/pii/S0045790622001598

[60] Ali, A., & Syed, A. M. (2020). Cyberbullying Detection Using Machine Learning. Pakistan Journal of Engineering and Technology, SI(01), 45-50. https://journals.uol.edu.pk/pakjet/article/view/430/274

[61] Moallem, A. (Ed.). (2021). HCI for Cybersecurity, Privacy, and Trust: Third International Conference, HCI-CPT 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings (Vol. 12788). Springer. https://link.springer.com/chapter/10.1007/978-3-030-77392-2_27

[62] Rajeswary, C., & Thirumaran, M. (2023). A comprehensive survey of automated website phishing detection techniques: A perspective of artificial intelligence and human behaviors. In Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023) (pp. 420-427). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10104988

[63] Gururaj, H. L., Janhavi, V., & Ambika, V. (2024). Social engineering in cybersecurity: Threats and defenses. CRC Press. https://newsletter.radensa.ru/wp-content/uploads/2024/05/Gururaj_H_Social_Engineering_in_Cybersecurity_Threats_and_Defenses.pdf#page=158

[64] CrowdStrike. (2022). CrowdStrike Falcon Intel Indicator Add-on Guide. Retrieved from https://www.crowdstrike.com/wp-content/uploads/2022/12/CrowdStrike-Falcon-Intel-Indicator-Add-on-Guide.pdf

[65] Trellix. (n.d.). Enterprise Security Manager 11.4.x Product Guide. Retrieved from https://docs.trellix.com/bundle/enterprise-security-manager-11.4.x-product-guide/page/GUID-88473528-B9BD-4799-B3A7-BC7A8C22B55D.html

[66] Chen, X., Liu, X., Zhang, L., & Tang, C. (2019). Optimal defense strategy selection for spear-phishing attack based on a multistage signaling game. IEEE Access, 7, 19907-19920. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8635453

[67] "What Is Angler Phishing And How Do I Avoid Becoming A Victim?" https://blog.knowbe4.com/what-is-anglerphishing-and-how-do-i-avoid-becoming-avictim (accessed Aug. 06, 2021).

[68] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[69] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[70] Adler, S., Hitzig, Z., Jain, S., Brewer, C., Chang, W., DiResta, R., ... & Zick, T. (2024). Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online. OpenAI, Harvard Society of Fellows, Microsoft, University of Oxford, SpruceID, a16z crypto, UL Research Institutes, Tucows, Collective Intelligence Project, Massachusetts Institute of Technology, Decentralization Research Center, Digital Bazaar, American Enterprise Institute, Center for Human-Compatible AI, University of California, Berkeley, OpenMined, Decentralized Identity Foundation, Goodfire, Partnership on AI, eGovernments Foundation, University of Minnesota Law School, Mina Foundation, ex/ante, School of Information, University of California, Berkeley, Berkman Klein Center for Internet & Society, Harvard University, Independent Researcher. Retrieved from https://arxiv.org/abs/2408.07892

[71] Mouton, F., Malan, M. M., Kimppa, K. K., & Venter, H. S. (2014). Social engineering attack framework. Artificial Intelligence Review, 50(4), 507-517. https://link.springer.com/article/10.1007/s10462-024-10973-2

[72] Saleem, J., Hammoudeh, M. (2018). Defense methods against social engineering attacks. In Daimi, K. (Ed.), Computer and Network Security Essentials (pp. 603–618). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-77392-2_27

[73] Moghimi, A., Wichelmann, J., Eisenbarth, T., & Sunar, B. (2019). Memjam: a false dependency attack against constant-time crypto implementations. International Journal of Parallel Programming, 47(4), 538–570. https://link.springer.com/article/10.1007/s42979-021-00557-0

[74] Chaudhary, A., & Sharma, A. (2022). Improving the digital data protection by embedding the features of deep learning and blockchain technology in steganographic process. NeuroQuantology, 20(20), 2778-2791. https://www.proquest.com/docview/2771607344?fromopenview=true&pq-origsite=gscholar&sourcetype=Scholarly%20Journals

[75] Albanese, M., Johnsgard, K. L., & Swarup, V. (2022). A formal model for credential hopping attacks. In V. Atluri, R. Di Pietro, C. D. Jensen, & W. Meng (Eds.), Computer Security – ESORICS 2022 (Vol. 13554, pp. 367-386). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-031-17140-6_18

[76] Alsubaei, F. S., Almazroi, A. A., & Ayub, N. (2024). Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. IEEE Access, 12, 8373-8389.https://ieeexplore.ieee.org/abstract/document/10384876

[77] Figueiredo, J., Carvalho, A., Castro, D., Gonçalves, D., & Santos, N. (2024). On the Feasibility of Fully AI-automated Vishing Attacks. INESC-ID/IST, Universidade de Lisboa. https://arxiv.org/abs/2409.13793

[78] Kaur, G., Singh, K. D., Arora, J., Bagchi, S., Debnath, S. K., & Kumar, A. V. S. (2024). Naive Bayes Classifier-Based Smishing Detection Framework to Reduce Cyber Attack. In N. K. Marriwala, S. Dhingra, S. Jain, & D. Kumar (Eds.), Mobile Radio Communications and 5G Networks (pp. 23-33). Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-97-0700-3_3

[79] Lenko, F. (2021). Specifics of RFID Based Access Control Systems Used in Logistics Centers. Transportation Research Procedia, 55, 1613-1619. https://www.sciencedirect.com/science/article/pii/S235214652100569X?via%3Dihub

[80] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., & Van Gool, L. (2020). SCAN: Learning to Classify Images Without Labels. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.), Computer Vision – ECCV 2020 (pp. 268–285). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-58607-2_16

[81] Niu, C., Shan, H., & Wang, G. (2022). SPICE: Semantic Pseudo-Labeling for Image Clustering. IEEE Transactions on Image Processing, 31, 7264-7277. https://ieeexplore.ieee.org/abstract/document/9952205

[82] Krüger, O. (2021). The Singularity is near! Visions of Artificial Intelligence in Contemporary Esoteric Discourses. Re-Unir.https://reunir.unir.net/handle/123456789/13025

[83] Prianto, Y., Sumantri, V. K., & Sasmita, P. Y. (2020). Pros and cons of AI robot as a legal subject. In Proceedings of the Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2019) (pp. 380-387). Atlantis Press. https://www.atlantis-press.com/proceedings/ticash-19/125940625

[84] Aldawood, H. A., & Skinner, G. (2018). A critical appraisal of contemporary cyber security social engineering solutions: Measures, policies, tools and applications. IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8638166

[85] F. Callegati, W. Cerroni, and M. Ramilli. 2009. Man-in-the-middle attack to the HTTPS protocol. IEEE Security and Privacy 7, 1, 78–81.

[86] S. Ford, M. Cova, C. Kruegel, and G. Vigna. 2020. Analyzing and detecting malicious flash advertisements. In Proceedings of the Annual Computer Security Applications Conference (ACSAC'09). IEEE, 363– 372

[87] J. R. Jacobs. 2011. Measuring the Effectiveness of the USB Flash Drive as a Vector for Social Engineering Attacks on Commercial and Residential Computer Systems. Master's thesis. Embry-Riddle Aeronautical University

[88] Renvall, A. (2018, November 26). Improving cybersecurity through ISO/IEC 27001 information security standard https://www.theseus.fi/handle/10024/157277

[89] Ivory, D., Loukas, G., & Gan, D. (2024). Enabling the Human-as-a-Security-Sensor Paradigm in the Internet of Things. In Security and Privacy in Smart Environments (pp. 75–97). Springer. https://doi.org/10.1007/978-3-031-66708-4_4

[90] Kolluri, V. (2019). Revolutionary research on the AI Sentry: An approach to overcome social engineering attacks using machine intelligence. International Journal of Creative Research Thoughts (IJCRT), 7(3), 590-593. https://www.researchgate.net/profile/Venkateswaranaidu-Kolluri/publication/380723807_REVOLUTIONARY_RESEARCH_ON_THE_AI_SENTRY_AN_APPROACH_TO_OVERCOME_SOCIAL_ENGINEERING_ATTACKS_USING_MACHINE_INTELLIGENCE/links/664b4247479366623afe388e/REVOLUTIONARY-RESEARCH-ON-THE-AI-SENTRY-AN-APPROACH-TO-OVERCOME-SOCIAL-ENGINEERING-ATTACKS-USING-MACHINE-INTELLIGENCE.pdf

[91] Yang, S., Li, S., Chen, W., & Liu, Y. (2020). A real-time and adaptive-learning malware detection method based on API-Pair Graph. IEEE Access, 8, 208120-208134. https://ieeexplore.ieee.org/abstract/document/9261431

[92] Alsmadi, I., Ahmad, K., Nazzal, M., Alam, F., Al-Fuqaha, A., Khreishah, A., & Algosaibi, A. (2021). Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions. arXiv preprint arXiv:2110.13980v1.https://arxiv.org/abs/2110.13980

[93] E., & John, G. (2024). AI for Cybersecurity. https://www.researchgate.net/profile/Emmanuel-Ok-2/publication/386171171_AI_for_Cybersecurity_Aouthors/links/67472765a7fbc259f1924ef0/AI-for-Cybersecurity-Aouthors.pdf

[94] Piconese, F. (2020). Deployment of Next Generation Intrusion Detection Systems against Internal Threats in a Medium-sized Enterprise (Master's thesis, University of Turku). University of Turku, Department of Future Technologies.

[95] Shah, B. (2017). Cisco Umbrella: A cloud-based secure internet gateway. International Journal of Advanced Research in Computer Science, 8(2), 4–7. https://www.academia.edu/download/106120169/2914-5799-1-SM.pdf

[96] Kumar, A., & Singh, R. (2021). Integration of Splunk Enterprise SIEM for DDoS Attack Detection in IoT. In 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA) (pp. 1-5). IEEE. https://ieeexplore.ieee.org/abstract/document/9685977?casa_token=P9bBldFohXoAAAAA:5blplWVk-pFX7Oxc8AKQ7tWPiiBWZ8-s2nPkef369Nj2ZLRE-1mm6zNwmmXsyOPXzX435Mpq

[97] Kure, H., Islam, S., & Razzaque, M. A. (2015). Security information and event management (SIEM) in the cloud computing environment. 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 1-6. https://ieeexplore.ieee.org/abstract/document/7028677 /

[98] IBM. (n.d.). QRadar SIEM 7.5.0 Documentation. Retrieved from https://www.ibm.com/docs/en/qsip/7.5?topic=SS42VS_7.5%2Fcom.ibm.qradar.doc%2Fc_qradar_pdfs.html

[99] Kure, H., Islam, S., & Razzaque, M. A. (2016). An analysis of information security event managers. 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 1-6.https://ieeexplore.ieee.org/abstract/document/7545921?casa_token=nQc_HeU5B3sA

AAAA:cDanq0jdxtnH_Dbd6KX28rHOYY5gx0vl2jeROu9zz2g3LVQF8mpq4nmV47PN9DXi-e7NIkXC

[100] Patel, H., Raval, M., & Chaudhary, S. (2022). AI/ML in Security Orchestration, Automation and Response: Future Research Directions. In Advances in Artificial Intelligence and Data Engineering (pp. 147-156). Springer. https://link.springer.com/chapter/10.1007/978-981-16-1335-7_14

[101] Andrade, E., Nogueira, B., Matos, R., Callou, G., & Maciel, P. (2017). Availability modeling and analysis of a disaster-recovery-as-a-service solution. Computing, 99(9), 929–954. https://link.springer.com/content/pdf/10.1007/s00607-017-0539-8.pdf