Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Ανασκόπηση της Ευρωστίας και της Ιδιωτικότητας συστημάτων Τεχνητής Νοημοσύνης με ανάλυση περιπτώσεων χρήσης σε κρίσιμα πεδία της έρευνας και των επιχειρήσεων

Διπλωματικη Εργασια

του

ΚΟΥΤΡΟΥΜΠΑΣ ΑΘΑΝΑΣΙΟΣ

Επιβλέπων: Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Ανασκόπηση της Ευρωστίας και της Ιδιωτικότητας συστημάτων Τεχνητής Νοημοσύνης με ανάλυση περιπτώσεων χρήσης σε κρίσιμα πεδία της έρευνας και των επιχειρήσεων

Διπλωματικη Εργασια

του

ΚΟΥΤΡΟΥΜΠΑΣ ΑΘΑΝΑΣΙΟΣ

Επιβλέπων: Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Μαρτίου 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Ευάγγελος Μαρινάκης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

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

(Υπογραφή)

.....
Αθανάσιος Κουτρούμπας

14 Μαρτίου 2024

Περίληψη

Η ραγδαία ανάπτυξη των της τεχνητής νοημοσύνης (AI), ειδικότερα τεχνικών μηχανικής μάθησης (ML) και βαθιάς μάθησης, έχει δημιουργήσει μοντέλα με εξαιρετικές επιδόσεις σε πληθώρα εργασιών, και για αυτό χρησιμοποιούνται ευρέως σε καθημερινές εφαρμογές και σε εργασίες όπως αναγνώριση και κατηγοριοποίηση εικόνων, ανίχνευση αντικειμένων ή αναγνώριση ψηφίων. Ακόμη, η χρήση του AI έχει αυξηθεί και επεκταθεί πέρα από τις κλασσικές εφαρμογές της σε τομείς της πληροφορικής, διεισδύοντας σε πιο κλασσικές και βαριές βιομηχανίες, όπως στο πεδίο της αυτοκινητοβιομηχανίας, για ανάπτυξη αυτόνομων συστημάτων πλοήγησης, στο πεδίο της ενέργειας για πρόβλεψη ενεργειακής ζήτησης, αλλά και σε πεδία όπως η υγεία για παροχή ιατρικής διάγνωσης. Παρά την ευρεία χρήση, αναδύονται ζητήματα ασφάλειας και ρίσκου όταν γίνεται εφαρμογή τέτοιων τεχνικών ειδικά σε κρίσιμα πεδία κυρίως λόγω της ευπάθειας των ML μοντέλων σε προσεκτικά κατασκευασμένα δείγματα εισόδου που περιέχουν διαταραχές, τα οποία ονομάζονται ανταγωνιστικά παραδείγματα και έχουν τη δυνατότητα να προκαλέσουν δυσλειτουργία στο σύστημα και στην απόφαση αυτού. Αντίστοιχες ανησυχίες υπάρχουν και για την ιδιωτικότητα των δεδομένων με τα οποία εκπαιδεύονται τα ML μοντέλα, ειδικά όταν αυτά περιέχουν ευαίσθητα προσωπικά δεδομένα (π.χ. ιατρικό ιστορικό) και το κατά πόσο μπορεί να διατηρηθεί όταν τα μοντέλα έχουν αποδειχθεί ότι απομνημονεύουν πολλά δεδομένα και μπορούν να διαρρεύσουν πληροφορίες για αυτά.

Στόχος της διπλωματικής εργασίας είναι η ανασκόπηση της ευρωστίας και της ιδιωτικότητας συστημάτων τεχνητής νοημοσύνης με ανάλυση περιπτώσεων χρήσης σε 3 κρίσιμα πεδία της έρευνας και των επιχειρήσεων: των οχημάτων και αυτόνομης οδήγησης, της υγείας και ιατρικής διάγνωσης, και της ηλεκτρικής ενέργειας και των έξυπνων δικτύων. Αυτό γίνεται με ταξινόμηση και ανάλυση των τρεχόντων και δημοφιλέστερων ανταγωνιστικών επιθέσεων και αμυνών με αξιολόγηση των κινδύνων, προστασιών και επιπτώσεων για το κάθε πεδίο. Αντίστοιχα, γίνεται ταξινόμηση και ανάλυση των επιθέσεων και προστασιών της ιδιωτικότητας των μοντέλων και δεδομένων με αξιολόγηση εφαρμογής των τρεχόντων τεχνικών στο κάθε πεδίο. Παρουσιάζονται επίσης τα προβλήματα και οι συμβιβασμοί που προκύπτουν με την προσπάθεια εφαρμογής τεχνικών για τη βελτίωση της ευρωστίας ή της προστασίας της ιδιωτικότητας σε ML μοντέλα, και γίνεται αναφορά σε μελλοντικές προεκτάσεις για την έρευνα και υλοποίηση μεθόδων ενίσχυσης ευρωστίας και ιδιωτικότητας σε κρίσιμα πεδία και μη.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Βαθιά Μάθηση, Νευρωνικά Δίκτυα, Ανταγωνιστικά Παραδείγματα, Ανταγωνιστική Μηχανική Μάθηση, Ασφάλεια, Ευρωστία, Ιδιωτικότητα, Differential Privacy, Federated Learning, Homomorphic Encryption, Secure Multi-party Computation, Οχήματα, Αυτόνομη Οδήγηση, Υγεία, Διάγνωση, Ιατρικές Εικόνες, Ενέργεια, Έξυπνα Δίκτυα

Abstract

The rapid development of artificial intelligence (AI) and particularly machine learning (ML) and deep learning techniques has led to the creation of models with exceptional performance across a multitude of tasks, widely used in everyday applications such as image recognition, object detection, and digit recognition. Moreover, the use of AI has expanded beyond its classical applications into heavy industries, such as the automotive sector for developing autonomous navigation systems, the energy sector for energy demand prediction, and healthcare for medical diagnosis. Despite their widespread use, issues of security and risk arise when applying such techniques, especially in critical fields, mainly due to the vulnerability of ML models to carefully crafted input samples containing perturbations, known as adversarial examples, capable of causing system malfunction and decision errors. Similar concerns exist for data privacy when training ML models, especially when they involve sensitive personal data, and the extent to which privacy can be maintained when models have been shown to memorize extensive data and potentially leak information.

The aim of the thesis is to review the robustness and privacy of artificial intelligence systems through case studies in three critical research and business fields: autonomous vehicles, healthcare and medical diagnosis, and electric energy and smart grids. This is achieved by classifying and analyzing current and popular adversarial attacks and defenses with risk assessment and impact evaluation for each field. Similarly, attacks and defenses on the privacy of models and data are classified and analyzed, with an evaluation of the application of current techniques in each field. The problems and trade-offs arising from the attempt to apply techniques to improve robustness or privacy protection in ML models are also presented, with reference to future extensions for research and implementation of methods to enhance robustness and privacy in critical and non-critical fields.

Keywords

Machine Learning, Deep Learning, Neural Networks, Adversarial Examples, Adversarial Machine Learning, Security, Robustness, Privacy, Differential Privacy, Federated Learning, Homomorphic Encryption, Secure Multi-party Computation, Vehicles, Autonomous Driving, Healthcare, Medical Diagnosis, Medical Images, Energy, Smart Grids

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Ασκούνη για την ανάθεση και επίβλεψη αυτής της διπλωματικής εργασίας. Επίσης, θέλω να ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Σωτήρη Πελέκη για την καθοδήγησή του, τη βοήθεια του, τις χρήσιμες συμβουλές του και την εξαιρετική συνεργασία που είχαμε. Τέλος, θα ήθελα να ευχαριστήσω θερμά τους γονείς μου, την αδερφή μου, την κοπέλα μου, και τους φίλους μου για την υπομονή και τη συμπαράστασή τους όλα αυτά τα χρόνια των σπουδών και ειδικότερα κατά την εκπόνηση της παρούσας διπλωματικής.

Αθήνα, Μάρτιος 2024

Αθανάσιος Κουτρούμπας

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Ασφάλεια Συστημάτων Μηχανικής Μάθησης και Προκλήσεις

Τα τελευταία χρόνια η χρήση τεχνητής νοημοσύνης (AI) και ειδικότερα των τεχνικών μηχανικής μάθησης (ML) έχει αυξηθεί και επεκταθεί πέρα από τις κλασσικές εφαρμογές της σε τομείς της πληροφορικής, διεισδύοντας σε πιο κλασσικές και βαριές βιομηχανίες. Στο πεδίο της αυτοκινητοβιομηχανίας, η χρήση AI έχει παίξει ρόλο στην ανάπτυξη αυτόνομων συστημάτων πλοήγησης, στο πεδίο της ενέργειας έχει ευρεία εφαρμογή στην ακριβέστερη πρόβλεψη ενεργειακής ζήτησης, αλλά και σε πεδίο όπως στην υγεία αξιοποιείται για την διάγνωση και φροντίδα των ασθενών. Στην ευρωπαϊκή ένωση (ΕΕ), σύμφωνα με έρευνα της Eurostat το 2021[1], το 28% των μεγαλύτερων επιχειρήσεων της ΕΕ κάνουν χρήση AI τεχνολογιών, με μόλις το 9% των επιχειρήσεων σε βαριές βιομηχανίες όπως την παροχής ρεύματος, φυσικού αερίου, ατμού, κλιματισμού και νερού κάνουν χρήση του AI. Όμως, σε παγκόσμιο επίπεδο, σύμφωνα με την δημοσκόπηση της IBM μεταξύ των παγκόσμιων ανώτερων στελεχών της πληροφορικής για την υιοθέτηση του AI το 2023[2], σε βιομηχανίες όπως των αυτοκινήτων, της ενέργεια, και της υγείας, πάνω από το 23% χρησιμοποιεί ήδη ενεργά AI στις επιχειρηματικές λειτουργίες τους και πάνω από το 44% εξερευνά τη χρήση του, όπου συγκεκριμένα στον κλάδο των αυτοκινήτων, το 73% των οργανισμών έχει επιταχύνει την επένδυση της χρήσης AI τους τελευταίους 24 μήνες.

Όμως, ενώ η προσοχή είναι στραμμένη στην ανάπτυξη ML μοντέλων και αλγορίθμων που θα είναι πιο ακριβείς και θα έχουν μεγαλύτερη εφαρμογή, αναπτύσσοντας νέες τεχνικές και αξιοποιώντας μεγαλύτερο όγκο δεδομένων, υπάρχουν ζητήματα ασφαλείας και ρίσκου όταν γίνεται εφαρμογή τέτοιων τεχνικών ειδικά σε κρίσιμα πεδία στα οποία ένα λάθος ή μια δυσλειτουργία μπορεί να προκαλέσει οικονομικές, υλικές ή επικίνδυνες για την ανθρώπινη ζωή ζημιές. Παρά την εντυπωσιακή απόδοση των ML αλγορίθμων και ειδικότερα των τεχνικών βαθιάς μάθησης (DL), πολλές πρόσφατες μελέτες έχουν εγείρει ανησυχίες σχετικά με την ασφάλεια και ευρωστία (robustness) των ML μοντέλων, τα οποία έχουν αποδειχθεί ότι είναι εν γένει ευάλωτα σε προσεκτικά κατασκευασμένα δείγματα εισόδου τα οποία περιέχουν διαταραχές (απαρατήρητες στο ανθρώπινο μάτι σε περιπτώσεις εικόνων), τα οποία ονομάζεται ανταγωνιστικά παραδείγματα (adversarial examples) και έχουν τη δυνατότητα να προκαλέσουν δυσλειτουργία στο σύστημα και στην απόφαση αυτού [1]. Αντίστοιχες ανησυχίες υπάρχουν για την ιδιωτικότητα των δεδομένων με τα οποία εκπαιδεύονται τα ML μοντέλα, ειδικά όταν

[1] https://ec.europa.eu/eurostat/statistics-explained

[2] https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters

αυτά περιέχουν ευαίσθητα προσωπικά δεδομένα (π.χ. ιατρικό ιστορικό) και το κατά πόσο μπορεί να διατηρηθεί όταν τα μοντέλα έχουν αποδειχθεί ότι απομνημονεύουν πολλά δεδομένα και μπορούν να διαρρεύσουν πληροφορίες για αυτά [2].

Λόγω αυτών των ζητημάτων, η έρευνα χρήσης AI σε κρίσιμους τομείς, πέρα από την επίδοση και την ακρίβεια, εστιάζει και στους τομείς ασφάλειας, όπως την ευρωστία των συστημάτων, δηλαδή τη δυνατότητα αντίστασης σε τέτοια κακόβουλα ή θορυβώδη δείγματα για την παραγωγή αξιόπιστων αποτελεσμάτων, αλλά και την ιδιωτικότητα των μοντέλων και των δεδομένων εκπαίδευσης και το κατά πόσο μπορεί να διατηρηθεί ή εγγυηθεί. Στην ίδια δημοσκόπηση της IBM για το 2022[3] και το 2023, το ποσοστό των οργανισμών που δεν έχουν καμία προφύλαξη από ανταγωνιστικές απειλές και πιθανές εισβολές πέφτει από το 59% στο 38% και το ποσοστό των οργανισμών που δεν έχουν καμία διαφύλαξη της ιδιωτικότητας των δεδομένων σε ολόκληρο τον κύκλο ζωής τους πέφτει από το 52% στο 44%, το οποίο δείχνει το ενδιαφέρον που αναπτύσσεται στην ανάπτυξη ασφαλών, εύρωστων και ιδιωτικών AI συστημάτων, δείχνοντας όμως ότι υπάρχει ανάγκη μεγαλύτερης προόδου.

Ειδικότερα, η ΕΕ έχει ασχοληθεί ενεργά με τον έλεγχο χρήσης του AI σε κρίσιμους τομείς και σε εφαρμογές υψηλού κινδύνου, αρχικά με τη δημοσίευση οδηγιών ηθικής για ανάπτυξη αξιόπιστων AI συστημάτων (*Ethics guidelines for trustworthy AI*) από το 2019[4], όπου μέσα στις βασικές απαιτήσεις που πρέπει να πληρούν αυτά τα συστήματα για να θεωρούνται αξιόπιστα είναι η τεχνική ευρωστία και ασφάλεια, δηλαδή πρέπει να είναι ακριβή, αξιόπιστα και επαναλήψιμα, διασφαλίζοντας ότι μια ακούσια βλάβη μπορεί να ελαχιστοποιηθεί και να προληφθεί, αλλά και η διασφάλιση της ιδιωτικότητας και προστασίας των δεδομένων. Επίσης, ακαδημαϊκή έρευνα της ΕΕ το 2020 [3], συνέβαλε στο κίνημα δημιουργίας ρυθμιστού πλαισίου για τη χρήση του AI, παρέχοντας μια αντικειμενική άποψη για το τρέχον τοπίο του AI και ML, αναφέροντας τους σχετικούς τεχνικούς κινδύνους και τους περιορισμούς, σε αναφορά με υπάρχουσες ρυθμίσεις για την ασφάλεια στον κυβερνοχώρο των ψηφιακών συστημάτων και την προστασία δεδομένων (π.χ. GDPR), δίνοντας έμφαση στην καθιέρωση μεθοδολογιών για την αξιολόγηση της ευρωστίας των αυτών συστημάτων. Έπειτα, το 2023 η ΕΕ εισάγει τον πρώτο ολοκληρωμένο νόμο, για το AI (*The AI Act*)[5], ο οποίος κατατάσσει τις εφαρμογές του AI σε 3 κατηγορίες κινδύνου και επιτρέπει ή απαγορεύει αντίστοιχα τη χρήση του: (i) Απαγόρευση χρήσης σε εφαρμογές και συστήματα τα οποία δημιουργούν απαράδεκτο κίνδυνο (π.χ. συστήματα κοινωνικής βαθμολόγησης, συστήματα ταξινόμησης βιομετρικών χαρακτηριστικών), (ii) Ελεγχόμενη χρήση σε εφαρμογές και συστήματα υψηλού κινδύνου (π.χ. deepfakes, εργαλεία σάρωσης και κατάταξης βιογραφικών), (iii) Εφαρμογές με ελάχιστο ρίσκο που δεν υπόκειται σε ρύθμιση (π.χ. βιντεοπαιχνίδια με χρήση AI, φίλτρα ανεπιθύμητης αλληλογραφίας). Στη 2η κατηγορία χρήσης, η οποία υπόκειται σε ρύθμιση, κάνει αναφορά για σχεδιασμό AI συστημάτων υψηλού ρίσκου τα οποία πρέπει να επιτυγχάνουν τα επιθυμητά επίπεδα ακρίβειας, ευρωστίας και κυβερνοασφάλειας. Επίσης, σε αυτήν την κατηγορία υψηλού κινδύνου ανήκουν και εφαρμογές του AI σε υποδομές ζωτικής σημασίας, όπως οδική κυκλοφορία, ηλεκτρική ενέργεια, παροχή νερού, φυσικού αερίου και θέρμανσης.

Με αυτές τις τρέχουσες συνθήκες, κρίνεται επιτακτική η ανάπτυξη τεχνικών οι οποίοι

[3] https://www.ibm.com/watson/resources/ai-adoption
[4] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[5] https://artificialintelligenceact.eu

ενισχύουν την ασφάλεια, ευρωστία και ιδιωτικότητα των ML μοντέλων, αλλά και οι μέθοδοι αξιόπιστης αξιολόγησης αυτών, ώστε να μπορέσει να επεκταθεί η χρήση του AI σε κρίσιμους τομείς με ασφάλεια, χωρίς οι συνέπειες μιας λανθασμένης εκτίμησης να είναι σοβαρές.

1.2 Αντικείμενο της διπλωματικής

Το αντικείμενο της συγκεκριμένης διπλωματικής είναι η ανασκόπηση και παράθεση των τρεχόντων ερευνητικών εξελίξεων στο πεδίο της ευρωστίας και ιδιωτικότητας συστημάτων τεχνητής νοημοσύνης, παρέχοντας το θεωρητικό υπόβαθρο για την αξιολόγηση αυτών σε μοντέλα μηχανικής μάθησης, ειδικότερα για κρίσιμα πεδία της έρευνας και των επιχειρήσεων. Συγκεκριμένα, γίνεται επεξήγηση και ανάλυση της ευρωστίας και ιδιωτικότητας των συστημάτων τεχνητής νοημοσύνης, και η ανάλυση χρήσης τους σε 3 κρίσιμα πεδία: πεδίο των οχημάτων και αυτόνομης οδήγησης, πεδίο της υγείας και ιατρικής διάγνωσης, και πεδίο της ηλεκτρικής ενέργειας και των έξυπνων δικτύων. Γίνεται συστηματική έρευνα και μελέτη της διαθέσιμης βιβλιογραφίας για τις πιθανές ανταγωνιστικές επιθέσεις και επιθέσεις ιδιωτικότητας που εμφανίζονται σε ML μοντέλα στα κρίσιμα αυτά πεδία, και αντίστοιχα για τις διαθέσιμες τεχνικές άμυνας για την προστασία από τέτοιες επιθέσεις και την ενίσχυση της ευρωστίας και της ιδιωτικότητας των ML μοντέλων σε κάθε πεδίο, όπως και μεθόδους αξιολόγησης αυτών των μεθόδων. Από το μεγάλο πεδίο έρευνας και την πληθώρα δημοσιεύσεων που είναι διαθέσιμες, γίνεται διαλογή των πιο διαδεδομένων και ισχυρών επιθέσεων και αμυνών, και παρουσιάζεται συγκεντρωτικά η μελέτη ευρωστίας για το κάθε πεδίο ξεχωριστά με ανάλυση εφαρμογής στο καθένα και δίνοντας παράλληλα το απαιτούμενο θεωρητικό και τεχνικό υπόβαθρο για την κάθε μέθοδο. Παρουσιάζονται επίσης, εργαλεία και βιβλιοθήκες ανοιχτού κώδικα οι οποίες είναι διαθέσιμες για την αξιολόγηση και ενίσχυση της ευρωστίας ή της ιδιωτικότητας των ML μοντέλων. Αυτή η ανάλυση και καταγραφή προβλημάτων, λύσεων και τρόπων αξιολόγησης μπορεί να προσφερθεί ως χρήσιμο έργο ανασκόπησης της ευρωστίας των AI/ML συστημάτων, σε πολλαπλά ενδιαφερόμενα μέρη (stakeholders) για κάθε πεδίο, όπως ερευνητές τεχνητής νοημοσύνης, ML μηχανικών και επαγγελματιών, καθώς και ειδικούς από το κάθε πεδίο, π.χ. μηχανικούς, γιατρούς.

1.3 Μεθοδολογία

1.3.1 Πηγές Αναζήτησης Βιβλιογραφίας

Για τη συγγραφή αυτής της εργασίας αναζητήθηκαν εργασίες, άρθρα και δημοσιεύσεις από αναγνωρισμένα επιστημονικά περιοδικά και συνέδρια πολλαπλών επιστημονικών πεδίων, λόγω της ανασκόπησης σε διάφορα πεδία έρευνας, όπως: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Ασφάλεια Υπολογιστών, Ιατρική, Ενέργεια και Μεταφορές. Η τελική βιβλιογραφία περιέχει δημοσιεύσεις από έγκυρα και αναγνωρισμένα περιοδικά ανά τομέα, όπως π.χ. το Journal of Machine Learning Research (JMLR) στον τομέα της Μηχανικής Μάθησης, το IEEE Security & Privacy στον τομέα της Ασφάλειας, περιοδικά του Nature όπως το Scientific Reports στον τομέα της Ιατρικής, περιοδικά του Elsevier όπως το Energy Reports στον τομέα της Ενέργειας και το IEEE Transactions on Intelligent Transportation Systems στον τομέα της Μετακίνησης. Πιο συγκεκριμένα στον πίνακα 1.1 παραθέτονται ενδεικτικά τα κύρια περιοδικά αλλά και συνέδρια στα οποία είναι δημοσιευμένες οι εργασίες της βιβλιογραφίας, ανά

επιστημονικό πεδίο.

Πεδία	Περιοδικό (Εκδότης)	Συνέδριο (Εκδότης)
AI/ML	Trans. on Neural Networks and Learning Systems (IEEE)	CVPR (IEEE)
AI/ML	Pattern Recognition (Elsevier)	ICCV (IEEE)
AI/ML	Algorithms (MDPI)	NeurIPS
AI/ML	Found. and Trends in Machine Learning (Now Publishers)	ICLR
AI/ML	Journal of Machine Learning Research (JMLR)	ICML
Security/Privacy	Security & Privacy (IEEE)	SP (IEEE)
Security/Privacy	Comput. Surv. (ACM)	Security (USENIX)
Security/Privacy	Trans. on Services Computing (IEEE)	CCS (ACM)
Οδήγηση	Trans. on Intelligent Transportation Systems (IEEE)	IV (IEEE)
Οδήγηση	Trans. on Industrial Informatics (IEEE)	PerCom (IEEE)
Οδήγηση	Journal of Systems Architecture (Elsevier)	CVPR (IEEE)
Υγεία	Scientific Reports (Nature)	PRIME (Springer)
Υγεία	npj Digital Medicine (Nature)	MICCAI (Springer)
Υγεία	Medical Image Analysis (Elsevier)	MLMI (Springer)
Ενέργεια	Trans. on Smart Grid (IEEE)	SmartGridComm (IEEE)
Ενέργεια	Smart Cities (MDPI)	ISGT (IEEE)
Ενέργεια	Renewable and Sustainable Energy Reviews (Elsevier)	PES (IEEE)

Πίνακας 1.1: *Πίνακας με αναφορά μερικών επιστημονικών περιοδικών και συνεδρίων στα οποία είναι δημοσιευμένες πολλές από τις επιλεγμένες δημοσιεύσεις της βιβλιογραφίας αυτής της διπλωματικής εργασίας, ανά επιστημονικό πεδίο.*

1.3.2 Αναζήτηση Βιβλιογραφίας

Για την αναζήτηση των δημοσιεύσεων, χρησιμοποιήθηκαν εργαλεία όπως το *Google Scholar* για την αναζήτηση από διάφορες πηγές, το *Science Direct* για την αναζήτηση και πλοήγηση πηγών δημοσιευμένων σε περιοδικά και συνέδρια από τον εκδοτικό οίκο Elsevier, και αντίστοιχα το *IEEE Xplore* για την αναζήτηση και πλοήγηση πηγών δημοσιευμένων σε περιοδικά και συνέδρια από τον οργανισμό IEEE. Για να βρεθούν δημοσιεύσεις που αφορούν τη θεματολογία της εργασίας, χρησιμοποιήθηκαν διάφορες λέξεις 'κλειδιά' για να καλυφθεί όλο το πιθανό φάσμα έρευνας γύρω από την αντίπαλη μηχανική μάθηση, την ευρωστία και ιδιωτικότητα συστημάτων τεχνητής νοημοσύνης στα πεδία που εξετάστηκαν. Για να ερευνηθεί γενικότερα το πεδίο των αντίπαλων επιθέσεων και των εύρωστων συστημάτων τεχνητής νοημοσύνης, χρησιμοποιήθηκαν λέξεις κλειδιά, όπως: 'adversarial machine learning', 'adversarial examples', 'adversarial attacks', 'adversarial perturbations', 'adversarial robustness', 'adversarial learning', 'adversarial defenses', 'privacy machine learning'. Για να ερευνηθεί η χρήση των παραπάνω στα πεδία που καλύφθηκαν, έγινε συνδυασμός αναζήτησης των παραπάνω με τους ερευνητικά και επιχειρησιακά πεδία, όπως: 'energy', 'smart grids', 'healtchare', 'medical imaging', 'autonomous vehicles', 'autonomous driving'.

1.3.3 Κριτήρια Επιλογής Βιβλιογραφίας

Αρχικά, επιλέχτηκαν ως βασικές πηγές για κάθε πεδίο, δημοσιεύσεις ανασκόπησης (Review Papers), αλλά και εργασίες συστηματοποίηση γνώσης (SoK: Systemization of Knowledge), καθώς τέτοιες πηγές ενοποιούν υπάρχουσα γνώση πάνω σε ένα θέμα, προσφέρουν περιεκτικές κριτικές, εκτελούν αξιολόγηση και προσφέρουν νέες οπτικές ή ιδέες πάνω στο θέμα το

οποίο ερευνούν, οπότε πρόκειται για ιδανικές πηγές για να υπάρχει μια ολοκληρωμένη άποψη γύρω από ένα θέμα, συγκεκριμένα για την ευρωστία και ιδιωτικότητα συστημάτων τεχνητής νοημοσύνης σε διάφορα πεδία.

Επειδή είναι μεγάλο το εύρος της εργασίας και υπάρχει μεγάλος αριθμός εργασιών, διευκρινίζονται τα κριτήρια επιλογής των δημοσιεύσεων:

- *Μέσο δημοσίευσης*: Επιλέγονται άρθρα δημοσιευμένα σε διεθνή, και αναγνωρισμένα επιστημονικά περιοδικά και εδραιωμένα συνέδρια τα όποια έχουν χρόνια παρουσίας (βλ. πίνακα 1.1) στην αγγλική γλώσσα.
- *Ημερομηνία συγγραφής*: Επειδή πρόκειται για σχετικά πρόσφατο τομέα έρευνας στην τεχνητή νοημοσύνη, δεν υπάρχει άμεσα σχετιζόμενη βιβλιογραφία πριν το 2014, και ειδικότερα για έρευνα ευρωστίας στα εξεταζόμενα πεδία πριν το 2018. Για αυτό επιλέγονται κυρίως πρόσφατες έρευνες (από 2020 και μετά), όμως επιλέγονται και κάποιες παλιότερες και εδραιωμένες δημοσιεύσεις.
- *Θεματολογία*: Επιλέγονται δημοσιεύσεις που σχετίζονται κυρίως με τις παραπάνω λέξεις 'κλειδιά' που αναφέραμε. Πιο συγκεκριμένα, εστιάζουμε σε μοντέλα μηχανικής μάθησης και βαθιάς μάθησης, καθώς δείχνουν να έχουν την μεγαλύτερη εφαρμογή. Επίσης, στο κομμάτι της ασφάλειας σε κάθε πεδίο, επιλέγονται έργα τα οποία αναφέρονται σε αντίπαλες επιθέσεις και ενίσχυση ασφάλειας των συστημάτων τεχνητής νοημοσύνης ως προς αυτές, και όχι αυτές που αναφέρονται στη γενικότερη ασφάλεια των συστημάτων από όλο το φάσμα των πιθανών κυβερνοεπιθέσεων.
- *Συγγραφείς*: Επιλέγονται δημοσιεύσεις που έχουν παραπάνω από 1 συγγραφέα και προέρχονται από πανεπιστήμια ή ερευνητικά ινστιτούτα, ιδανικά σχετιζόμενα με τα πεδία που σχετίζεται η εργασία.
- *Αναφορές*: Επιλέγονται δημοσιεύσεις που διαθέτουν ένα ικανοποιητικό αριθμό αναφορών, ο οποίος εξαρτάται από το πόσο πρόσφατη είναι ή το πόσο ειδικευμένη είναι. Για πιο παλιές και γενικές δημοσιεύσεις ιδανικά υπάρχουν πάνω από 100-150 αναφορές, ενώ για πιο καινούριες ή ειδικευμένες έρευνες μπορεί να είναι ιδανικά πάνω από 10-20.

1.3.4 Μεθοδολογία Συλλογής Βιβλιογραφίας και Συγγραφής

Για τη συγκέντρωση της βιβλιογραφίας και της συγγραφής της διπλωματικής εργασίας, ακολουθήθηκε η εξής μεθοδολογία. Πρώτα έγινε γενική μελέτη των αντίπαλων παραδειγμάτων, επιθέσεων, αμυνών, τεχνικών ενίσχυσης ευρωστίας και ιδιωτικότητας στο πεδίο της τεχνητής νοημοσύνης, η οποία χρησιμοποιήθηκε για τη συγγραφή των αρχικών κεφαλαίων όπου δίνεται το θεωρητικό υπόβαθρο για την κατανόηση αυτών των όρων. Συνολικά, γίνεται αναφορά σε πάνω από 80 έργα στον τομέα της αντίπαλης μηχανικής μάθησης και ευρωστίας, και πάνω από 40 στον τομέα της ιδιωτικότητας των συστημάτων τεχνητής νοημοσύνης και μηχανικής μάθησης.

Έπειτα, με αφετηρία τις δημοσιεύσεις ανασκόπησης, έγινε μελέτη της εφαρμογής των παραπάνω στα εξεταζόμενα πεδία. Από τις παραπομπές αυτών των δημοσιεύσεων, αλλά και με ανεξάρτητη αναζήτηση σύμφωνα με τα προαναφερθέντα εργαλεία και κριτήρια, επιλέχθηκαν έργα τα οποία παρουσιάζουν τεχνικές επιθέσεων και αμυνών στα πεδία αυτά. Μελετήθηκαν πάνω από 50 δημοσιεύσεις σε κάθε πεδίο, δηλαδή συνολικά πάνω από 150 δημοσιεύσεις, κάποιες αναλυτικά και άλλες συνοπτικά. Τελικά, σε κάθε πεδίο, αναλύονται πάνω από 20 τεχνικές

επίθεσης και άμυνας, δηλαδή συνολικά πάνω από 60 δημοσιεύσεις, για την κατανόηση των κινδύνων των αντιπάλων επιθέσεων σε κάθε πεδίο και τις τεχνικές για ενίσχυση της ευρωστίας και της ιδιωτικότητα των συστημάτων τεχνητής νοημοσύνης, φτάνοντας σε συνολικό αριθμό.

1.4 Σχετική Έρευνα και Συμβολή της Εργασίας

Η συμβολή της παρούσας εργασίας, είναι η συγκεντρωτική, αναλυτική παρουσίαση και ταξινόμηση των τεχνικών επίτευξης και αξιολόγηση ευρωστίας και ιδιωτικότητας για ML μοντέλα σε 3 διαφορετικά και ξεχωριστά κρίσιμα πεδία της βιομηχανίας και των επιχειρήσεων, όπου παρουσιάζονται οι ξεχωριστές ιδιαιτερότητες του κάθε πεδίου, αλλά και αναδεικνύονται οι κοινοί κίνδυνοι και τρόποι αντιμετώπισης.

Οι περισσότερες εργασίες ανασκόπησης σχετικές με το θέμα επικεντρώνονται είτε αυστηρά στους κινδύνους των ανταγωνιστικών επιθέσεων και στους τρόπους αύξησης ευρωστίας χωρίς όμως αναφορά σε κάποιο συγκεκριμένο πεδίο (βλ. [4] όπου γίνεται διερεύνηση των κινδύνων και των μεθόδων διατήρησης της ασφάλειας και της ιδιωτικότητας στη μηχανική μάθηση, και [5] όπου γίνεται έρευνα των πιο πρόσφατων και επικρατέστερων μεθόδων στον τομέα της ανταγωνιστικής μάθησης για εργασίας ταξινόμησης εικόνων), είτε με αναφορά σε ένα μόνο αποκλειστικά πεδίο (βλ. [6] όπου γίνεται έρευνα ασφαλής και εύρωστης μηχανική μάθηση στην υγεία, και [7] όπου γίνεται ανάλυση ανταγωνιστικών επιθέσεων και αμυνών σε μοντέλα αυτόνομης οδήγησης).

Η εργασία αυτή θα μπορούσε να χρησιμεύσει ως μια περιεκτική σύνθεση της υπάρχουσας έρευνας και βιβλιογραφίας για τους κινδύνους των ανταγωνιστικών επιθέσεων, τις τεχνικές αύξησης ευρωστίας και ιδιωτικότητας, παρέχοντας τις απαραίτητες πληροφορίες για μετέπειτα μελλοντική έρευνα, εστιάζοντας σε ένα τομέα ή γνωστικό αντικείμενο με σκοπό τη δημιουργία και εξασφάλιση, ασφαλών και αξιόπιστων AI συστημάτων, καθώς πρόκειται για μια κατεύθυνση που έχει ακόμα χαμηλή αξιοποίηση σε επιχειρηματικούς τομείς σε σχέση με την υπάρχουσα ακαδημαϊκή έρευνα.

1.5 Οργάνωση Κειμένου

Η παρούσα διπλωματική εργασία έχει οργανωθεί ως εξής:

- Στο **Κεφάλαιο 1**, γίνεται η εισαγωγή στην εργασία, δίνοντας μια πλήρη εικόνα του σκοπού, του αντικείμενου, της συμβολής της εργασίας καθώς και τη μεθοδολογία που ακολουθήθηκε και την οργάνωσή της.
- Στο **Κεφάλαιο 2**, δίνεται το θεωρητικό υπόβαθρο των θεματικών που σχετίζονται με τη διπλωματική, όπως για τη μηχανική μάθηση, τα νευρωνικά δίκτυα, τη βαθιά μάθηση, και την ασφάλεια απλών συστημάτων και συστημάτων τεχνητής νοημοσύνης.
- Στο **Κεφάλαιο 3**, γίνεται ανάλυση του πεδίου της ανταγωνιστικής μηχανικής μάθησης, όπου παρουσιάζονται και αναλύονται ανταγωνιστικές επιθέσεις και άμυνες για ML μοντέλα.
- Στο **Κεφάλαιο 4**, γίνεται ανάλυση της ευρωστίας των ML μοντέλων, παρουσιάζοντας τα διαφορετικά είδη ευρωστίας, τους τρόπους αξιολόγησης της, τους συμβιβασμούς που προκύπτουν και τα διαθέσιμα εργαλεία για την ενίσχυση και αξιολόγηση της.
- Στο **Κεφάλαιο 5**, γίνεται ανάλυση του πεδίου της ιδιωτικότητας μοντέλου και δεδομένων στη μηχανική μάθηση, όπου παρουσιάζονται και αναλύονται επιθέσεις ιδιωτικότη-

τας σε ML μοντέλα, μέθοδοι προστασίας για κάθε επίθεση αλλά και γενικές μέθοδοι εγγυημένης προστασίας της ιδιωτικότητας των δεδομένων.

- Στο **Κεφάλαιο 6**, γίνεται η μελέτη ευρωστίας ανά κρίσιμο πεδίο, όπου παρουσιάζονται και ταξινομούνται οι εφαρμόσιμες επιθέσεις, οι τρόποι αντιμετώπισης και τα συμπεράσματα αξιολόγησης στο καθένα.
- Στο **Κεφάλαιο 7**, γίνεται σύνοψη των κύριων συμπερασμάτων που προέκυψαν από την εργασία, αναδεικνύοντας την ανάγκη ενίσχυσης της ευρωστίας των συστημάτων τεχνητής νοημοσύνης σε κρίσιμους τομείς, καθώς και ανοιχτά ζητήματα και μελλοντικές κατευθύνσεις για τους τομείς αυτούς.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό εισάγονται οι απαραίτητες έννοιες για τα επόμενα κεφάλαια ορίζοντας βασικές έννοιες της τεχνητής νοημοσύνης (AI), της μηχανικής μάθησης (ML), και των νευρωνικών δικτύων (NN), καθώς και βασικές έννοιες της ασφάλειας με βασική περιγραφή του πεδίου της ασφάλειας συστημάτων τεχνητής νοημοσύνης.

2.1 Μηχανική Μάθηση και Νευρωνικά Δίκτυα

2.1.1 Τεχνητή Νοημοσύνη

Η *Τεχνητή Νοημοσύνη (AI)*, αντιπροσωπεύει την προσπάθεια των υπολογιστικών συστημάτων να αναπαράγουν ανθρώπινες γνωστικές διαδικασίες, αναλύοντας πληροφορία και χρησιμοποιώντας τεχνικές μάθησης. Η χρήση του AI είναι πλέον διαδεδομένη σε πολλούς βιομηχανικούς τομείς, με εφαρμογές σε αυτό-οδηγούμενα αυτοκίνητα, πρόβλεψη ενεργειακής ζήτησης, ανακάλυψη φαρμάκων, χρηματιστηριακές επενδύσεις, και έχει βοηθήσει στην εξέλιξη αυτών παρέχοντας λύσεις σε σύνθετα και πολύπλοκα προβλήματα γρήγορα και αποτελεσματικά. Επίσης, η χρήση της είναι εδραιωμένη και σε καθημερινές εργασίες, όπως ψηφιακούς βοηθούς, μετάφραση και παραγωγή κειμένου, προσωποποιημένες προτάσεις περιεχομένου και πλοήγηση, βελτιώνοντας την καθημερινότητα την καθημερινότητα πολλών ανθρώπων και επηρεάζοντας τον σύγχρονο τρόπο ζωής.

Το AI είναι ένας όρος 'ομπρέλα', στον οποίο ανήκουν πολλές τεχνικές, μοντέλα και αλγόριθμοι. Παρακάτω και σε όλη την εργασία εστιάζουμε σε μία κατηγορία, τη *Μηχανική Μάθηση (ML)* και κυρίως στην υποκατηγορία της *Βαθιάς Μάθησης (Deep Learning)*, που αποτελεί το κομμάτι του AI με τη μεγαλύτερη εφαρμογή και τις καλύτερες επιδόσεις μέχρι σήμερα.

2.1.2 Μηχανική Μάθηση

Ο όρος *Μηχανική Μάθηση (ML)*, αντιπροσωπεύει ένα ολόκληρο πεδίο το οποίο δίνει τη δυνατότητα στους υπολογιστές να μαθαίνουν από δεδομένα χωρίς να είναι ρητά προγραμματισμένοι [8]. Ο στόχος των ML αλγορίθμων είναι να μάθουν να εκτελούν ορισμένες εργασίες γενικεύοντας από τα δεδομένα, όπως την παροχή ακριβών προβλέψεων και αποφάσεων ή την εύρεση δομών και εξαγωγή χρήσιμων πληροφοριών από δεδομένα.

Η βασική είσοδος των ML αλγορίθμων είναι τα δεδομένα, τα οποία αντιπροσωπεύονται ως ένα σύνολο δειγμάτων, όπου το κάθε δείγμα περιέχει ένα σύνολο τιμών *χαρακτηριστικών (features)*, π.χ. σε μια φωτογραφία 100x100 pixel, όπου κάθε pixel αντιπροσωπεύεται από έναν αριθμό (0-255). Στο παράδειγμα αυτό, όλες αυτές οι τιμές μπορούν να συγκεντρωθούν για να σχηματίσουν ένα διάνυσμα μήκους 10.000 το οποίο ονομάζεται *διάνυσμα χαρακτηριστικών (feature vector)*, και κάθε φωτογραφία που αναπαρίσταται ως διάνυσμα χαρακτηριστικών,

μπορεί να συσχετιστεί με μία ετικέτα (label), π.χ. το όνομα ενός ατόμου στη φωτογραφία.

2.1.2.1 Εκπαίδευση και Εξαγωγή Συμπερασμάτων

Ένας ML αλγόριθμος χρησιμοποιεί ένα σύνολο από πολλαπλά διανύσματα χαρακτηριστικών και τις σχετικές ετικέτες τους, το οποίο ονομάζεται *σύνολο δεδομένων εκπαίδευσης (training dataset)*, για να εκπαιδεύσει ένα ML μοντέλο ώστε όταν του παρουσιαστεί ένα νέο δείγμα να μπορεί να δώσει την προβλεπόμενη ετικέτα. Η ικανότητα ενός ML μοντέλου να προβλέπει με ακρίβεια την ετικέτα, είναι ένα μέτρο του πόσο καλά το μοντέλο αυτό γενικεύει σε δεδομένα που δεν έχει ξαναδεί, και ονομάζεται *ικανότητας γενίκευσης (generalization)* του μοντέλου. Αυτό μπορεί να μετρηθεί, κυρίως για μοντέλα ταξινόμησης, με την *ακρίβεια (accuracy)*, όπου μετριέται το ποσοστό των σωστών προβλέψεων έναντι των συνολικών, αλλά και εμπειρικά με το *σφάλμα δοκιμής (test error)*, το οποίο αντιπροσωπεύει την μέση απώλεια που σημειώνεται κατά την εκπαίδευση, και εξαρτώνται από πολλές παράγοντες, όπως την ποιότητά, την ποσότητα των δεδομένων εκπαίδευσης αλλά και των αλγόριθμων και των παραμέτρων που επιλέχθηκαν.

Γενικά, οι ML αλγόριθμοι μπορούν να χωριστούν σε 2 διαφορετικές φάσεις [8]:

- Εκπαίδευση (**Training**): Εάν ένα ML μοντέλο αναπαρασταθεί ως μια παραμετρική συνάρτηση $h_\theta(x)$ που δέχεται ως είσοδο ένα διάνυσμα χαρακτηριστικών x και έχει ως παράμετρο το διάνυσμα θ, τότε η διαδικασία της εκπαίδευσης είναι αυτή που έχει στόχο την ανάλυση των δεδομένων εισόδου για να ευρεθούν οι κατάλληλοι παράμετροι θ, οι οποίοι αντιστοιχίζουν καλύτερα την είσοδο x στην κατάλληλη έξοδο $h_\theta(x)$, συνήθως με μείωση του ρίσκου της κατάλληλης συνάρτησης στόχου (objective function). Η απόδοση του μοντέλου επιβεβαιώνεται σε ένα *σύνολο δεδομένων δοκιμής (test dataset)*, τα οποία είναι διαφορετικά από τα δεδομένα εκπαίδευσης, για να μετρηθεί η ικανότητα γενίκευσης του μοντέλου.
- Εξαγωγή Συμπερασμάτων (**Inference**): Όταν ολοκληρωθεί η εκπαίδευση ενός ML μοντέλου, τότε ακολουθεί η διαδικασία εξαγωγής συμπερασμάτων ή προβλέψεων για δεδομένα που δεν έχει ξανασυναντήσει στην εκπαίδευση, με σταθερή τιμή των παραμέτρων θ και υπολογίζοντας τιμή εξόδου $h_\theta(x)$ για κάθε νέα είσοδο x.

2.1.2.2 Τεχνικές Εκπαίδευσης

Υπάρχουν διάφοροι τρόποι να εκτελεστεί η διαδικασία της εκπαίδευσης των αλγορίθμων μηχανικής μάθησης σχετικά με το πως να μαθαίνουν να είναι πιο ακριβής στην πρόβλεψή τους. Παρακάτω παρουσιάζονται συνοπτικά οι ποιο βασικές κατηγορίες εκπαίδευσης [8].

Στην Επιβλεπόμενη Μάθηση (**Supervised Learning**), το μοντέλο μαθαίνει τη συσχέτιση μεταξύ εισόδων x και εξόδων $y = h_\theta(x)$ βασισμένο σε δεδομένα εκπαίδευσης $\mathcal{D}$ τα οποία περιέχουν τις εισόδους και τις αντίστοιχες ετικέτες που τους αντιστοιχούν $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$. Οι κύριες διεργασίες των μοντέλων επιβλεπόμενης μάθησης είναι η *Ταξινόμηση (Classification)* όπου η τιμή εξόδου είναι μια ετικέτα κλάσης, και η *Παλινδρόμηση (Regression)* όπου τιμή εξόδου είναι μια συνεχής τιμή.

Στη Μη Επιβλεπόμενη Μάθηση (**Unsupervised Learning**), δεν υπάρχουν οι ετικέτες y διαθέσιμες κατά την εκπαίδευση και το σύνολο δεδομένων εκπαίδευσης $\mathcal{D}$ αποτελείται μόνο από τις εισόδους x_i. Εφόσον δεν μπορεί να γίνει ακριβής αντιστοίχηση ή πρόβλεψη, ο στόχος συνήθως είναι η εύρεση δομής στα δεδομένα. Μια από τις πιο κοινές εργασίας των μοντέλων

μην επιβλεπόμενης μάθησης είναι η *Ομαδοποίηση (Clustering)* όπου τα δείγματα που είναι σχετικά παρόμοια ανήκουν στην ίδια ομάδα.

Στην Ημιεπιβλεπόμενη Μάθηση (**Semi-supervised Learning**), μόνο ένα μέρος των δεδομένα εκπαίδευσης περιέχουν και τις ετικέτες εξόδου, καθώς σε πραγματικές συνθήκης η επισήμανση των δεδομένων είναι μια χρονοβόρα διαδικασία. Οπότε είναι μια συνδυαστική τεχνική, όπου τα μη επισημασμένα δεδομένα χρησιμοποιούνται για την εκμάθηση αναπαραστάσεων υψηλότερου επιπέδου και στη συνέχεια χρησιμοποιούν τα επισημασμένα παραδείγματα για να καθοδηγήσουν την εργασία μεταγενέστερης μάθησης.

Στην Ενισχυτική Μάθηση (**Reinforcement Learning**), ορίζονται πράκτορες οι οποίοι κάνουν παρατηρήσεις του περιβάλλοντος τους και τις χρησιμοποιούν για να αναλάβουν ενέργειες με στόχο τη μεγιστοποίηση ενός σήματος ανταμοιβής. Στην πιο γενική διατύπωση, το σύνολο των ενεργειών δεν είναι προκαθορισμένο και οι ανταμοιβές δεν είναι απαραίτητα άμεσες, αλλά μπορούν να προκύψουν μετά από μια σειρά ενεργειών.

Οι περισσότερες εφαρμογές επιθέσεων σχετικά με την ασφάλεια και την ιδιωτικότητα των ML μοντέλων, διεξάγονται σε μοντέλα εκπαιδευμένα με επίβλεψη που εκτελούν εργασίες ταξινόμησης, οπότε η εργασία αυτή επικεντρώνεται κυρίως σε αυτά τα μοντέλα και αλγορίθμους.

2.1.2.3 Μοντέλα Μηχανικής Μάθησης

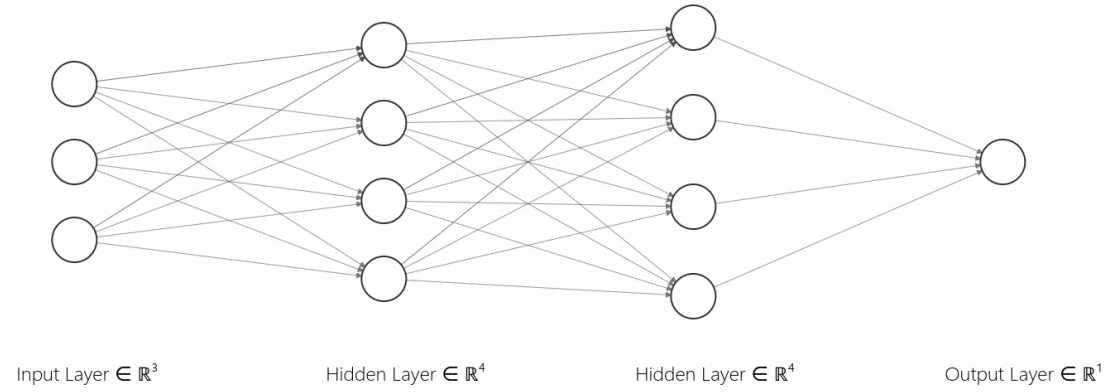
Κάποια από τα πιο γνωστά και δημοφιλή ML μοντέλα και αλγόριθμοι είναι τα εξής [8]:

- Μοντέλα Γραμμικής Παλινδρόμησης (**Linear Regression**): Στατιστική μέθοδος που χρησιμοποιείται για την εκτίμηση πραγματικών αξιών με βάση συνεχής μεταβλητές.
- Μοντέλα Λογιστικής Παλινδρόμησης (**Logistic Regression**): Στατιστική μέθοδος που χρησιμοποιείται για την εκτίμηση διακριτών τιμών (δυαδικές τιμές όπως 0/1, ναι/όχι, αλήθεια/ψέμα) με βάση ένα σύνολο ανεξάρτητων μεταβλητών, δηλαδή προβλέπει την πιθανότητα να συμβεί ένα συμβάν προσαρμόζοντας δεδομένα σε μια λογιστική συνάρτηση δίνοντας τιμές εξόδου πιθανότητες πρόβλεψης μεταξύ 0 και 1.
- Δέντρα Αποφάσεων (**Decision Tree**): Δομή δέντρου όπου μοντελοποιούνται αποφάσεις που βασίζονται σε χαρακτηριστικά των δεδομένων εισόδου, οι οποίες παίρνονται σε κάθε εσωτερικό κόμβο, οδηγώντας σε μια τελική απόφαση ή αποτέλεσμα στους κόμβους των φύλλων.
- Μηχανές Διανυσμάτων Στήριξης (**SVM: Support Vector Machines**): Μέθοδος ταξινόμησης σχεδιάζοντας κάθε δεδομένο ως σημείο σε χώρο n-διαστάσεων, με βάση των αριθμό των χαρακτηριστικών, και βρίσκοντας το βέλτιστο υπερεπίπεδο που διαχωρίζει καλύτερα διαφορετικές κλάσεις στον χώρο, μεγιστοποιώντας το περιθώριο μεταξύ των κλάσεων.
- Ταξινομητής Bayes (**Naive Bayes**): Πιθανοτικός αλγόριθμος που βασίζεται στο θεώρημα του Bayes, το οποίο υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο και χρησιμοποιείται σε εργασίες ταξινόμησης, υπολογίζοντας την πιθανότητα κάθε κλάσης με βάση την παρουσία ορισμένων χαρακτηριστικών στα δεδομένα εισόδου.
- k-Nearest Neighbors (**kNN**): Απλός αλγόριθμος που χρησιμοποιείται κυρίως για εργασίες ταξινόμησης, όπου κάνοντας προβλέψεις βρίσκει την πλειοψηφική τάξη μεταξύ των k πλησιέστερων γειτόνων στον χώρο χαρακτηριστικών για ταξινόμηση.

2.1.3 Νευρωνικά Δίκτυα και Βαθιά Μάθηση

2.1.3.1 Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (**ANN** ή απλά **NN**) [8], πρόκειται για τα πιο δημοφιλή μοντέλα μηχανικής μάθησης, τα οποία αποτελούνται από συνδεδεμένους κόμβους που ονομάζονται *νευρώνες (neurons)*, όπου ο κάθε κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων είτε από το περιβάλλον, είτε από άλλους νευρώνες, επιτελεί ένα υπολογισμό με βάση τα βάρη του και παράγει μια έξοδο η οποία καταλήγει είτε στο περιβάλλον είτε ως είσοδος σε άλλους νευρώνες. Οι νευρώνες είναι διατεταγμένοι κάθετα σε διαφορετικές ομάδες κόμβων τα οποία ονομάζονται *στρώματα ή επίπεδα (layers)* και στην πιο απλή περίπτωση νευρωνικού δικτύου υπάρχουν 3 στρωματά: το *στρώμα εισόδου (input layer)* στο οποίο τροφοδοτείται απλά η είσοδος στο νευρωνικό, το *κρυφό στρώμα (hidden layer)* στο οποίο επιτελούνται οι υπολογισμοί, το *στρώμα εξόδου (output layer)* που διοχετεύει στο περιβάλλον την τελική έξοδο. Τα νευρωνικά δίκτυα τα οποία αποτελούνται από συνδεδεμένους νευρώνες πολλαπλών επιπέδων ονομάζονται *Multi Layer Perceptron (MLP)* (βλ. σχήμα 2.1).



Σχήμα 2.1: *Σχηματική αναπαράσταση νευρωνικού δικτύου πολλαπλών επιπέδων (MLP) με 1 επίπεδο εισόδου, 1 εξόδου, και 2 κρυμμένα επίπεδα (δημιουργήθηκε με τη βοήθεια του NN-SVG εργαλείου*[1]*).*

Ο υπολογισμός που επιτελείται σε κάθε νευρώνα, είναι ο πολλαπλασιασμός του διανύσματος εισόδου $X = [x_1, x_2, ..., x_n]$ με το αντίστοιχο βάρος του νευρώνα $W = [w_1, w_1, ..., w_n]$ και το ολικό άθροισμα των γινομένων $\sum_{i=0}^{n} x_i * w_i$. Η τελική έξοδος θα προκύψει αφού περαστεί το αποτέλεσμα από μια *συνάρτηση ενεργοποίησης (activation function)* ϕ η οποία θα αποφασίσει αν περνάει ένα συγκεκριμένα όριο που έχει οριστεί, $y = \phi(\sum_{i=0}^{n} x_i * w_i)$. Αυτή η συνάρτηση μπορεί να είναι γραμμική, οπότε και η έξοδος θα είναι αντίστοιχα γραμμική εξίσωση των εξόδων προσαρμοσμένη από τα βάρη του δικτύου, όμως πλέον χρησιμοποιούνται αποκλειστικά μη γραμμικές συναρτήσεις ώστε να κάνουν το δίκτυο δυναμικό με τη δυνατότητα να εξάγει πολύπλοκα χαρακτηριστικά από τα δεδομένα και να μπορούν να αναπαριστούν μη γραμμικές και τυχαίες συσχετίσεις. Τέτοιες συναρτήσεις είναι για παράδειγμα: η *Sigmoid*, η *Tanh*, η *ReLU*, και η *Softmax*.

Απαραίτητο στοιχείο για την εκπαίδευση των NN, αλλά και άλλων ML αλγορίθμων είναι η *συνάρτηση κόστους (loss function)*, η οποίο ορίζει το πόσο καλά ένα μοντέλο κάνει προβλέψεις για ένα δεδομένο σενάριο και διαθέτει τη δικιά της *καμπύλη (curve)* και *κλίσεις*

[1] https://alexlenail.me/NN-SVG/index.html

(gradients). Ο στόχος της εκπαίδευσης είναι η ενημέρωση των βαρών του δικτύου μέσω ελαχιστοποίησης της συνάρτησης κόστους, το οποίο μπορεί να μοντελοποιηθεί ως γενικό πρόβλημα ελαχιστοποίησης. Υπάρχουν πολλοί αλγόριθμοι βελτιστοποίησης που λύνουν αυτό το πρόβλημα και χρησιμοποιούνται για την εκπαίδευση των NN, με πιο γνωστό την *Κλίση Κατάβασης (Gradient Descent)*.

2.1.3.2 Βαθιά Νευρωνικά Δίκτυα

Τα Βαθιά Νευρωνικά Δίκτυα (**DNN**) [8], πρόκειται απλά για πολυεπίπεδα νευρωνικά δίκτυα, που αποτελούνται από πολλούς νευρώνες σε σειρά και παράλληλα, τα οποία βρίσκουν μεγαλύτερη εφαρμογή και καλύτερες επιδόσεις από τα απλά NN, απαιτώντας όμως μεγαλύτερους υπολογιστικούς πόρους για την εκπαίδευση τους. Η αρχιτεκτονική των δικτύων αυτών, όπου η είσοδος τροφοδοτείται στα πολλαπλά κρυφά επίπεδα μέχρι την έξοδο ονομάζεται *εμπρόσθια τροφοδότηση (feed forward)*.

Για την εκπαίδευση των DNN, ο πιο διαδεδομένος αλγόριθμος είναι αυτός της *Οπισθοδιάδοσης (Backpropagation)*, όπου στόχος του είναι ο υπολογισμός της κλίσης της συνάρτησης κόστους ως προς τις παραμέτρους, με σκοπό την ανανέωση τους των βαρών για βέλτιστη απόδοση, στην ουσία 'ταξιδεύοντας' με την αντίθετη φορά, από την έξοδο προς της είσοδο, διαδίδοντας προς τα πίσω το σφάλμα ("backward propagation of errors").

Το πλήθος των διάφορων τύπων συναρτήσεων που μπορεί ένα NN να προσεγγίσει αναφέρεται ως *χωρητικότητα (capacity)* του δικτύου και εξαρτάται από το πλήθος των νευρώνων και επιπέδων σε ένα δίκτυο, και όσο πιο πολύπλοκο ένα δίκτυο τόσο πιο δύσκολες συναρτήσεις μπορεί να υλοποιήσει. Αν το δίκτυο έχει χαμηλή χωρητικότητά δεν θα είναι αρκετά ισχυρό να συλλάβει περίπλοκες σχέσεις των δεδομένων με τις εξόδους, οδηγώντας το σε *υποπροσαρμογή (underfitting)*. Αντίθετα, αν έχει υψηλότερη χωρητικότητά από ότι χρειάζεται αντί να προσεγγίσει τις συναρτήσεις, θα αποστηθίσει απλά τις σχέσεις, μην αποδίδοντας καλά σε νέα δεδομένα, οδηγώντας το δίκτυο σε *υπερπροσαρμογή (overfitting)*, το οποίο αποτελεί και μεγαλύτερο πρόβλημα πλέον στα σύγχρονα δίκτυα.

Για την αντιμετώπιση του προβλήματος της υπερπροσαρμογής, και αντίστοιχα τη βελτίωση της γενίκευσης του δικτύου, υπάρχουν τεχνικές *κανονικοποίησης (regularization)* οι οποίες εφαρμόζονται στα DNN για τη μείωση της διακύμανσης του μοντέλου. Τέτοιες τεχνικές είναι:

- Η *Κανονικοποίηση Βαρών (Weight Regularization)* [9] για τον περιορισμό της συνεχής αύξησης των τιμών ορισμένων βαρών.
- Η *Απόρριψη (Dropout)* [10] όπου κάποιοι νευρώνες αγνοούνται κατά τη διάρκεια της εκπαίδευσης με σκοπό την απλοποίηση του DNN.
- Η *Επαύξηση δεδομένων (Data augmentation)* [11] όπου ενισχύεται το σύνολο δεδομένων με παραλλαγμένα δεδομένα από τα αρχικά, κυρίως για σύνολα εικόνων.
- Η *Πρόωρη διακοπή (Early Stopping)* [12] για τη διακοπή της εκπαίδευσης πριν ολοκληρωθούν όλες οι εποχές όταν ικανοποιεί μια ορισμένη συνθήκη.

2.1.3.3 Μοντέλα Βαθιάς Μάθησης

Από τις πιο σημαντικές κατηγορίες βαθιών νευρωνικών δικτύων τα οποία έχουν τη μεγαλύτερη εφαρμογή είναι τα *Συνελικτικά Νευρωνικά Δίκτυα (**CNN**)*, καθώς προσφέρουν εξαιρετικές αποδόσεις σε εφαρμογές με επεξεργασία εικόνας, τα οποία χρησιμοποιούν την πράξη της *συνέλιξης (convolution)* μεταξύ των εικόνων και ενός φίλτρου σε συγκεκριμένα

κρυφά στρώματα [13]. Επίσης, σε αντίθεση με τα NN οι νευρώνες κάθε επιπέδου δεν είναι πλήρως συνδεδεμένοι με τους νευρώνες των γειτονικών επιπέδων, οι νευρώνες μοιράζονται κοινά βάρη μεταξύ τους, αντί να έχει κάθε νευρώνας ξεχωριστό βάρος και ανάμεσα στα επιμέρους στρώματα των νευρώνων υπάρχουν διατάξεις που πραγματοποιούν δειγματοληψία ώστε να μειωθούν η διαστατικότητα τις εισόδου. Κάποια από τα πιο γνωστά CNN μοντέλα είναι τα AlexNet [14], VGGNet [15] και ResNet [16].

Μια άλλη δημοφιλής κατηγορία βαθιών νευρωνικών δικτύων είναι και τα *Ανατροφοδοτούμενα Νευρωνικά Δίκτυα (**RNN**)*, τα οποία σε αντίθεση με τα δίκτυα εμπρόσθιας τροφοδότησης, η πληροφορία διοχετεύεται και με τις δύο κατευθύνσεις στα στρώματα του δικτύου [17]. Η δυνατότητα τους να κρατάνε εσωτερική κατάσταση (μνήμη), διατηρώντας πληροφορίες σχετικά με προηγούμενες εισόδους, για την επεξεργασία αυθαίρετων ακολουθιών εισόδων, τα καθιστά ιδιαίτερα χρήσιμα σε εργασίες όπως αναγνώριση χειρόγραφης γραφής ή αναγνώριση ομιλίας.

Υπάρχουν όμως και άλλα μοντέλα βαθιάς μάθησης τα οποία δεν ανήκουν στην κλασσική αρχιτεκτονική των NN. Κάποια από τα πιο δημοφιλή είναι:

- Autoencoders (**AE**) [18]: Τύπος NN το οποίο χρησιμοποιείται για την εκμάθηση αποτελεσματικών κωδικοποιήσεων από δεδομένα χωρίς ετικέτα (μη επιβλεπόμενη μάθηση), με σκοπό τη μείωση των διαστάσεων, διατηρώντας μόνα τα σημαντικά χαρακτηριστικά των δεδομένων ή για την παραγωγή δεδομένων.
- Παραγωγικά Ανταγωνιστικά Δίκτυα (**GAN**) [19]: Δύο NN (μια γεννήτρια και μια συσκευή διάκρισης) τα οποία διαγωνίζεται σε ένα παίγνιο με σκοπό τη βελτίωση της αναπαράστασης δεδομένων, για εφαρμογές μη επιβλεπόμενης ή ημιεπιβλεπόμενης μάθησης. Προτείνονται για τη δημιουργία νέων δεδομένων με ίδια στατιστικά στοιχεία με τα δεδομένα εκπαίδευσης.
- Μετασχηματιστές (**Transformers**) [20]: Αρχιτεκτονική NN όπου το μοντέλο μαθαίνει το πλαίσιο (context) ανιχνεύοντας σχέσει μεταξύ των διαδοχικών δεδομένων, όπως λέξεις σε μια πρόταση, χωρίς την ανάγκη ανατροφοδοτούμενων μονάδων όπως τα RNN. Πρόκειται για την κύρια αρχιτεκτονική που χρησιμοποιείται από τα σύγχρονα μεγάλα γλωσσικά μοντέλα (LLMs).
- Μοντέλα Διάχυσης (**Diffusion Models**) [21]: Τύπος παραγωγικών μοντέλων αποτελούμενο από 3 μονάδες, την εμπρόσθια διαδικασία, την αντίστροφη διαδικασία και τη διαδικασία της δειγματοληψίας, όπου μαθαίνοντας την κατανομή πιθανότητας των δεδομένων εκπαίδευσης παράγουν νέα δεδομένα, αποδίδοντας πολλές φορές καλύτερα από τα GANs σε εργασίες π.χ. σύνθεσης εικόνας.

2.2 Ασφάλεια συστημάτων Τεχνητής Νοημοσύνης

2.2.1 Ασφάλεια Συστημάτων και Κυβερνοασφάλεια

Με την εκθετική αύξηση χρήσης του Διαδικτύου και των συστημάτων τα οποία εκτίθενται σε αυτό, έχει αυξηθεί και ο κίνδυνος των *κυβερνοεπιθέσεων (cyber attacks)*, δηλαδή επιθέσεις σε συστήματα τα οποία μπορούν να συμβούν απομακρυσμένα από κακόβουλους χρήστες, χωρίς την ανάγκη φυσικής παρουσίας [22]. Έτσι η ασφάλεια συστημάτων έχει αποκτήσει όλο και μεγαλύτερη σημασία, με εταιρείες και βιομηχανίες να επενδύουν μεγάλα ποσά και να δίνουν μεγάλη προσοχή στην ασφάλεια των συστημάτων και των λογισμικών που χρησιμοποιούν ή

παράγουν.

Η *κυβερνοασφάλεια cybersecurity* αφορά την προστασία οποιασδήποτε ψηφιακής τεχνολογίας και τεχνολογίας πληροφοριών (π.χ. δίκτυα, υπολογιστές, διακομιστές, δεδομένα), παίρνοντας τα κατάλληλα μέτρα άμυνας για τη διατήρηση της *εμπιστευτικότητας (confidentiality)*, της *ακεραιότητας (integrity)*, και της *διαθεσιμότητας (availability)* αυτών.

2.2.2 Επιθέσεις σε Συστήματα Τεχνητής Νοημοσύνης

Η διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων, μαζί με την πρόοδο στην υπολογιστική ισχύ, έχει επιτρέψει την ανάπτυξη ισχυρών AI εφαρμογών και εφόσον η τεχνητή νοημοσύνη έχει πλέον ευρεία εφαρμογή και ειδικότερα τα μοντέλα μηχανικής μάθησης χρησιμοποιούνται σε διάφορα συστήματα για την παροχή υπηρεσιών ή την εκτέλεση εργασιών, είναι επόμενο να αποτελέσουν στόχο επιθέσεων. Ενώ μπορούν να εκτελεστούν επιθέσεις σε τέτοια συστήματα με τον ίδιο τρόπο όπως σε άλλα συστήματα, όπως π.χ. με εύρεση και με εκμετάλλευση ευπαθειών στον πηγαίο κώδικα, αναβάθμιση δικαιωμάτων χρήστη, χρήση ιών και κακόβουλου λογισμικού, αυτές οι επιθέσεις στην ουσία έχουν ως στόχο πάλι τα ίδια τα υπολογιστικά συστήματα και όχι τα μοντέλα τεχνητής νοημοσύνης.

Σε αυτήν την εργασία *γίνεται εστίαση στις επιθέσεις που έχουν εφαρμογή και στόχο τα ίδια τα μοντέλα και όχι στο γενικότερο πεδίο των κυβερνοεπιθέσεων*, το οποίο όμως είναι αρκετά σημαντικό όταν γίνεται αναφορά στη γενικότερη ασφάλεια συστημάτων.

Μια από τις πρώτες καταγεγραμμένες ευπάθειες σε μοντέλο μηχανικής μάθησης εμφανίζεται το 2020, υπό μορφή CVE στην εθνική βάση ευπαθειών των ΗΠΑ[2], για ένα εμπορικό σύστημα που χρησιμοποιούσε ένα ML σύστημα ταξινόμησης για spam email, από το οποίο επιτιθέμενοι μπορούσαν να εξάγουν πληροφορίες, με σκοπό την δημιουργία κακόβουλων email για την παράκαμψη του έλεγχου. Επίσης, την ίδια χρονιά, το συντονιστικό κέντρο αντιμετώπισης καταστάσεων έκτακτης ανάγκης υπολογιστών (CERT/CC) του πανεπιστήμιου Carnegie Mellon[3], έθεσε το πρόβλημα της ευπάθειας των ML συστημάτων από επιθέσεις με χρήση *διαταραχών (perturbations)* ή αλλιώς *ανταγωνιστικών παραδειγμάτων (adversarial examples)*, οι οποίες εκμεταλλεύονται εγγενείς ευπάθειες των ML μοντέλων για πρόκληση εσφαλμένης ταξινόμησης.

Η αμερικάνικη συμβουλευτική εταιρεία τεχνολογικής έρευνας Gartner, σε λίστα με τις 10 κορυφαίες στρατηγικές τεχνολογικές τάσεις για το 2020 κατέταξε στην 10η θέση την ασφάλεια AI συστημάτων[4], δείχνοντας την προσοχή και σημασία που έχει αναπτυχθεί αυτόν τον κλάδο. Όμως την ίδια χρονιά έρευνα από τη Microsoft σε 28 επιχειρήσεις που κάνουν χρήση ML συστημάτων, οι 25 από αυτές δεν έχουν τα κατάλληλα εργαλεία, γνώσεις ή χρόνο για την εκπαίδευση ασφαλών μοντέλων, και η ασφάλεια περιορίζεται σε παραδοσιακές πρακτικές, παρά στην ενίσχυση των ίδιων των μοντέλων [23].

2.2.3 Πεδίο Επιθέσεων σε Συστήματα Τεχνητής Νοημοσύνης

Γενικά, η ασφάλεια οποιουδήποτε συστήματος μετριέται πάντα σε σχέση με τους στόχους και τις δυνατότητες ενός που πρόκειται να επιτεθεί σε ένα σύστημα, τον οποίο ονομάζουμε *αντίπαλος (adversary)* ή *επιτιθέμενο (attacker)* [4]. Με βάση τους αντιπάλους σχεδιάζεται και

[2] https://nvd.nist.gov/vuln/detail/CVE-2019-20634
[3] https://kb.cert.org/vuls/id/425163
[4] https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020

το μοντέλο απειλής *(threat model)* του συστήματος. Το ίδιο ισχύει και για την ασφάλεια των συστημάτων τεχνητής νοημοσύνης, οπότε πρέπει πρώτα να γίνει αναγνώριση του μοντέλου απειλών σε αυτά, αναγνωρίζοντας το πως οι αντίπαλοι μπορούν να επιχειρήσουν επιθέσεις σε ένα τέτοιο σύστημα και τι σκοπούς έχουν.

Το μοντέλο απειλής σε ένα ML σύστημα, τα στοιχεία που είναι ευαίσθητα και δυνητικά υπόκεινται σε επίθεση είναι:

1. Το σύνολο δεδομένων εκπαίδευσης.
2. Το ίδιο το μοντέλο.
3. Οι παράμετροι, υπερπαράμετροι και η αρχιτεκτονική του μοντέλου.

Αντίστοιχα, τα άτομα τα όποια επηρεάζονται σε αυτό το μοντέλο απειλής και πρέπει να προστατευτούν είναι [2]:

- Οι *κάτοχοι δεδομένων*, των οποίων τα δεδομένα ενδέχεται να είναι ευαίσθητα.
- Οι *κάτοχοι μοντέλων*, οι οποίοι μπορεί να είναι κάτοχοι ή όχι των δεδομένων και μπορεί να θέλουν ή να μη θέλουν να μοιραστούν πληροφορίες σχετικά με τα μοντέλα τους.
- Οι *χρήστες των μοντέλων* που χρησιμοποιούν τις υπηρεσίες που εκθέτει ο ιδιοκτήτης του μοντέλου, συνήθως μέσω κάποιου είδους προγραμματισμού ή διεπαφής χρήστη (API).
- Οι *αντίπαλοι*, που μπορεί επίσης να έχουν πρόσβαση στις διεπαφές του μοντέλου όπως έχει ένας κανονικός καταναλωτής. Εάν το επιτρέπει ο κάτοχος του μοντέλου, μπορεί να έχουν πρόσβαση και στο ίδιο το μοντέλο.

Τέλος, το *πεδίο επιθέσεων (attack surface)* σε ένα σύστημα αποτελούμενο από δεδομένα και ML μοντέλα, κατά τη φάση εξαγωγής συμπερασμάτων, σε γενικές γραμμές μπορεί να συνοψιστεί σε αυτές τις διαδικασίες στις οποίες μπορεί να εξαχθεί επίθεση με στόχο κάποια από τα παραπάνω στοιχεία [4]:

- Χαρακτηριστικά εισόδου συλλέγονται από αισθητήρες ή αποθετήρια δεδομένων.
- Γίνεται επεξεργασία των δεδομένων σε ψηφιακό επίπεδο.
- Το μοντέλο χρησιμοποιεί τα δεδομένα για την παραγωγή εξόδου.
- Η έξοδος κοινοποιείται σε κάποιο εξωτερικό σύστημα ή χρήστη.

2.2.4 Ασφάλεια και Ευρωστία Συστημάτων Τεχνητής Νοημοσύνης

Για την προστασία των συστημάτων τεχνητής νοημοσύνης, δηλαδή όλων των παραπάνω παραγόντων, χρειάζονται διαφορετικές προσεγγίσεις και λύσεις, καθώς δεν υπάρχει ένας μοναδικός τρόπος να προστατευθούν καθολικά, όπως και γενικότερα στον κλάδο της ασφάλειας συστημάτων.

Οι στόχοι της προστασίας και της ασφάλειας τέτοιων συστημάτων, μπορούν να κατηγοριοποιηθούν στους γενικότερους στόχους της εμπιστευτικότητας, διαθεσιμότητας, και ακεραιότητας [24], όμως πιο συγκεκριμένα οι προκλήσεις και στόχοι για ασφαλή και εύρωστα συστήματα τεχνητής νοημοσύνης μπορούν να συνοψιστούν στους παρακάτω:

- *Ηθική (**Ethical**)* και έμπιστη τεχνητή νοημοσύνη: Περιλαμβάνει σχεδιασμό και ανάπτυξη AI συστημάτων τα οποία δίνουν προτεραιότητα στη δικαιοσύνη, τη διαφάνεια, την αμεροληψία, τη μη κακοήθεια, δίνοντας εξηγήσεις στη λήψη αποφάσεων, ελαχιστοποιώντας τις πιθανές προκαταλήψεις και τις ανεπιθύμητες συνέπειες [25].
- *Ασφαλή (**Secure**)* συστήματα τεχνητής νοημοσύνης: Περιλαμβάνει τον σχεδιασμό ασφαλών συστημάτων με την κλασσική έννοια, ακολουθώντας τις καλύτερες πρακτικές,

κατά την ανάπτυξη λογισμικού (όπως έλεγχο των εισόδων ή απολύμανση των εξόδων από ευαίσθητα δεδομένα), κατά τη λειτουργία (όπως συχνή ενημέρωση εφαρμογών ή έλεγχο πρόσβασης και δικαιωμάτων στα συστήματα) και στη διαχείριση δεδομένων (όπως αποθήκευση με κρυπτογράφηση, ή αυστηρούς ελέγχους πρόσβασης και περιορισμό έκθεσης των δεδομένων) [24].

- *Εύρωστα (**Robust**)* μοντέλα μηχανικής μάθησης: Περιλαμβάνει τον σχεδιασμό και την ανάπτυξη ML μοντέλων τα οποία είναι ανθεκτικά σε απρόβλεπτα γεγονότα και ανταγωνιστικές επιθέσεις, όπου με χρήση των κατάλληλων εισόδων μπορεί να προκληθεί βλάβη ή να χειραγωγηθούν τέτοια συστήματα. Πρόκειται για ασφάλεια των ίδιων των μοντέλων, είτε από αντιπάλους που μεταλλάσσουν τις εισροές του μοντέλου, είτε για περιπτώσεις *μετατόπισης τις κατανομής των δεδομένων (distribution shift)*, είτε ακόμα για ασταθής περιπτώσεις στις οποίες το μοντέλο δεν έχει εκπαιδευτεί [26].
- *Ιδιωτικά (**Private**)* μοντέλα μηχανικής μάθησης: Περιλαμβάνει τον σχεδιασμό και την ανάπτυξη ML μοντέλων τα οποία δίνουν προτεραιότητα στην ιδιωτικότητα και την εμπιστευτικότητα των δεδομένων, για να διασφαλιστεί ότι οι ευαίσθητες πληροφορίες παραμένουν ασφαλείς και ανώνυμες κατά τις διαδικασίες εκπαίδευσης και συμπερασμάτων, είτε από επιτιθέμενους που έχουν στόχο στην εξαγωγή δεδομένων από το σύστημα, είτε ακούσια διαρρέοντας προσωπικά δεδομένα κατά την παραγωγή αποτελεσμάτων [4].

Πολλοί από αυτούς τους στόχους είναι ακόμα σε ερευνητικό επίπεδο και παραμένουν ανοιχτά προβλήματα [4], ενώ άλλοι έχουν ήδη εφαρμογή σε κρίσιμους κλάδους της βιομηχανίας. Σε αυτήν την εργασία *γίνεται εστίαση στους τρόπους ενίσχυσης της ευρωστίας και ιδιωτικότητας των μοντέλων μηχανικής μάθησης και όχι στους υπόλοιπους τομείς*, οι οποίοι όμως είναι αρκετά σημαντικοί όταν εξετάζεται η γενικότερη ασφάλεια των συστημάτων τεχνητής νοημοσύνης.

Κεφάλαιο 3

Ανταγωνιστική Μηχανική Μάθηση - Επιθέσεις και Άμυνες

Στο κεφάλαιο αυτό ορίζεται η έννοια των ανταγωνιστικών παραδειγμάτων (adversarial examples), καθώς και η έννοια της ανταγωνιστικής μηχανικής μάθησης (adversarial machine learning. Επίσης, αναλύονται οι ανταγωνιστικές επιθέσεις ενάντια σε μοντέλα μηχανικής μάθησης καθώς και τεχνικές άμυνας ενάντια σε τέτοια δείγματα και επιθέσεις. Επίσης, γίνεται μια αναφορά σε κάποια παραδείγματα πραγματικών ανταγωνιστικών επιθέσεων.

3.1 Το πεδίο της Ανταγωνιστικής Μηχανικής Μάθησης

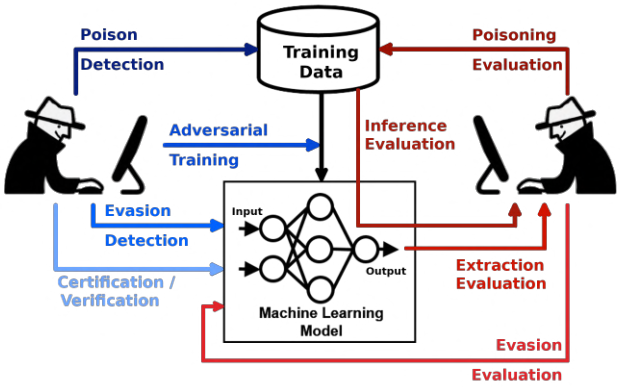
Τα τελευταία χρόνια τα νευρωνικά δίκτυα (Neural Networks) και η μηχανική μάθηση (Machine Learning ή αλλιώς ML) έχουν κάνει πολύ μεγάλη πρόοδο, αυξάνοντας τις επιδόσεις τους και βρίσκοντας εφαρμογή σε όλο και περισσότερους τομείς. Παρόλα αυτά, τα συστήματα αυτά παραμένουν ευάλωτα σε διάφορα είδη επιθέσεων, τα οποία εκμεταλλεύονται τα τρωτά σημεία στη διαδικασία λήψης αποφάσεων των μοντέλων. Ένα μεγάλο μέρος αυτών των επιθέσεων είναι και οι ανταγωνιστικές επιθέσεις (**Adversarial Attacks**), όπου ένας επιτιθέμενος προσαρμόζει τα δεδομένα εισόδου του συστήματος ώστε να το προκαλέσει να κάνει εσφαλμένη ή μη βέλτιστη επιλογή. Μία τέτοια προσεκτικά παραλλαγμένη είσοδος, μπορεί να δοθεί είτε στα δεδομένα κατά την εκπαίδευση του νευρωνικού (**training**), είτε και κατά τη δοκιμή του (**testing**), ώστε να προκαλέσουν ή συγκεκριμένες επιλογές με την κατάλληλη είσοδο ή γενικότερα εσφαλμένες επιλογές [27].

Όλες αυτές οι επιθέσεις μπορούν να γίνουν ιδιαίτερα επικίνδυνες, ειδικά όταν έχουν εφαρμογή σε τομείς όπου οι αποφάσεις ενός AI συστήματος είναι κρίσιμες, όπως στην αυτόνομη οδήγηση, σε υποδομές ενέργειας, σε νοσοκομεία και ιατρικές διαγνώσεις ή σε άλλες δημόσιες υποδομές και υπηρεσίες.

Αντίστοιχα με την ανάπτυξη αυτών των επιθέσεων έχουν αναπτυχθεί και άμυνες ή αντεπιθέσεις (**Adversarial Defenses**) που μπορούν να εφαρμοστούν ώστε να προστατεύσουν τα ML συστήματα, αυξάνοντας την ευρωστία (**Robustness**) τους. Όμως πρόκειται για ένα δύσκολο πρόβλημα και είναι ένας τομέας ενεργούς έρευνας με λίγες αποδεδειγμένες γενικεύσιμες λύσεις. Η ερευνητική κοινότητα πάνω στο συγκεκριμένο πεδίο, ενώ εργάζεται σε συγκεκριμένες άμυνες, βρίσκεται σε μια κούρσα όπου προτείνονται και καταρρίπτονται συνεχώς άμυνες, εφευρίσκοντας νέες επιθέσεις. Η εστίαση πλέον είναι περισσότερο στον έλεγχο της ευρωστίας ενός νευρωνικού δικτύου, για τον εντοπισμό ελαττωμάτων του καθώς και την αξιολόγηση των νέων αμυνών [28].

Από την πρώτη εμφάνιση των ανταγωνιστικών επιθέσεων έως και σήμερα, υπάρχει ένας διαρκής κύκλος ανατροφοδότησης, για την ανακάλυψη νέων επιθέσεων καθώς και μέτρων

αντιμετώπισης. Αυτός ο κύκλος για την *ανακάλυψη νέων επιθέσεων* περιλαμβάνει: (α) τη δημιουργία νέας επίθεσης πάνω σε ένα νευρωνικό, (β) τη δοκιμή υπαρχόντων αμυνών πάνω σε αυτή, (γ) την αξιολόγηση της με βάση κάποιες μετρικές ευρωστίας. Αντίθετα, για την *ανακάλυψη νέων αμυνών* περιλαμβάνει την αντίστροφη διαδικασία: (α) τη δημιουργία νέας άμυνας, (β) δοκιμή υπαρχόντων επιθέσεων σε αυτήν, (γ) την αξιολόγηση της με βάση κάποιες μετρικές ευρωστίας (βλέπε σχήμα 3.1). Αυτός ο κύκλος συμβαίνει αρχικά από κάποιον ή μια ομάδα ερευνητών κατά τη διάρκεια των δοκιμών (testing) μιας μεθόδου για την αξιολόγηση της, καθώς και μετά τη δημοσίευσή της από άλλους ερευνητές για την ενίσχυση της αποτελεσματικότητας της, ή για την κατάρριψη της. Περισσότερα για τις μεθόδους και τον τρόπο αξιολόγησης αυτών των τεχνικών, στο Κεφάλαιο 4 για **εύρωστα μοντέλα μηχανικής μάθησης**.



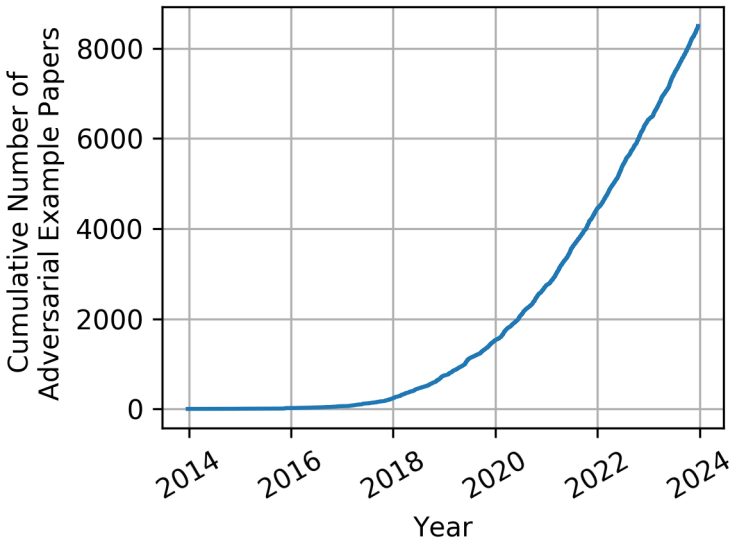
Σχήμα 3.1: *Ο κύκλος της ανταγωνιστικής μηχανικής μάθησης: Επίθεση, Άμυνα, Αξιολόγηση*[1]

Σύμφωνα με έρευνα από ερευνητή της DeepMind [29], και του πανεπιστημίου Carnegie Mellon [28], το πεδίο της ανταγωνιστικής μηχανικής μάθησης (**Adversarial Machine Learning**) είναι αρκετά καινούριο και με λίγη έρευνα, η οποία έχει αυξηθεί κατά πολύ τα τελευταία 5 χρόνια, ξεκινώντας όμως πριν από 10 χρόνια σχεδόν, όπως φαίνεται και στο διάγραμμα 3.2. Οπότε υπάρχουν λίγοι καθολικά αποδεκτοί τρόποι και μέθοδοι που μπορούν να μειώσουν την αντίκτυπο των ανταγωνιστικών επιθέσεων και άρα να αυξήσουν την ευρωστία ενός AI συστήματος. Νέες τεχνικές και άμυνες προτείνονται κάθε χρόνο, αλλά πολλές από αυτές καταρρίπτονται πολύ εύκολα και γρήγορα, ύστερα από ανεξάρτητη αξιολόγηση από άλλους ερευνητές [26].

Ακόμη όμως και για τις πιο καθολικά αποδεκτές μεθόδους που υπάρχουν για να προστατεύουν από αυτές τις επιθέσεις, το ποσοστό επιτυχίας επίθεσης (Attack Success Rate - ASR), παραμένει σχεδόν στο 50%, όταν σε άλλους τομείς της ασφάλειας, όπως η κρυπτογραφία, το ASR είναι της τάξεως του 2^{-128} [30].

Επίσης, η συνεχής αυτή έρευνα έχει προσπαθήσει με τα χρόνια να κατηγοριοποιήσει και να τυποποιήσει αυτό το πεδίο, είτε αναφερόμαστε στις διάφορες ανταγωνιστικών επιθέσεις, είτε στις άμυνες, είτε ακόμα και στις μεθόδους και δείκτες αξιολόγησης αυτών. Στις επόμενες ενότητες αναλύονται οι διάφορες επιθέσεις (βλ. 3.4) και άμυνες (βλ. 3.6) που έχουν αναπτυ-

[1] https://github.com/Trusted-AI/adversarial-robustness-toolbox



Σχήμα 3.2: *Αριθμός των Adversarial Example Papers ανά έτος [29]*

χθεί τα τελευταία χρόνια, καθώς και κάποια παραδείγματα από τον πραγματικό κόσμο (βλ. 3.5).

3.2 Ανταγωνιστικά Παραδείγματα

3.2.1 Προέλευση όρου

Στην ιστορία της τεχνητής νοημοσύνης, όπως αναφέρεται στο [27] η πρώτη εμφάνιση του όρου των ανταγωνιστικών ή αντιπαραθετικών ή αντίπαλων δειγμάτων (**adversarial examples**), έγινε περίπου το 2004 στο πλαίσιο των φίλτρων ανεπιθύμητης αλληλογραφίας (spam filtering), όπου παρατηρήθηκε ότι οι μικρές τροποποιήσεις στα spam email τα επέτρεψαν να περάσουν μέσα από τα φίλτρα, τα οποία επρόκειτο για γραμμικούς ταξινομητές (linear classifiers), χωρίς να επηρεάσουν σημαντικά το περιεχόμενο του μηνύματος. Από κει και έπειτα τα adversarial examples έχουν επεκταθεί σε άλλους κλάδους της τεχνητής νοημοσύνης, όπως π.χ. η ταξινόμηση εικόνων και κακόβουλου λογισμικού (malware) μέσω μηχανικής μάθησης.

Η πρώτη εμφάνιση του όρου στο πλαίσιο της των τεχνικών βαθιάς μάθησης (deep learning), όπου έπειτα το ενδιαφέρον για αυτόν τον τομέα συνέχισε να αυξάνεται, με όλο και περισσότερες δημοσιεύσεις κάθε χρόνο, έγινε από τον Szegedy [31], όπου εφαρμόζοντας, μικρές και με τα βίας αντιληπτές από το ανθρώπινο μάτι, διαταραχές ή παραλλαγές (**perturbations**) σε εικόνες εισόδου του νευρωνικού, παρατηρήθηκε μεγιστοποίηση του σφάλματος πρόβλεψης (prediction error). Αυτές τις μικρές παραλλαγές οι οποίες βρίσκονται προσαρμόζοντας την είσοδο ώστε να μεγιστοποιηθεί το σφάλματος πρόβλεψης, ονομάζονται adversarial examples. Οι εφαρμογές των adversarial examples, γίνονται σε διάφορα είδη εισόδων πέρα από εικόνες, όπως κείμενο [32], ήχος [33], εκτελέσιμα αρχεία [34] κ.α., αλλά το μεγαλύτερο μέρος της βιβλιογραφίες εστιάζει σε διαταραχές εικόνων, καθώς οι ταξινομητές εικόνων έχουν και την μεγαλύτερη απήχηση και εφαρμογή. Το πιο σημαντικό όμως συμπέρασμα, ήταν ότι η φύση αυτών των παραλλαγών, δεν είναι τυχαίο αποτέλεσμα κάποιας μηχανικής μάθησης, συγκεκριμένων παραμέτρων, αλλά μπορεί να εμφανιστεί σε όλα τα νευρωνικά δίκτυα ανεξαρτήτως

στρώματος, συνάρτησης ενεργοποίησης, ή δεδομένων εκπαίδευσης.

3.2.2 Μαθηματικός ορισμός

3.2.2.1 Ανταγωνιστικά παραδείγματα σε γραμμικά μοντέλα

Όπως έχει αναφερθεί τα adversarial examples, πέρα από τα νευρωνικά δίκτυα που είναι μη γραμμικά μοντέλα, έχουν επίπτωση και στα γραμμικά μοντέλα, όποτε γίνεται μια πρώτη αναφορά και εξήγηση σε αυτά. Σε πολλά προβλήματα η ακρίβεια μιας μεμονωμένης εισόδου είναι περιορισμένη. Για παράδειγμα, στις ψηφιακές εικόνες χρησιμοποιούμε συνήθως μόνο 8 bit ανά pixel, οπότε απορρίπτονται όλες οι πληροφορίες κάτω από το 1/255 του δυναμικού εύρους. Λόγου αυτού του περιορισμού της ακρίβειας των χαρακτηριστικών, είναι λογικό ένας ταξινομητής να μην απαντά διαφορετικά σε μία είσοδο x από ότι σε μία ανταγωνιστική είσοδο $\tilde{x} = x+\delta$, εάν κάθε στοιχείο της διαταραχής δ είναι μικρότερο από ότι η ακρίβεια των χαρακτηριστικών. Τυπικά δηλαδή, για προβλήματα με καλά διαχωρισμένες κλάσεις, είναι αναμενόμενο από έναν ταξινομητή να αντιστοιχήσει στην ίδια κλάση το x και το $\tilde{x}$, εφόσον $\|\delta\|_\infty < \epsilon$, όπου ϵ αρκετά μικρό ώστε να απορριφθεί από τον αισθητήρα ή τη συσκευή αποθήκευσης δεδομένων που σχετίζεται με το πρόβλημα αυτό.

Εάν θεωρήσουμε το εσωτερικό γινόμενο μεταξύ ενός διανύσματος βάρους w και μιας ανταγωνιστικής εισόδου $\tilde{x}$:

$$w^T\tilde{x} = w^Tx + w^T\delta$$

Το adversarial perturbation προκαλεί αύξηση της ενεργοποίησης κατά $w > \delta$, η οποία μπορεί να μεγιστοποιηθεί αντικαθιστώντας τη διαταραχή με τη συνάρτηση πρόσημου του διανύσματος βάρους $\delta = sign(w)$. Για προβλήματα υψηλών διαστάσεων, απειροελάχιστες αλλαγές στην είσοδο μπορούν να οδηγήσουν σε μία μεγάλη αλλαγή εξόδου, δημιουργώντας ένα είδος «τυχαίας στεγανογραφίας», όπου ένα γραμμικό μοντέλο αναγκάζεται να παρακολουθεί αποκλειστικά το σήμα που ευθυγραμμίζεται περισσότερο με τα βάρη του, ακόμα και αν πολλαπλά σήματα έχουν μεγαλύτερο πλάτος.

Αυτή η εξήγηση σύμφωνα με το [1] δείχνει ότι ένα απλό γραμμικό μοντέλο είναι ευάλωτο σε adversarial examples, εφόσον η είσοδος έχει επαρκείς διαστάσεις, απλοποιώντας την εξήγηση του γιατί οι softmax ταξινομητές είναι ευπαθείς σε adversarial examples.

3.2.2.2 Ανταγωνιστικά παραδείγματα σε μη-γραμμικά μοντέλα

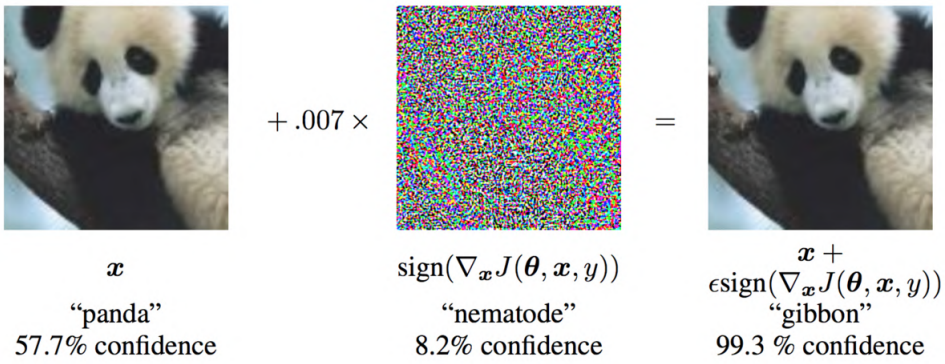
Τα μη γραμμικά μοντέλα όπως τα νευρωνικά δίκτυα είναι εξίσου ευάλωτα σε adversarial examples, καθώς η μη γραμμικότητα τους τα βοηθάει να μπορούν να μοντελοποιούν οποιαδήποτε συνάρτηση με τον κατάλληλο αριθμό στρωμάτων και νευρώνων, καταλήγουν να συμπεριφέρονται με γραμμικό τρόπο, λόγο κυρίως των συναρτήσεων ενεργοποίησης (π.χ. ReLU) ώστε να μπορούν να βελτιστοποιηθούν εύκολα.

Αν θέσουμε θ τις παραμέτρους ενός μοντέλου, x την είσοδο του, y τους στόχους σχετιζόμενους με τη x, και $J(\theta, x, y)$ το κόστος εκπαίδευσης του νευρωνικού, μπορούμε να 'γραμμικοποιήσουμε' τη συνάρτηση κόστους γύρω από την τρέχουσα τιμή του θ, λαμβάνοντας ένα βέλτιστο μέγιστο κανόνα περιορισμένης διαταραχής με τιμή:

$$\delta = \epsilon sign(\nabla_x J(\theta, x, y))$$

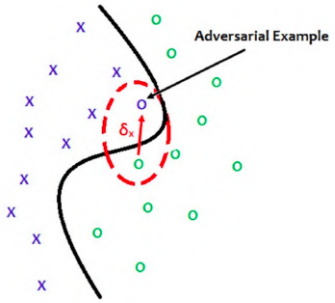
Η απαιτούμενη κλίση (gradient) μπορεί να υπολογιστεί αποτελεσματικά χρησιμοποιώντας backpropagation [1]. Η συγκεκριμένη μέθοδος για τη δημιουργία adversarial examples ονομάζεται **Fast Gradient Sign Method** και παρουσιάζεται αναλυτικά στην ενότητα 3.4.2.1 μαζί με άλλα είδη ανταγωνιστικών επιθέσεων.

Αυτή η μέθοδος αποδείχθηκε ότι προκαλεί πλήθος διαφορετικών μοντέλων να κάνουν, αρκετά αξιόπιστα, λανθασμένη ταξινόμηση (misclassify) των εισόδων τους. Ένα κλασσικό παράδειγμα είναι αυτό που φαίνεται στο σχήμα 3.3 με τη φωτογραφία ενός πάντα, το οποίο αναγνωρίζεται και ταξινομείται σωστά από το GoogLeNet DNN με βεβαιότητά 57.7%. Όμως με την εισαγωγή θορύβου υπό τη μορφή μικρών διαταραχών εφαρμόζοντάς την παραπάνω μέθοδο, ώστε να δημιουργηθεί ένα adversarial example, αυτό καταλήγει να αναγνωρίζεται ως γίββωνας με βεβαιότητα 99.3% [1].



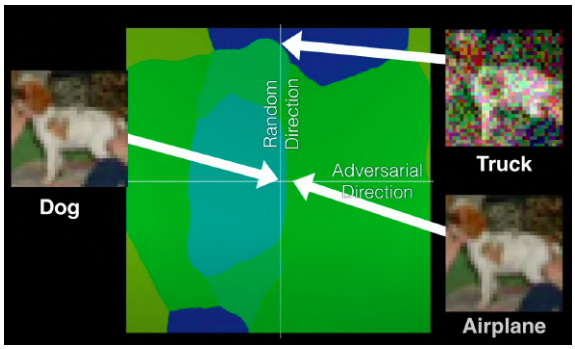
Σχήμα 3.3: *Παράδειγμα εισαγωγής θορύβου για την παραγωγή ενός adversarial example το οποίο γίνεται misclassify με μεγαλύτερη σιγουριά από ότι η αρχική εικόνα [1]*

Η διαταραχή αυτή που ονομάσαμε δ_x, μπορεί να ερμηνευτεί ως το ένα διάνυσμα $\overrightarrow{\delta}_x$ όπου το μέγεθος του $\left\|\overrightarrow{\delta}_x\right\|$ αντιπροσωπεύει την ποσότητα της διαταραχής που χρειάζεται ώστε να μεταφραστεί το σημείο που αντιπροσωπεύεται από την είσοδο x σε χώρο πέρα από το όριο απόφασης (decision bounday). Στο σχήμα 3.4, φαίνεται και οπτικά αυτή η εισαγωγή της διαταραχής δ_x σε μία κανονική είσοδο εικόνας x πάνω στον δισδιάστατο (2D) χώρο [5]. Ένα άλλο τέτοιο οπτικό παράδειγμα φαίνεται και στο σχήμα 3.5, όπου φαίνεται και στον τρισδιάστατο (3D) χώρο, το πώς εισάγοντας την κατάλληλη διαταραχή, μετατοπίζει την είσοδο πέρα από το όριο απόφασης.

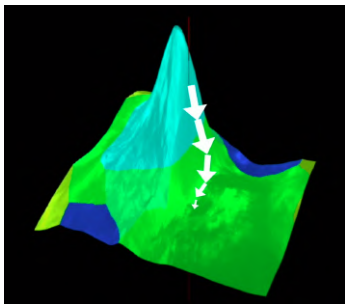


Σχήμα 3.4: *Ο στόχος ενός μιας ανταγωνιστικής επίθεσης είναι να παράξει μια διαταραχή δ_x η οποία όταν εισαχθεί στην κανονική είσοδο x, να δημιουργήσει το adversarial example $\tilde{x} = x + \delta_x$ το οποίο θα καταφέρει να περάσει το όριο απόφασης του νευρωνικού [5]*

Οι περισσότερες τεχνικές δημιουργίας adversarial examples, προτείνουν την ελαχιστοποίηση της απόστασης μεταξύ του adversarial example και του στιγμιότυπου που πρόκειται να χειραγωγηθεί, μετατοπίζοντας παράλληλα την πρόβλεψη στο επιθυμητό (adversarial) αποτέλεσμα. Κάποιες μέθοδοι, εκμεταλλεύονται όπως η προηγούμενη τις κλήσεις του μοντέλου, ενώ άλλες π.χ. τη συνάρτηση πρόβλεψης (prediction function) του μοντέλου, όμως υπάρχουν και άλλες επιθέσεις που λειτουργούν ανεξάρτητα αν υπάρχει επίγνωση ή πρόσβαση στα χαρακτηριστικά του μοντέλου. Περισσότερη ανάλυση για τα είδη των επιθέσεων γίνεται στην ενότητα 3.4.



(α′) *Διάγραμμα 2D του decision boundary*



(β′) *Διάγραμμα 3D του decision boundary*

Σχήμα 3.5: *Διάγραμμα του decision boundary ενός NN, όπου τα δείγματα που ανήκουν στις μπλε περιοχές αναγνωρίζονται ως σκύλος. Παράδειγμα ενός adversarial example όπου προστίθεται τέτοιος θόρυβος που οδηγεί την απόφαση στην χειρότερη περιοχή, με τον λιγότερο δυνατό θόρυβο [30]*

Συνοψίζοντας, τα χαρακτηριστικά τα οποία έχουν τα adversarial examples και σύμφωνα με το [35] είναι αυτά και τα οποία χαρακτηρίζουν αν είναι πετυχημένο example τα εξής:

1. Ομοιότητα σε σχέση με αρχική είσοδο, ώστε με το ανθρώπινο μάτι να μην διακρίνεται διαφορά σε σχέση με την αρχική είσοδο.
2. Οι διαταραχές να είναι ικανές να κάνουν το νευρωνικό να ταξινομήσει την είσοδο σε λανθασμένη κλάση και ιδανικά με υψηλή βεβαιότητα.

3.3 Ανταγωνιστική Εκπαίδευση

3.3.1 Προέλευση όρου

Όλη η έρευνα και ενασχόληση γύρω από την ευπάθεια της μηχανικής μάθησης σε τέτοιες επιθέσεις που σχετίζονται δηλαδή στο πως θα ξεγελάσουν ένα νευρωνικό να ξεφύγει από την κανονική του λειτουργία, μαζί με τον σχεδιασμό κατάλληλων αντεπιθέσεων, είναι το αντικείμενο της ανταγωνιστικής ή αντιπαραθετικής ή αντίπαλης μηχανικής μάθησης (**Adversarial Machine Learning**). Όπως είναι λογικό μαζί με την εμφάνιση των adversarial examples και την ανάδειξη της ευπάθειας των νευρωνικών δικτύων σε τέτοια παραδείγματα, αναπτύχθηκε ταυτόχρονα και ένας κλάδος για την προστασία από αυτά, που βασίζεται πάνω στην εκπαίδευση με τέτοια δείγματα και ονομάζεται **Adversarial Training** ή **Adversarial Learning**. Πιο συγκεκριμένα, adversarial training ονομάζουμε την μέθοδο εκπαίδευσης ενός μοντέλου πάνω σε adversarial examples, με σκοπό να γίνουν πιο εύρωστα σε τέτοιες επιθέσεις ή να μειωθεί το test error πάνω σε καθαρές εισόδους [36].

Το πεδίο της ανταγωνιστικής μηχανικής μάθησης ξεκινάει πολύ πιο πριν από το 2014 με την εισαγωγή των adversarial examples στα DNNs, αλλά από το 2004, όπως αναφέραμε και στην προηγούμενη ενότητα, και σύμφωνα με το [27], η ερευνητική κοινότητα στο πεδίο της ανταγωνιστικής μηχανικής μάθησης στα DNNs έχει καταλήξει να ξαναανακαλύψει ανεξάρτητη πολλά φαινόμενα τα το οποία είχαν ερευνηθεί παλαιότερα σε άλλα πεδία του ML.

Παρακάτω στο κεφάλαιο 3.6 αναφέρονται αναλυτικά οι κατηγορίες και οι μέθοδοι για την προστασία των νευρωνικών από ανταγωνιστικές επιθέσεις, οι οποίες δε βασίζονται μόνο σε τεχνικές adversarial training, αλλά και σε άλλες μεθόδους άλλες πιο ντετερμινιστικές και άλλες πιο ευριστικές (heuristic).

3.3.2 Μαθηματικός ορισμός

Αρχικά οι αλγόριθμοι του adversarial training, μπορούν να οριστούν σαν ένα πρόβλημα βελτιστοποίησης μεγίστου-ελαχίστου (**Minimax**), όπου τα adversarial examples παράγονται για να μεγιστοποιήσουν την απώλεια (loss), ενώ ταυτόχρονα το μοντέλο εκπαιδεύεται για να την μειώσει [37]. Οπότε για να φτιάξουμε εύρωστα μοντέλα, με δεδομένα ζεύγη εισόδων/εξόδων S, θέλουμε να λύσουμε το παρακάτω πρόβλημα ελαχιστοποίησης:

$$minimize_\theta \frac{1}{|S|} \sum_{x,y \in S} \max_{\|\delta\| \leq \epsilon} \ell(h_\theta(x+\delta), y)$$

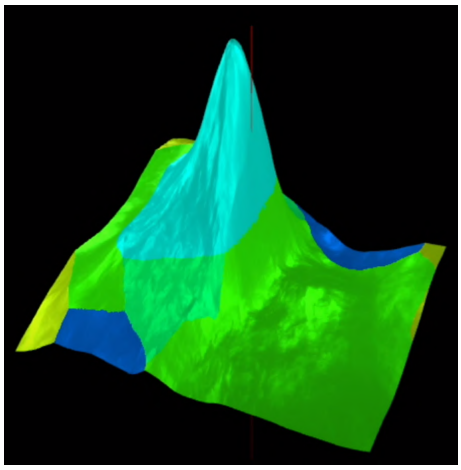
όπου, $h_\theta(x)$ αντιπροσωπεύει ένα νευρωνικό δίκτυο πολλαπλών επιπέδων και ℓ η συνάρτηση κόστους (loss function). Η σειρά των λειτουργιών min-max είναι κρίσιμη, καθώς το max βρίσκεται μέσα στην ελαχιστοποίηση, οπότε ο αντίπαλος (adversary) έχει το πλεονέκτημα να 'κινηθεί' δεύτερος. Ο στόχος αυτής της εύρωστης συνθήκης βελτιστοποίησης (**robust optimization formulation**) είναι να αποτραπούν οι επιθέσεις μοντέλων ακόμα κι αν ο αντίπαλος έχει πλήρη γνώση του μοντέλου (παράμετρος θ). Μπορούν να γίνουν αρκετές υποθέσεις σχετικά με την ισχύ και τη γνώση του αντιπάλου, αλλά είναι δύσκολο να προσδιοριστεί ένας ακριβής ορισμός, επομένως απαιτείται πρόσθετη προσοχή κατά την αξιολόγηση μοντέλων έναντι ρεαλιστικών adversaries [38].

Υπάρχουν 2 τρόποι κυρίως για να λύσουμε κατά προσέγγιση το πρόβλημα της εξωτερικής ελαχιστοποίησης:

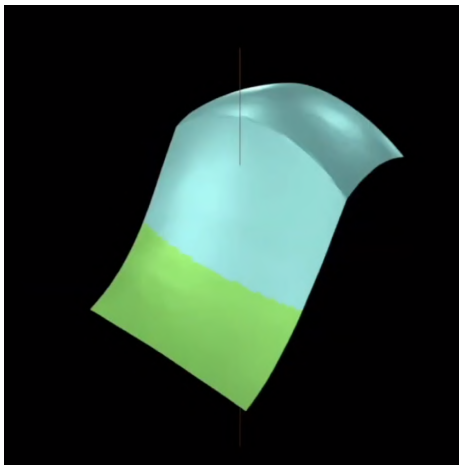
- **Lower-bound solutions**: Χρησιμοποιώντας κατώτερα όρια και παραδείγματα που κατασκευάστηκαν μέσω μεθόδων τοπικής αναζήτησης, για να εκπαιδεύσουμε έναν εμπειρικά (**empirical**) ανταγωνιστικά εύρωστο ταξινομητή.
- **Upper-bound solutions**: Χρησιμοποιώντας κυρτά (convex) άνω όρια, για να εκπαιδεύσουμε έναν αποδεδειγμένα (**certified**) ανταγωνιστικά εύρωστο ταξινομητή.

Σχετικά με τον πρώτο τρόπο επίλυσης, δηλαδή της ενσωμάτωσης adversarial examples στα δεδομένα εκπαίδευσης, χρησιμοποιούνται διάφορες τεχνικές για να βρεθούν ευάλωτα κενά στα μοντέλα, και έτσι δημιουργούνται παραδείγματα τα οποία θα μπορούν να καταλήξουν στην adversarial περιοχή. Η προσθήκη αυτών των δειγμάτων στα δεδομένα εκπαίδευσης έχει ως αποτέλεσμα ένα νέο μοντέλο που δεν ακολουθεί τόσο πιστά τα αρχικά σημεία εκπαίδευσης, δημιουργώντας μια πιο ομαλή και ακριβή λειτουργία [39]. Αυτό φαίνεται και οπτικά στο σχήμα 3.6α′, όπου πρόκειται για το ίδιο παράδειγμα με το σχήμα 3.5, όπου η αρχική συνάρτηση κόστους μετά το adversarial training ιδανικά θα έχει τη μορφή του σχήματος 3.6β′, δηλαδή μια

ομαλή συνάρτηση η οποία δεν έχει εξαφανίσει τελείως το decision boundary, καθώς υπάρχουν ακόμα περιοχές που θα οδηγηθεί σε λάθος απόφαση, αλλά αυτές βρίσκονται ομοιόμορφα προς την ίδια κατεύθυνση. Αυτό το είδος εκπαίδευσης έχει βέβαια έχει και τη δυσκολία ότι το μοντέλο θα πρέπει να εκπαιδευτεί με κάθε νέο είδος adversarial example για να μπορέσει να αντιμετωπίσει ικανοποιητικά νέες επιθέσεις, για αυτό και προσφέρει εμπειρική ασφάλεια και όχι αποδεδειγμένη [30].



(α′) *Διάγραμμα του decision boundary* ***πριν*** *το adversarial training*



(β′) *Διάγραμμα του decision boundary* ***μετά*** *το adversarial training*

Σχήμα 3.6: *Διάγραμμα του decision boundary ενός NN, όπου τα δείγματα που ανήκουν στις μπλε περιοχές αναγνωρίζονται ως σκύλος, πριν και μετά το adversarial training [30]*

Περισσότερα για τις τεχνικές που χρησιμοποιούμε για να εκπαιδεύσουμε και να αξιολογήσουμε αν ένα νευρωνικό δίκτυο είναι **Empirical Robust** ή **Certified Robust**, αναφέρονται στο κεφάλαιο 4.

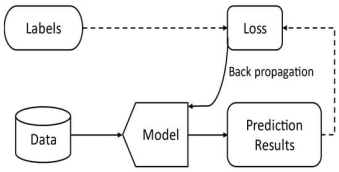
3.3.3 Διαδικασία εκπαίδευσης

Για την ενσωμάτωση της ανταγωνιστικής μάθησης σε ένα μοντέλο μηχανικής μάθησης απαιτούνται δύο βήματα: (i) η δημιουργία adversarial examples και (ii) η ενσωμάτωση αυτών των παραδειγμάτων στη διαδικασία της εκπαίδευσης του νευρωνικού. Ο σκοπός της ανταγωνιστικής μάθησης είναι ο εκπαιδευμένος ταξινομητής να μπορεί να γενικεύει τα **ανταγωνιστικά (adversarial) δείγματα** όσο καλά και τα **καθαρά (clean) δείγματα**.

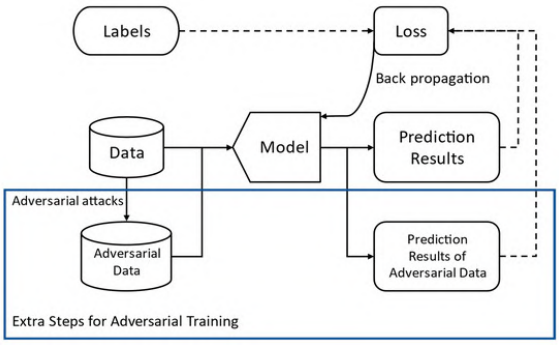
Στη συμβατική διαδικασία μάθησης, τα δεδομένα εκπαίδευσης διαβιβάζονται στο μοντέλο και η πρόβλεψη απώλειας (prediction loss) διαδίδεται εκ των υστέρων μέσω της διαδικασίας του backpropagation για τη βελτίωση των αποτελεσμάτων ταξινόμησης, όπως φαίνεται στο σχήμα 3.7α′.

Το **adversarial training**, επεκτείνει τη συμβατική διαδικασία μάθησης, εισάγοντας ένα επιπλέον βήμα, το οποίο εισάγει adversarial examples είτε στη διαδικασία της εκπαίδευσης ώστε να ανανεώσει τις παραμέτρους του μοντέλου, είτε στα δεδομένα εκπαίδευσης (training data), για να ενισχύσει την ευρωστία του μοντέλου. Τα adversarial examples μπορούν να δημιουργηθούν με διάφορους τρόπους, όπως με τον υπολογισμό της κλίσης της συνάρτησης κόστους που περιγράψαμε παραπάνω αλλά και με άλλες τεχνικές, όπως με χρήση γενετικών αλγορίθμων, ενισχυτικής μάθησης κ.α. Συγκεκριμένα για την πρώτη διαδικασία εκπαίδευσης,

τα adversarial example δημιουργούνται με βάση την τρέχουσα κατάσταση του μοντέλου ή την προηγούμενη παρτίδα (batch) βημάτων εκπαίδευσης για οποιαδήποτε μέθοδο ανταγωνιστικής επίθεσης επιλεχθεί, και τα αποτελέσματα της απώλειας πρόβλεψης των adversarial examples συμπεριλαμβάνονται στα αποτελέσματα των καθαρών δεδομένων, ώστε να διαδοθούν πάλι μέσω backpropagation, όπως φαίνεται και στο σχήμα 3.7β′. Αυτή η απλή και αποτελεσματική διαδικασία, εκπαιδεύει το μοντέλο να προβλέπει σωστές ετικέτες κλάσεων (class labels) τόσο για αρχικά όσο και για adversarial examples, καθιστώντας το μοντέλο πιο ανθεκτικό σε παραλλαγές και παραμορφώσεις δεδομένων [40].



(α′) *Παραδοσιακή εκπαίδευση ενός NN*



(β′) *Ανταγωνιστική εκπαίδευση ενός NN*

Σχήμα 3.7: *Διάγραμμα ροής εκπαίδευσης ενός νευρωνικού παραδοσιακά και με adversarial training [40]*

3.3.4 Αποτελέσματα εκπαίδευσης

3.3.4.1 Πλεονεκτήματα

Το πρώτο αποτέλεσμα που προκύπτει από το adversarial training ενός μοντέλου είναι ότι αυτό γίνεται πιο ανθεκτικό σε adversarial examples και μπορεί να γενικεύσει καλύτερα. Το μοντέλο εκπαιδεύεται σε ένα ευρύ φάσμα παραμορφώσεων, κάτι που το καθιστά πιο ανθεκτικό σε δεδομένα που δεν έχει ξανασυναντήσει και το βοηθάει να αναγνωρίσει και να προσαρμοστεί στην ίδια τη δομή των δεδομένων, κάτι που είναι απαραίτητο και κρίσιμο για την ευρωστία του.

Πιο συγκεκριμένα, σύμφωνα με τα αποτελέσματα του [1], ένα εκπαιδευμένο DNN με το MNIST σύνολο δεδομένων είχε ποσοστό σφάλματος (error rate) 89.4% σε adversarial examples βασισμένα στη Fast Gradient Sign μέθοδο, ενώ μετά από adversarial training έπεσε στο 17.9%.

Μια ακόμη ιδιότητα του adversarial training είναι ότι μπορεί να οδηγήσει σε εξομάλυνση (**regularization**), η οποία προσφέρεται για να περιορίσει την υπερπροσαρμογή (overfitting) ενός μοντέλου, και είναι πολλές φορές ακόμη καλύτερη από άλλες γενικευμένες τεχνικές που υπάρχουν για να κάνουν ακριβώς αυτό, όπως το Dropout [36]. Επίσης, αυτές οι τεχνικές εξομάλυνσης, δεν προσφέρουν καμία μείωση στην ευπάθεια ενός μοντέλου σε adversarial examples, σε αντίθεση με το adversarial training.

3.3.4.2 Προβλήματα

Όμως το adversarial training δημιουργεί και κάποια καινούρια προβλήματα στην απόδοση του μοντέλου, καθώς συνήθως η ακρίβεια (accuracy) του πάνω στα καθαρά δείγματα (βλ. ορισμό clean accuracy στο 4.2.1) δείγματα πέφτει, το οποίο οφείλεται κυρίως στη λειτουργία του

ως regularizer [36]. Συνήθως θα υπάρχει ένας συμβιβασμός (trade-off) μεταξύ της καθαρής και της εύρωστης ακρίβειας (βλ. ορισμό robust accuracy στο 4.2.2), σε ένα μοντέλο εκπαιδευμένο ανταγωνιστικά με τον ίδιο αριθμό δεδομένων. Γενικά, η ανταγωνιστική εκπαίδευση χρειάζεται πολύ μεγαλύτερο αριθμό δεδομένων για να φτάσει ικανοποιητικό αριθμό ακρίβειας, ειδικότερα όσο αυξάνεται και η πολυπλοκότητα ενός συνόλου δεδομένων [40].

Επίσης, είναι δύσκολο να αμυνθούν τα μοντέλα σε adversarial examples, επειδή είναι δύσκολο να κατασκευαστεί ένα θεωρητικό μοντέλο της διαδικασίας δημιουργίας τους. Τα adversarial examples είναι λύσεις σε ένα πρόβλημα βελτιστοποίησης που είναι μη γραμμικό και μη κυρτό για πολλά μοντέλα ML, συμπεριλαμβανομένων των νευρωνικών δικτύων. Λόγω αυτού και της έλλειψης καλών θεωρητικών εργαλείων για την περιγραφή των λύσεων σε αυτά τα περίπλοκα προβλήματα βελτιστοποίησης, είναι πολύ δύσκολο να βρεθεί μία άμυνα που θεωρητικά θα καλύπτει όλες τις πιθανές παραμορφώσεις και δείγματα. Όλες οι τεχνικές adversarial training που έχουν προταθεί δουλεύουν καλά μόνο για τις επιθέσεις που έχουν σχεδιαστεί να προστατεύουν, μη προσφέροντας κάποια ασφάλεια για νέα δείγματα που δεν έχουν βρεθεί κατά την εκπαίδευση. Αυτό κάνει τα μοντέλα ευάλωτα σε προσαρμοστικές (**adaptive**) επιθέσεις και επιτιθέμενους [41].

3.3.4.3 Συμπεράσματα

Όπως φαίνεται και από τα προβλήματα της ανταγωνιστικής εκπαίδευσης, ο σχεδιασμός μιας άμυνας που μπορεί να προστατεύσει από ένα ισχυρό και προσαρμοστικό επιτιθέμενο, ενώ παράλληλα να διατηρεί την ακρίβεια του μοντέλου και την ποσότητα δεδομένων εκπαίδευσης, είναι ακόμα ανοιχτό πεδίο έρευνας.

Αυτό μας οδηγεί στο συμπέρασμα ότι η ανταγωνιστική εκπαίδευση θα πρέπει να χρησιμοποιείται κυρίως σε **2 σενάρια**:

1. Όταν ένα μοντέλο παρουσιάζει υπερπροσαρμογή και χρειάζεται εξομάλυνση, παράλληλα με την ανάγκη προστασίας από adversarial examples
2. Όταν η ασφάλεια απέναντι σε adversarial examples κρίνεται αναγκαία παρά την κάποια πτώση της ακρίβειας, καθώς είναι η πιο αξιόπιστη μέθοδος προφύλαξης απέναντι σε τέτοιες επιθέσεις.

3.4 Ανταγωνιστικές Επιθέσεις

Μέχρι στιγμής έχουμε αναφερθεί στην ευαλωτότητα των ML μοντέλων σε adversarial examples, χωρίς να αναφερθούμε στους διαφορετικούς τρόπους με τους οποίους ένα επιτιθέμενος μπορεί να επιτεθεί ανάλογα με τις εκάστοτε συνθήκες. Επίσης, ο τρόπος παραγωγής adversarial examples που έχουμε περιγράψει αποτελεί τη βάση για άλλες επιθέσεις που έχουν προταθεί και συνεχίζουν να εφευρίσκονται και να έχουν μεγαλύτερη επίπτωση ή πιο γενικευμένη εφαρμογή. Οι ανταγωνιστικές επιθέσεις (**adversarial attacks**) σε ML μοντέλα είναι ένας υπαρκτός κίνδυνος και οι επιθέσεις αυτές αλλάζουν και μπορούν να προσαρμόζονται ανάλογα με τις γνώσεις, τον στόχο και την ικανότητα του επιτιθέμενου, αλλά και ανάλογα το ίδιο το μοντέλο και τα χαρακτηριστικά του, καθώς και το περιβάλλον στο οποίο ορίζεται και τις συνθήκες αυτού.

Σε αυτήν την ενότητα κάνουμε μια αναφορά στις πιο διαδεδομένες ανταγωνιστικές επιθέσεις ενάντια σε ML μοντέλα και νευρωνικά δίκτυα, και γίνεται μια προσπάθεια συγκέντρωσης και

κατηγοριοποίησης τον πιο δημοφιλών ειδών επιθέσεων.

3.4.1 Κατηγοριοποίηση Επιθέσεων

Σε αυτήν την ενότητα γίνεται μια κατηγοριοποίηση και μοντελοποίηση των ανταγωνιστικών απειλών ενάντια στα συστήματα μηχανικής μάθησης. Με την κατάλληλη μοντελοποίηση, μπορούμε να εξάγουμε καλύτερα αποτελέσματα για το μέγεθος και την επίπτωση της κάθε επιθέσεις και αντίστοιχα να εφαρμοστούν τα καλύτερα αντίμετρα για να μετριάσουν αυτούς τους κινδύνους.

3.4.1.1 Ανά Γνώση Επιτιθέμενου

Οι ανταγωνιστικές επιθέσεις ανάλογα με τη γνώση (**knowledge**) του επιτιθέμενου μπορεί να χωριστούν σε:

1. **White-Box Attacks**: Οι επιθέσεις πλήρης γνώσης (**white-box**) πρόκειται για επιθέσεις στις οποίες ο επιτιθέμενος έχει πλήρη γνώση και πρόσβαση στο μοντέλου στο οποίο επιτίθεται. Αυτό σημαίνει ότι έχει πλήρη γνώση της αρχιτεκτονικής, των παραμέτρων και των κλίσεων του μοντέλου. Σε κάποιες περιπτώσεις μπορεί να υπάρχει και πρόσβαση στα δεδομένα και δοκιμής του μοντέλου. Αυτές οι επιθέσεις σχεδιάζονται με βάση τα χαρακτηριστικά του κάθε μοντέλου, αλλά ενώ μπορεί να έχουν μεγάλα ποσοστά επιτυχίας σε ένα, να μην μπορούν να μεταφερθούν σε κάποιο άλλο μοντέλο [27].
2. **Black-Box Attacks**: Οι επιθέσεις χωρίς γνώση (**black-box**) πρόκειται για επιθέσεις στις οποίες ο επιτιθέμενος δεν έχει καμία γνώση της εσωτερικής λειτουργίας του μοντέλου στο οποίο επιτίθεται. Αυτό σημαίνει ότι δε γνωρίζει την αρχιτεκτονικής, τις παραμέτρους του μοντέλου ή στα δεδομένα εκπαίδευσης και δοκιμής. Η μόνη πρόσβαση που διαθέτει είναι στις προβλέψεις ή απαντήσεις του μοντέλου, κατά τη χρήση του. Λόγο αυτών των χαρακτηριστικών οι επιθέσεις αυτές τείνουν να μην έχουν τα ίδια ποσοστά επιτυχίας έναντι white-box επιθέσεων που είναι σχεδιασμένες για ένα μοντέλο, αλλά μπορούν να έχουν μεγαλύτερο εύρος εφαρμογής σε πολλαπλά μοντέλα [27].
3. **Gray-Box Attacks**: Οι επιθέσεις μερικής γνώσης (**gray-box**) είναι ένα ενδιάμεσο των white-box και black-box, καθώς σημαίνει ότι ο επιτιθέμενος έχει κάποια γνώση για το μοντέλο ή τα δεδομένα, άλλα όχι πλήρη γνώση. Δηλαδή μπορεί να μην υπάρχει πρόσβαση στις παραμέτρους του μοντέλου, αλλά πρόσβαση στα δεδομένα εκπαίδευσης ή δοκιμής του. Άλλη περίπτωση θα μπορούσε να είναι, η γνώση του αλγορίθμου που χρησιμοποιείται για το μοντέλο άλλα η μη επίγνωση των εκπαιδευμένων παραμέτρων του ή δεδομένων του [27].

Σχετικά με τον *τρόπο απόκτησης της γνώσης του μοντέλου*, υπάρχουν διάφορες περιπτώσεις όταν πρόκειται να γίνει μοντελοποίηση των απειλών (**threat modeling**) στον πραγματικό κόσμο. Αρχικά, μπορεί το μοντέλο να έχει αρχιτεκτονική ανοιχτού κώδικα (open-source), είτε οι λεπτομέρειες του να είναι δημοσιευμένες, οπότε και να είναι άμεσα διαθέσιμο στους τελικούς χρήστες άρα και στους επιτιθέμενους. Στην περίπτωση που δεν είναι ελεύθερα διαθέσιμο, αλλά οι επιτιθέμενοι έχουν φυσική πρόσβαση στο μοντέλο, μπορεί να αποκτηθεί γνώση, μέσω reversing του κώδικα από τη συσκευή στην οποία τρέχει, η οποία μπορεί να είναι σε υπολογιστές ή κινητά τηλέφωνα των χρηστών ή ακόμα και IoT συσκευές. Ακόμη, αν το μοντέλο προσφέρει δυνατότητα επερωτήσεων (querying) ή πρόσβαση μέσω κάποιας προγραμματιστικής διεπαφής (API), τότε με χρήση τεχνικών model extraction/stealing (βλ. ενότητα

5.2) μπορεί να εξαχθούν πληροφορίες για το μοντέλο από τους επιτιθέμενους με προσεκτικά δημιουργημένα ανταγωνιστικά ερωτήματα.

3.4.1.2 Ανά Ειδικότητα Επιτιθέμενου

Οι ανταγωνιστικές επιθέσεις ανάλογα με την ειδικότητα (**specificity**) του επιτιθέμενου μπορεί να χωριστούν σε:

1. **Targeted Attacks**: Οι στοχευμένες (**targeted**) επιθέσεις ή αλλιώς σκόπιμες (**intentional**) επιθέσεις είναι αυτές στις οποίες ο επιτιθέμενος στοχεύει στο να ταξινομηθεί η είσοδος του εσφαλμένα σε μία καθορισμένη κλάση ή να κλέψει τον υποκείμενο αλγόριθμο [42] [43].
2. **Untargeted Attacks**: Οι μη στοχευμένες (**untargeted**) επιθέσεις ή αλλιώς μη σκόπιμες (**unintentional**) επιθέσεις είναι αυτές στις οποίες ο επιτιθέμενος στοχεύει στο να ταξινομηθεί η είσοδος του εσφαλμένα χωρίς περιορισμούς στο ποια θα είναι η νέα κλάση, απλά για να παραχθεί ένα ανασφαλές αποτέλεσμα [42] [43].

Στις περιπτώσεις των *σκόπιμων επιθέσεων*, μπορούμε να κατατάξουμε κακόβουλες διάφορες επιθέσεις με κυρίαρχο στόχο το μοντέλο ή τα δεδομένα εκπαίδευσης, ενώ στις περιπτώσεις των *μη σκόπιμων επιθέσεων*, μπορούμε να κατατάξουμε πέρα από επιθέσεις και ακούσιες αποτυχίες που μπορεί να συμβούν από μη κακόβουλους χρήστες λόγω κακού σχεδιασμού, μη επαρκών δοκιμών ή έλλειψης προφύλαξης από ανταγωνιστικές επιθέσεις [43].

3.4.1.3 Ανά Στόχο Επιτιθέμενου

Οι ανταγωνιστικές επιθέσεις ανάλογα με τον στόχο (**objective**) του επιτιθέμενου μπορεί να χωριστούν σε:

1. **Integrity violation**: Οι επιθέσεις παραβίασης της ακεραιότητας του συστήματος (**integrity**) αποσκοπούν στο να αποφευχθεί ο εντοπισμός των επιτιθεμένων χωρίς να διακυβευθεί η κανονική λειτουργία του συστήματος [27]. Πρόκειται για την πιο κοινό είδος παραβίασης που προκαλείται από ανταγωνιστικές επιθέσεις, καθώς με τη δημιουργία των adversarial examples είναι σε θέση να παρακάμψουν κρυφά τα υπάρχοντα αντίμετρα και να οδηγήσουν τα μοντέλα σε εσφαλμένη ταξινόμηση, χωρίς να διακυβεύεται η λειτουργικότητα του συστήματος [5].
2. **Availability violation**: Οι επιθέσεις παραβίασης της διαθεσιμότητας του συστήματος (**availability**) αποσκοπούν στο να τεθούν σε κίνδυνο οι συνήθεις λειτουργίες του συστήματος που είναι διαθέσιμες σε κανονικούς χρήστες [27]. Αυτές συμβαίνουν συνήθως όταν παραβιάζεται και η λειτουργικότητα του συστήματος, προκαλώντας άρνηση εξυπηρέτησης (denial of service). Οι παραβιάσεις διαθεσιμότητας επηρεάζουν κυρίως την αξιοπιστία των μοντέλων, αυξάνοντας την αβεβαιότητα των προβλέψεών τους [5].
3. **Privacy violation**: Οι επιθέσεις παραβίασης της ιδιωτικότητας του συστήματος (**privacy**) αποσκοπούν στο να αποκτηθούν προσωπικές πληροφορίες σχετικά με το σύστημα, τους χρήστες ή δεδομένα του [27]. Αυτές συμβαίνουν συνήθως όταν ο επιτιθέμενος μπορεί να αποκτήσει πρόσβαση σε κάποια χαρακτηριστικά του μοντέλου όπως παραμέτρους, αρχιτεκτονική και αλγορίθμους εκμάθησης ή σε δεδομένα εκπαίδευσης του μοντέλου. Αυτό μπορεί να επιτευχθεί είτε με κατάλληλα σχεδιασμένα ερωτήματα στο μοντέλο ή με τεχνικές αντίστροφης μηχανικής (reverse-engineering) ώστε να παραχθεί ένα υποκατάστατο μοντέλου (**surrogate**) το οποίο προσομοιάζει τη λειτουργία

του αρχικού μοντέλου και την αρχική κατανομή δεδομένων. [5].

Όλες οι παραπάνω παραβιάσεις ασφαλείας μπορούν να επιτευχθούν μεμονωμένα ή και συνολικά ανάλογα με τις επιθέσεις που θα επιλεχθούν. Στις *επιθέσεις κατά της ακεραιότητας του συστήματος*, κατατάσσονται και οι περισσότερες επιθέσεις με adversarial examples, π.χ. επιθέσεις που προσπαθούν να δημιουργήσουν false positives σε ένα σύστημα αναγνώρισης προσώπων [44], καθώς ο σκοπός τους είναι η αστοχία πρόβλεψης ή εσφαλμένης ταξινόμησης, χωρίς να θέσουν απαραίτητα τη λειτουργία του συστήματος σε κίνδυνο. Για τις *επιθέσεις κατά τις ιδιωτικότητας του συστήματος* γίνεται εκτενής ανάλυση στο κεφάλαιο 5.

3.4.1.4 Ανά Επιρροή Επιτιθέμενου

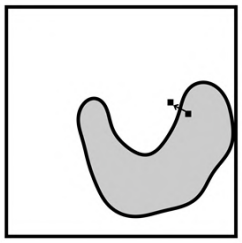
Οι ανταγωνιστικές επιθέσεις ανάλογα με την επιρροή (**influence**) ή ικανότητα (**capability**) του επιτιθέμενου μπορεί να χωριστούν σε:

1. **Evasion Attacks**: Οι επιθέσεις εισβολής (**evasion**) είναι επιθέσεις που συμβαίνουν κατά τη δοκιμή (testing) ή κατά το τρέξιμο (inference) του μοντέλου, και ο επιτιθέμενος έχει σκοπό να χειριστεί τα δεδομένα εισόδου για να δημιουργήσει ένα σφάλμα σε ένα ML σύστημα [27].
2. **Poisoning Attacks**: Οι επιθέσεις δηλητηρίασης (**poisoning**) είναι επιθέσεις που συμβαίνουν κατά την εκπαίδευση (training) του μοντέλου, και ο επιτιθέμενος έχει σκοπό να εισάγει μικρά δείγματα παραποιημένων δεδομένων στα δεδομένα εκπαίδευσης για να αυξήσει τα λανθασμένα ταξινομημένα δείγματα κατά τον χρόνο της δοκιμής και να μειώσει τη συνολική απόδοση του μοντέλου ή να εισαγάγει backdoors τα οποία θα αξιοποιηθούν όταν δοθεί η κατάλληλη είσοδος κατά το τρέξιμο του μοντέλου από τον επιτιθέμενο [27].

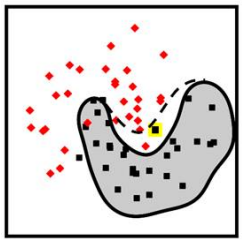
Η βασική διαφορά μεταξύ των δυο αυτών ειδών είναι ότι στις evasion επιθέσεις οι επιτιθέμενοι έχουν πρόσβαση στο μοντέλο, ενώ στις poisoning επιθέσεις έχουν και τη δυνατότητα να το τροποποιήσουν [4]. Ένας άλλος τρόπος για να κατανοήσουμε τη διαφορά μεταξύ των δύο τύπων επίθεσης είναι μέσω του *ορίου απόφασης του μοντέλου*. Οι poisoning επιθέσεις μετατοπίζουν το όριο απόφασης προς την κατεύθυνση που θέλει ο επιτιθέμενος, ενώ οι evasion επιθέσεις μετατοπίζουν τα σημεία δεδομένων κατά μήκος του ορίου απόφασης με τρόπους είναι δύσκολο να εντοπιστούν [45], όπως φαίνεται και στο σχήμα 3.8. Οι poisoning επιθέσεις έχουν αρκετές ομοιότητες με το πρόβλημα του **distribution shift** που εμφανίζεται στα περισσότερα ML μοντέλα, όπου συνήθως με την πάροδο του χρόνου, η κατανομή των δεδομένων που βλέπουν κατά το inference δεν ταιριάζει με τα δεδομένα που εκπαιδεύτηκαν [45].

Επίσης, και στις δύο περιπτώσεις επιθέσεων, αυτές μπορεί να είναι targeted ή untargeted, δηλαδή για τις evasion επιθέσεις να στοχεύουν στην ταξινόμηση ενός ανταγωνιστικού δείγματος σε συγκεκριμένη κλάση ή όχι, και αντίστοιχα για τις evasion επιθέσεις τα ανταγωνιστικά δείγματα εκπαίδευσης να οδηγούν στην αλλαγή του ορίου απόφασης προς συγκεκριμένη κατεύθυνση ή όχι [27]. Ειδικότερα, στις περιπτώσεις των poisoning επιθέσεων, τα δηλητηριασμένα δεδομένα μπορεί να προκύψουν είτε σκόπιμα από επιτιθέμενους που έχουν τη δυνατότητα να επηρεάσουν τα δεδομένα εκπαίδευσης, είτε μη σκόπιμα από μη ελεγμένα δεδομένα εκπαίδευσης (π.χ. δημόσια δεδομένα από μη έγκυρες πηγές) τα οποία θα εκπαιδεύσουν ένα αναξιόπιστο μοντέλο [43].

Τέλος, ανάλογα με τον στόχο του επιτιθέμενου και την επιρροή του, θα καθοριστεί το είδος της επίθεσης και της επίπτωσης. Για παράδειγμα, ένας adversary που έχει ως στόχο να



(α′) *evasion επίθεση*



(β′) *poisoning επίθεση*

Σχήμα 3.8: *Διάγραμμα ορίων απόφασης σε evasion και poisoning επιθέσεις. Στην πρώτη περίπτωση ένα δείγμα γίνεται διαταραγμένο ώστε να βρεθεί πέρα από το όριο απόφασης, ενώ στη δεύτερη περίπτωση εισάγονται νέα δεδομένα εκπαίδευσης που θα μετατοπίσουν το όριο απόφασης προς τη διακεκομμένη γραμμή στο νέο μοντέλο [45]*

παραβιάσει την ακεραιότητα του συστήματος, αν έχει πρόσβαση στα δεδομένα εκπαίδευσης τότε πρόκειται για μια backdoor poisoning επίθεση, ενώ έχει πρόσβαση μόνο στο μοντέλο, τότε πρόκειται για μια evasion επίθεση με χρήση adversarial examples. Η πλήρης κατηγοριοποίηση των επιθέσεων ανά στόχο και επιρροή του επιτιθέμενου, φαίνεται στον πίνακα 3.1.

	Integrity	**Availability**	**Privacy**
Test data	Evasion (adversarial examples)	-	Model Extraction, Model Stealing, Model Inversion
Training data	Poisoning (backdooring)	Poisoning (maximize classification error)	-

Πίνακας 3.1: *Κατηγοριοποίηση ML επιθέσεων με βάση το καθορισμένο μοντέλο απειλής. Οι* ***γραμμές*** *καθορίζουν την επιρροή του ή ικανότητα του επιτιθέμενου (**Attacker's Capability**) και οι* ***στήλες*** *τον στόχο του επιτιθέμενου (**Attacker's Goal**) [27].*

3.4.2 Τεχνικές Επιθέσεων (White-Box)

Σε αυτήν την ενότητα παρουσιάζουμε τις πιο δημοφιλείς white-box evasion ανταγωνιστικές επιθέσεις από τη βιβλιογραφία. Όλες αυτές οι επιθέσεις προϋποθέτουν γνώση του μοντέλου (π.χ. τις παραμέτρους του, τις κλίσεις του, τη συνάρτηση κόστους, κ.λπ.). Στον πίνακα 3.2 μπορούμε να βρούμε συγκεντρωτικά όσες επιθέσεις αναλύουμε εδώ αλλά και άλλες state-of-the-art (white-box, black-box) από τη βιβλιογραφία.

3.4.2.1 Fast Gradient Sign Method (FGSM)

Η **Fast Gradient Sign Method (FGSM)** [1], πρόκειται για μια από τις πρώτες και δημοφιλέστερες ανταγωνιστικές επιθέσεις που προτάθηκαν και χρησιμοποιούνται για να ξεγελάσουν νευρωνικά δίκτυα, η οποία μπορεί να λειτουργήσει σαν **targeted** και σαν **untargeted** επίθεση και μπορεί να λειτουργήσει για μεγέθη ℓ_∞, ℓ_1, ℓ_2 adversarial perturbation [42]. Η FGSM, λειτουργεί χρησιμοποιώντας τις κλίσεις ενός νευρωνικού δικτύου για να δημιουργήσει τα adversarial examples. Για μια εικόνα εισόδου, η μέθοδος υπολογίζει την κλίση της συνάρτησης κόστους σε σχέση με τα εικόνα εισόδου, για να δημιουργήσει μια νέα εικόνα η οποία μεγιστοποιεί την απώλεια, την οποία ονομάζουμε ανταγωνιστική εικόνα (adversarial image)

[1]. Το παραπάνω συνοψίζεται με την παρακάτω μαθηματική έκφραση:

$$\tilde{x} = x + \epsilon sign(\nabla_x J(\theta, x, y))$$

όπου όπως είδαμε και στην ενότητα 3.2.2 για τα adversarial examples, το $\tilde{x}$ συμβολίζει το παραγόμενο adversarial example, x την αρχική είσοδο, y την ετικέτα της αρχικής εισόδου x, ϵ το threshold που διασφαλίζει ότι οι διαταραχές είναι μικρές, θ τις παραμέτρους ενός μοντέλου, και $J(\theta, x, y)$ τη συνάρτηση κόστους του νευρωνικού.

Στην ουσία η μέθοδος διασχίζει την καμπύλη κόστους κινούμενη προς την αντίθετη κατεύθυνση της κλίσης της συνάρτησης κόστους, λειτουργώντας ως μια αντίστροφη μέθοδος βελτιστοποίησης. Επίσης, πρόκειται για προσέγγισης της πραγματικής κλίσης καθώς είναι πιο αποδοτικός ο υπολογισμός του πρόσημου της κλίσης από τον υπολογισμό της πραγματική κλίσης [39]. Το πρόσημο ('Sign') στο όνομα της επίθεσης, αναφέρεται σε συγκεκριμένο υπολογισμό του ℓ_∞ μεγέθους για το οποίο λειτουργεί βέλτιστα, καθώς για τον υπολογισμό των ℓ_1, ℓ_2 μεγεθών συνήθως η επίθεση αναφέρεται ως FGM [42].

Το πλεονέκτημα της, είναι ότι είναι πολύ αποδοτική για να υπολογιστεί, καθώς απαιτείται μόνο μια αξιολόγηση της κλίσης και η επίθεση μπορεί να εφαρμοστεί απευθείας σε μια παρτίδα (batch) εισόδων [42], αλλά συνήθως δεν είναι τόσο ισχυρή όσο άλλες μέθοδοι που αναλύουμε παρακάτω [5].

Η FGSM όντας και η πρώτη μέθοδος παραγωγής adversarial example δείχνει την ευθραυστότητα των ML συστημάτων, ειδικά μέσω και τον χαρακτηριστικών παραδειγμάτων όπως στην εικόνα 3.3 με το πάντα.

3.4.2.2 Basic Iterative Method (BIM)

Η **Basic Iterative Method (BIM)** [36] είναι μια ευθύς επέκταση του FGSM, στην οποία εκτελούνται πολλαπλές επαναλήψεις αντί για μία, τα οποία ονομάζει βήματα α, όπου σε κάθε βήμα διαταράσσει τα δεδομένα εισόδου κατά μικρή ποσότητα προς την κατεύθυνση της κλίσης. Τυπικά, αν θέσουμε ως πρώτη συνθήκη $\tilde{x}_0 = x$, τότε έχουμε για κάθε επανάληψη:

$$\tilde{x}_{N+1} = Clip_{x,\epsilon}\{\tilde{x}_N + \alpha sign(\nabla_x J(\tilde{x}_N, y_{true}))\}$$

όπου, η συνάρτηση $Clip$, εφαρμόζει αποκοπή ανά pixel σε κάθε ανταγωνιστική εικόνα $\tilde{x}$, περιορίζοντας τις τιμές στη γειτονιά (ℓ_∞, ϵ) της αρχικής εικόνας x.

Στη BIM γίνεται προσπάθεια αύξηση του κόστους ταξινόμησης στη σωστή κλάση χωρίς να καθορίζεται σε ποια λάθος κλάση θα γίνει η λάθος ταξινόμηση (**untargeted** μέθοδος), υπάρχει και μια παραλλαγή της μεθόδου που ονομάζεται **Iterative Targeted Fast Gradient Sign Method (IT-FGSM)** ή αλλιώς **Iterative least-likely class method** [46] στην οποία προσπαθεί να γίνει ταξινόμηση προς τη λιγότερο πιθανή κλάση (least-likely class). Αυτή η λιγότερο πιθανή κλάση για μία είσοδο x που έχει πραγματική ετικέτα y, ορίζεται ως:

$$y_{LL} = arg\min_y \{p(y|x)\}$$

Οπότε αντίστοιχα με πριν, αν θέσουμε ως πρώτη συνθήκη $\tilde{x}_0 = x$, τότε έχουμε για κάθε επανάληψη:

$$\tilde{x}_{N+1} = Clip_{x,\epsilon}\{\tilde{x}_N - \alpha sign(\nabla_x J(\tilde{x}_N, y_{LL}))\}$$

Αυτές οι μέθοδοι, έχει παρουσιάσει πολύ καλύτερα αποτελέσματα από την απλή της έκδοση, αλλά είναι πιο αργές λόγω των επαναλήψεων που χρειάζονται για την παραγωγή των adversarial examples [26].

3.4.2.3 Projected Gradient Descent (PGD)

Η **Projected Gradient Descent (PGD)** [47] είναι και αυτή μια επαναληπτική επέκταση του FGSM, αρκετά παρόμοια με τη BIM. Η κύρια διαφορά με την BIM είναι ότι η PGD προσπαθεί να φτιάξει ένα adversarial example $\tilde{x}$ από μια είσοδο x που να ικανοποιεί τη συνθήκη $\|\tilde{x} - x\|_p \leq \epsilon$, δηλαδή να προβάλλει το αποτέλεσμα της επίθεσης πίσω στη σφαίρα (ℓ_p, ϵ) γύρω από την αρχική είσοδο σε κάθε επανάληψη [42]. Τυπικά, αν ορίσουμε B τη ℓ_∞ σφαίρα ακτίνας ϵ με κέντρο το x, η επίθεση ξεκινάει από ένα τυχαίο σημείο $x_0 \in B$ και θέτει επαναληπτικά [48]:

$$x_{i+1} = Proj_B(x_i + asign(\nabla_{x_i} J(x_i, y)))$$

όπου $Proj_B$ είναι η προβολή της εισόδου στη σφαίρα B. Σε μια πιο γενική περίπτωση, όπου ορίζουμε B τη ℓ_p σφαίρα ακτίνας ϵ με κέντρο το x, τότε η μετακίνηση προς την κατεύθυνση με τη μεγαλύτερη απώλεια ορίζεται ως:

$$x_{i+1} = Proj_B(x_i + a \cdot \underset{\|v\|_p \leq 1}{arg\max}\, v^\top \nabla_{x_i} J(x_i, y))$$

3.4.2.4 Jacobian Saliency Map Attack (JSMA)

Η **Jacobian Saliency Map Attack (JSMA)** [49] είναι μια **targeted** επίθεση που προσδιορίζει τα πιο σημαντικά χαρακτηριστικά των δεδομένων εισόδου σε σχέση με τις εξόδους, αναλύοντας τον πίνακα Τζακόμπι (Jacobian matrix), και τροποποιεί αυτά τα χαρακτηριστικά για να δημιουργήσει adversarial examples.

Πρόκειται για μια gradient-based επίθεση που στόχο έχει να ελέγξει το ℓ_∞ μέγεθος, δηλαδή των αριθμό των συνιστωσών της εισόδου x που θα τροποποιηθούν για να φτιαχτεί το adversarial example $\tilde{x}$. Χρησιμοποιεί την κλίση για να υπολογίσει μια βαθμολογία σημαντικών χαρακτηριστικών (saliency score) για κάθε pixel, όπου η βαθμολογία αυτή αντιπροσωπεύει το πόσο ισχυρό είναι το κάθε pixel για να επηρεάσει την ταξινόμηση. Αφού φτιαχτεί ένας χάρτης των σημαντικών αυτών χαρακτηριστικών (saliency map) μέσω του Τζακόμπι πίνακα του μοντέλου, η επίθεση προσπαθεί με άπληστο τρόπο (greedy) να τροποποιήσει το πιο σημαντικό pixel σε κάθε επανάληψη, ένα κάθε φορά, έως ότου επιτευχθεί είτε η στοχευμένη εσφαλμένη ταξινόμηση είτε ο συνολικός αριθμός των τροποποιημένων στοιχείων υπερβεί έναν καθορισμένο προϋπολογισμό [50].

Τυπικά, ο συγκεκριμένος αλγόριθμος περιλαμβάνει τα εξής βήματα:

1. Υπολογισμός της κλίσης: $\nabla Z(X)_l = \frac{\partial Z(X)_l}{\partial X} = \left[\frac{\partial Z_j(X)_l}{\partial x}\right]_{i \in 1..M, j \in 1..N}$, όπου Z η συνάρτηση που μαθαίνει το μοντέλο N-διαστάσεων, X η είσοδος M-διαστάσεων, και l η κλάση στόχος.
2. Κατασκευή του saliency map με βάση τις υπολογισμένες κλίσεις του βήματος 1

3. Τροποποίηση του πιο σημαντικού pixel επαναληπτικά μέχρι την ταξινόμηση της εικόνας στη στοχευμένη κλάση.

3.4.2.5 Carlini & Wagner (C&W)

Η **Carlini & Wagner (C&W)** [26] είναι μια **targeted** επίθεση (υπάρχει και **untargeted** έκδοση) η οποία δημιουργήθηκε αρχικά για να αντιμετωπίσει την τότε πιο δυνατή τεχνική άμυνας Defense Distillation [51], αλλά έχει γενικότερη εφαρμογή και είναι μια από τις πιο ισχυρές (state-of-the-art) white-box ανταγωνιστικές επιθέσεις. Η επίθεση αυτή έχει 3 διαφορετικές μορφές, η κάθε μια διαμορφωμένη για να δουλεύει για ℓ_0, ℓ_2, και ℓ_∞ μεγέθη διαταραχής. Μάλιστα, η ℓ_∞ επίθεση πρόκειται για την πρώτη δημοσιευμένη που κατάφερε να προκαλέσει εσφαλμένη ταξινόμηση στο ImageNet.

Αυτές οι επιθέσεις διατυπώνουν το πρόβλημα της ανταγωνιστικής επίθεσης ως ένα πρόβλημα βελτιστοποίησης, όπου αντί να μεγιστοποιήσουν μια συνάρτηση κόστους κάτω από ένα δεδομένο περιορισμό διαταραχής (όπως γίνεται με π.χ. το PGD), στοχεύουν στο να βρουν το μικρότερο επιτυχημένο adversarial perturbation [48].

Τυπικά, η επίθεση μπορεί να μοντελοποιηθεί για όλα τα είδη ℓ_p επιθέσεων (ℓ_0, ℓ_2, ℓ_∞) ως:

$$\text{minimize } \|\delta\|_p + c \cdot f(x+\delta) \text{ such that } x+\delta \in [0,1]^n$$

όπου $\delta = x - \tilde{x}$ η μικρή διαταραχή, και το $c > 0$ είναι μια προσεκτικά διαλεγμένη σταθερά η οποία μπορεί να βρεθεί με μια τροποποιημένη δυαδική αναζήτηση.

Για να είναι βέβαιο ότι η τροποποίηση θα παράξει μια έγκυρη εικόνα, υπάρχει ο εξής περιορισμός στο δ: $0 \leq x_i + \delta_i \leq 1$ για κάθε i (box constraint). Υπάρχουν πολλοί τρόποι να λυθεί το παραπάνω πρόβλημα βελτιστοποίησης, αλλά η C&W επίθεση επιλέγει τον Adam optimizer [52], καθώς βρέθηκε να είναι ο πιο αποτελεσματικός στην εύρεση των adversarial examples.

3.4.2.6 DeepFool

Η **DeepFool** [53] είναι μια **untargeted** επίθεση που βρίσκει επαναληπτικά την ελάχιστη διαταραχή που χρειάζεται για να μετακινηθεί ένα δείγμα εισόδου πέρα από το όριο απόφασης του μοντέλου, βελτιστοποιημένη για ℓ_2 μεγέθη, μετατοπίζοντας σταδιακά την είσοδο προς την εσφαλμένη ταξινόμηση. Η βασική ιδέα είναι ότι το DeepFool προσεγγίζει, για κάθε επανάληψη, τη λύση αυτού του προβλήματος θεωρώντας ότι ο ταξινομητής είναι γραμμικός και άρα βρίσκει τη βέλτιστη λύση στο απλοποιημένο πρόβλημα, κατασκευάζοντας το adversarial example. Όμως, επειδή τα νευρωνικά δίκτυα δεν είναι στην πραγματικότητα γραμμικά, γίνονται βήματα προς τη βέλτιστη λύση, και επαναλαμβάνεται η διαδικασία μέχρι να βρεθεί εάν πραγματικό adversarial example, το οποίο θα το κάνει να διασχίσει το όριο απόφασης [26].

Τυπικά, για την περίπτωση των δυαδικών ταξινομητών (binary classifiers), θεωρώντας $f(x) = w^T x + b$ ο γραμμικός δυαδικός ταξινομητής, η παραμόρφωση αυτού ορίζεται ως

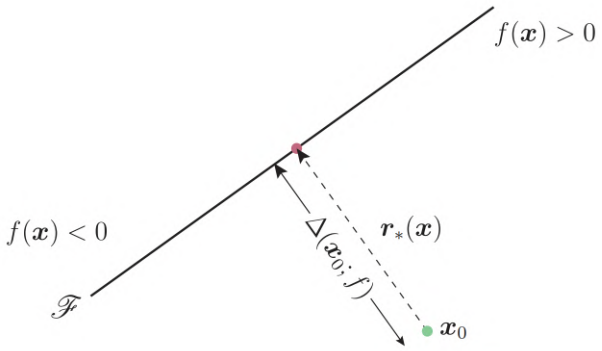
$$r_*(x) = -\frac{f(x)}{\|w\|_2^2} w$$

Τώρα για έναν γενικό δυαδικό διαφορίσιμο (μη γραμμικό) ταξινομητή (binary differentiable classifier) f, χρησιμοποιείται μια επαναληπτική μέθοδο για να προσεγγισθεί η παραμόρ-

φωση θεωρώντας ότι ο f είναι γραμμικός σχετικά με τις εισόδους σε κάθε επανάληψη και υπολογίζεται η ελάχιστη παραμόρφωση ως εξής:

$$\underset{r_i}{arg\,\min} \|r_i\|_2 \text{ subject to } f(x_i) + \nabla f(x_i)^T r_i = 0$$

Αντίστοιχα, αυτή η διαδικασία μπορεί να επεκταθεί και για ταξινομητές πολλών κλάσεων.



Σχήμα 3.9: *Σχηματική αναπαράσταση εύρεσης adversarial example για έναν γραμμικό δυαδικό ταξινομητή [53]*

Πρόκειται για μια αποδοτική μέθοδο που υπολογίζει αποτελεσματικά τις διαταραχές που παραπλανούν τα DNN και προκύπτει ότι είναι επίσης μια αξιόπιστη μέθοδος αξιολόγησης της ευρωστίας αυτών των ταξινομητών.

3.4.2.7 Universal Adversarial Perturbations

Η **Universal Adversarial Perturbations** [54] είναι μια ειδικού τύπου **untargeted** επίθεση, η οποία παράγει ένα μοναδικό adversarial example για να προκαλέσει εσφαλμένη ταξινόμηση σε κάθε δεδομένο από το σύνολο δεδομένων, καθιστώντας την εξαιρετικά μεταβιβάσιμη σε διαφορετικά δείγματα.

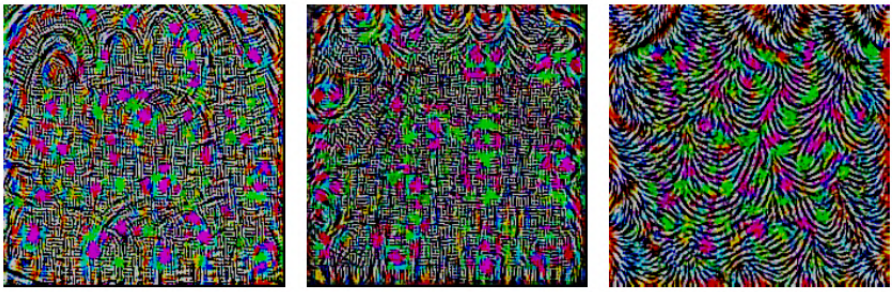
Η επίθεση λειτουργεί, δημιουργώντας μια σταθερή διαταραχή υ το οποίο αλλάζει με επιτυχία την ταξινόμηση ενός συγκεκριμένου κλάσματος εισόδων. Αυτό η καθολική διαταραχή δημιουργείται χρησιμοποιώντας μια μη στοχευμένη επίθεση, όπου για όσο διάστημα δεν έχει επιτευχθεί ο στόχος της παραπλάνησης ή δεν έχει επιτευχθεί ο μέγιστος αριθμός επαναλήψεων, ο αλγόριθμος προσαρμόζει επαναληπτικά την καθολική διαταραχή προσθέτοντας βελτιώσεις που βοηθούν στη διαταραχή πρόσθετων δειγμάτων από το σύνολο εισόδου. Μετά από κάθε επανάληψη, η καθολική διαταραχή προβάλλεται σε σφαίρα ℓ_p με ακτίνα ϵ για να ελεγχθεί η δύναμη επίθεσης [42].

Τυπικά, το πρόβλημα μοντελοποιείται ως η εύρεση ενός καθολικού διανύσματος παραμόρφωσης υ που να ικανοποιεί τις εξής συνθήκες:

$$\|\upsilon\|_p \leq \xi$$

$$\underset{x\sim\mu}{\mathbb{P}}(\widehat{k}(x+\upsilon) \neq \widehat{k}(x)) \geq 1-\delta$$

όπου $\widehat{k}$ ορίζει μια συνάρτηση ταξινόμησης που βγάζει για κάθε εικόνα εισόδου $x \in \mathbb{R}$ μια εκτιμώμενη ετικέτα $\widehat{k}(x)$, ξ περιορίζει το μέγεθος της διαταραχής του διανύσματος υ,



(a) CaffeNet (b) VGG-F (c) VGG-16

Σχήμα 3.10: *Παράδειγμα universal perturbations υπολογισμένα για διαφορετικές DNN αρχιτεκτονικές για* $p = \infty$ *[54]*

και δ καθορίζει το ποσοστό αποτυχίας των adversarial δειγμάτων πάνω σε δείγματα εικόνων κατανομής $\mu \in \mathbb{R}$.

Τα universal perturbations δείχνουν να επιδρούν δραστικότερα από κάθε άλλο τύπο (όπως π.χ. random perturbation), αφού αξιοποιούν γεωμετρικές συσχετίσεις μεταξύ διαφόρων σημείων του ορίου απόφασης (decision boundary) του εκάστου ταξινομητή [54].

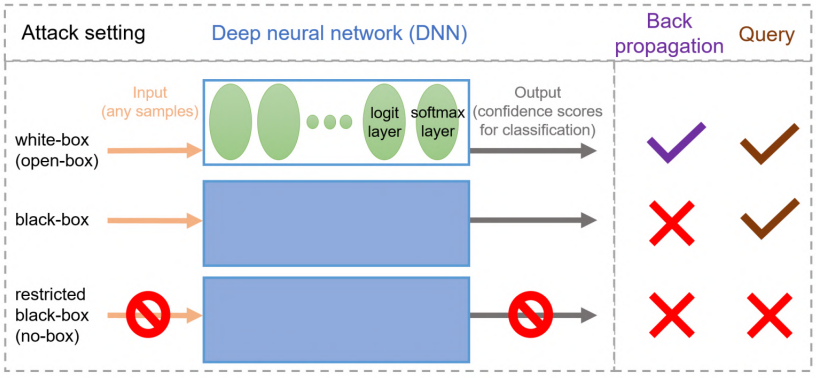
3.4.3 Τεχνικές Επιθέσεων (Black-Box)

Σε αυτήν την ενότητα παρουσιάζουμε ενδεικτικά κάποιες από τις πιο δημοφιλείς black-box evasion ανταγωνιστικές επιθέσεις από τη βιβλιογραφία. Καμία επίθεση δεν απαιτεί γνώση του μοντέλου και συνήθως ακολουθούνται οι εξής προσεγγίσεις για την παραγωγή black-box adversarial δειγμάτων: (i) μέθοδοι βασισμένοι σε μεταφορά (**transfer-based**) με κατασκευή υποκατάστατου μοντέλου (surrogate ή substitute model), (ii) μέθοδοι βασισμένοι σε βαθμολογία (**score-based**), (iii) μέθοδοι βασισμένοι σε αποφάσεις (**decision-based**) [55]. Στον πίνακα 3.2 μπορούμε να βρούμε συγκεντρωτικά όσες επιθέσεις αναλύουμε εδώ αλλά και άλλες state-of-the-art (white-box, black-box) από τη βιβλιογραφία.

3.4.3.1 Zeroth Order Optimization Attack (ZOO)

Η **Zeroth Order Optimization Attack (ZOO)** [56], πρόκειται για μια από τις πρώτες black-box επιθέσεις, η οποία μπορεί να λειτουργήσει σαν **targeted** και σαν **untargeted** επίθεση, και μπορεί να θεωρηθεί ως μια black-box έκδοση της C&W επίθεσης, η οποία όμως δε χρησιμοποιεί υποκατάστατο μοντέλο, αλλά βασίζεται σε ερωτήματα σχετικά με τις πιθανότητες εξόδου του ταξινομητή.

Όπως φαίνεται και στο σχήμα 3.11, σε ένα black-box περιβάλλον επιτρέπεται η άσκηση ερωτήσεων ή εισαγωγή δειγμάτων και η παρατήρηση της αντίστοιχης εξόδου ή αποτελέσματος, αλλά δεν υπάρχει καμία πρόσβαση στις λεπτομέρειες του μοντέλου, το οποίο μπορεί να είναι μια πολύ συχνή περίπτωση, εφόσον οι περισσότερες υπηρεσίες που χρησιμοποιούν ML μοντέλα, δε μοιράζονται τις λεπτομέρειες τους, αλλά είναι προσβάσιμα για χρήση. Οπότε οι adversaries μέσω κατάλληλων επερωτήσεων, μπορούν να εκπαιδεύσουν ένα υποκατάστατο μοντέλο (**surrogate ή substitute model**) το οποίο να είναι αντιπροσωπευτικό του στοχευμένου ML μοντέλου. Σε αυτό το υποκατάστατο μοντέλο μπορούν να χρησιμοποιηθούν οποιεσδήποτε white-box τεχνικές επίθεσης και οι παραγόμενες ανταγωνιστικές εικόνες χρησιμοποιούνται μετέπειτα για επίθεση στο πραγματικό μοντέλο στόχο.



Σχήμα 3.11: *Κατηγοριοποίηση των adversarial attacks ανάλογα τη γνώση του επιτιθέμενου (white-box, black-box). Με τον όρο Back propagation εννοείται η δυνατότητα πρόσβασης στην αρχιτεκτονική του μοντέλου, ενώ με τον όρο Query εννοείται η δυνατότητα εισαγωγής δείγματος εισόδου και παρατήρησης της αντίστοιχης εξόδου [56]*

Το μεγάλο πλεονέκτημα της εκπαίδευσης ενός υποκατάστατου μοντέλου είναι η πλήρης διαφάνεια σε έναν adversary και ως εκ τούτου όλες οι βασικές διαδικασίες που χρησιμοποιούνται στις white-box επιθέσεις (πχ. υπολογισμός κλίσεων), μπορούν να εφαρμοστούν στο υποκατάστατο μοντέλο. Επίσης, εφόσον τα adversarial examples που παράγονται μπορούν να μεταφερθούν σε μεγάλο βαθμό στο μοντέλο στόχος και άρα χαρακτηρίζονται ως **highly transferable**.

Η ZOO επίθεση όμως, παρουσιάζεται ως ένα πρόβλημα βελτιστοποίησης με χρήση τεχνικών βελτιστοποίησης μηδενικής τάξης (zeroth order optimization), χωρίς να χρειάζεται να εκπαιδεύσει κάποιο substitute model, επιτρέποντας έτσι ένα ψευδό-backpropagation στο μοντέλο στόχος. Η επίθεση λοιπόν, λειτουργεί με τον ίδιο τρόπο όπως και μια white-box επίθεση, και το πλεονέκτημά της σε σχέση με τις άλλες black-box μεθόδους είναι ότι αποφεύγει οποιαδήποτε πιθανή απώλεια στη μεταφορά από ένα υποκατάστατο μοντέλο.

Ενώ η χρήση μεθόδων μηδενικών τάξεων είναι διαισθητική, η αφελής χρήση του σε μεγάλα μοντέλα και σύνολα δεδομένων είναι μη πρακτική. Το ZOO όμως διαθέτοντας τεχνικές όπως μείωση διάστασης του χώρου επίθεσης, ιεραρχικών επιθέσεων και δειγματοληψία σπουδαιότητας (importance sampling), καταφέρνει να επιδείξει αξιόλογες επιδόσεις σε μεγάλα σύνολα δεδομένων, όπως το ImageNet, όταν άλλες black-box τεχνικές που βασίζονται σε υποκατάστατα μοντέλα έχουν ικανοποιητικές επιδόσεις σε μικρά μοντέλα και σύνολα δεδομένων, όπως το MNIST.

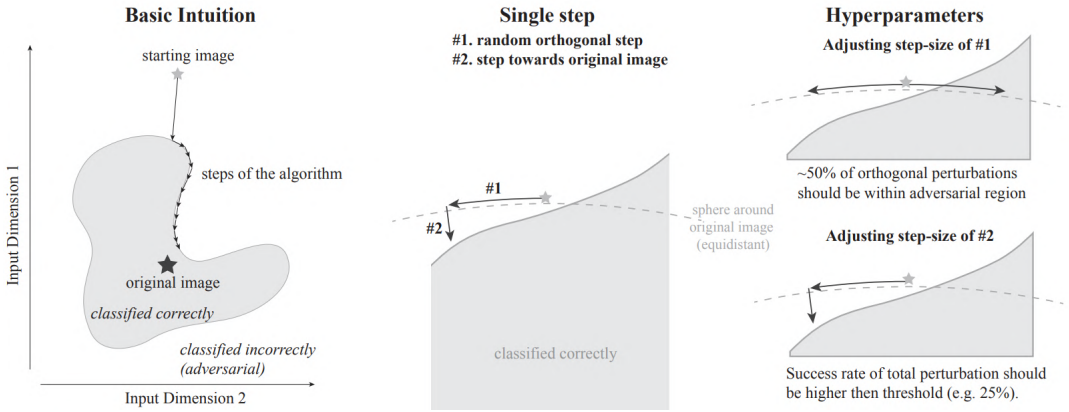
3.4.3.2 Boundary Attack (BA)

Η **Boundary Attack (BA)** [57], πρόκειται για μια black-box επίθεση που μπορεί να λειτουργήσει σαν **targeted** και σαν **untargeted** επίθεση, και απαιτεί μόνο ερωτήματα της κλάσης εξόδου, και όχι των logit (οι έξοδοι ενός NN πριν εφαρμοστεί η συνάρτηση ενεργοποίησης) ή των πιθανοτήτων εξόδου του μοντέλου.

Η συγκεκριμένη άμυνα δε βασίζεται ούτε σε βαθμολογίες, ούτε σε μεταφορές επιθέσεων, αλλά μόνο σε αποφάσεις, και συγκεκριμένα στην τελική απόφαση του μοντέλου (**decision-based**). Αυτό κάνει τις decision-based επιθέσεις: (i) πιο σχετικές σε σχέση με τις score-based επιθέσεις, καθώς σε πραγματικά περιβάλλοντα είναι σπάνια προσβάσιμες οι βαθμολογίες

εμπιστοσύνης ή τα logit των μοντέλων, (ii) είναι πιο ισχυρές απέναντι σε τυπικές άμυνες, όπως gradient masking ή robust training, από άλλες επιθέσεις, και (iii) χρειάζονται πολύ λιγότερες πληροφορίες για το μοντέλο σε σχέση με τις transfer-based επιθέσεις και είναι πιο απλές στην εφαρμογή τους.

Η Boundary Attack είναι μια απλή επίθεση που ξεκινάει από ένα μεγάλο adversarial perturbation και στη συνέχεια επιδιώκει να μειώσει τη διαταραχή παραμένοντας παράλληλα ανταγωνιστικό. Πιο αναλυτικά, ακολουθείται ένας αλγόριθμος (βλ. σχήμα 3.12) ο οποίος ξεκινάει από ένα σημείο το οποίο είναι ήδη ανταγωνιστικό και στη συνέχεια κινείται στο όριο που χωρίζει τον ανταγωνιστικό από τον μη ανταγωνιστικό χώρο έτσι ώστε (1) να παραμένει στον ανταγωνιστικό χώρο, (2) η απόσταση από την εικόνα στόχο να μειώνεται συνεχώς. Το αρχικό adversarial example προκύπτει από μια ομοιόμορφη κατανομή δεδομένων (uniform distribution), της οποίας το εύρος τιμών καθορίζεται κάθε φορά από το είδος και τις διαστάσεις των εισόδων του κάθε προβλήματος (π.χ. για εικόνες έχουμε εύρος $[0, 255]$ για κάθε pixel). Η αποδοτικότητα του αλγορίθμου εξαρτάται από την κατανομή όπου θα προκύψει το αρχικό adversarial example και για αυτό τον λόγο είναι πολύ σημαντική η διαδικασία κατασκευής της. Για τις περισσότερες εφαρμογές σε πεδία με εικόνες, προτείνεται η παραγωγή adversarial example από την κανονική κατανομή και στη συνέχεια ανακατανομή του δείγματος στο σωστό εύρος τιμών, μέσω προβολής του δείγματος στον χώρο γύρω από το αρχικό δείγμα.



Σχήμα 3.12: *(Αριστερά) Στην ουσία η BA εκτελεί δειγματοληψία απόρριψης κατά μήκος του ορίου ανάμεσα σε ανταγωνιστικές και μη εικόνες. (Κέντρο) Σε κάθε βήμα σχεδιάζεται μια νέα τυχαία κατεύθυνση. (Δεξιά) Τα δύο μεγέθη βημάτων προσαρμόζονται δυναμικά σύμφωνα με την τοπική γεωμετρία του ορίου [57].*

Η επίθεση αυτή είναι απλή στη βάση της, δεν απαιτεί σχεδόν καθόλου hyperparameter tuning, δε βασίζεται σε substitute model, και είναι ανταγωνιστική με τις καλύτερες white-box gradient-based επιθέσεις και σε targeted και σε untargeted σενάρια, σε εργασίες εικόνων και μεγάλα σύνολα δεδομένων όπως το ImageNet.

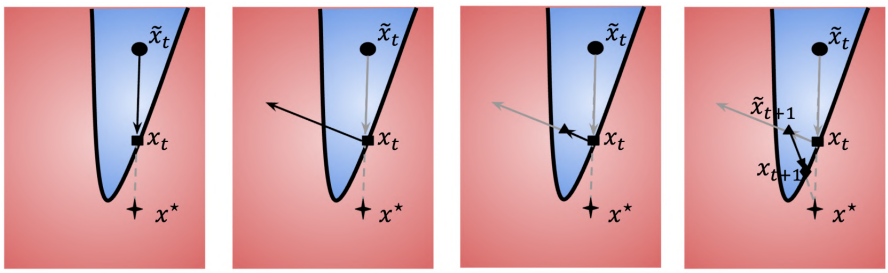
3.4.3.3 HopSkipJump Attack (HSJA)

Η **HopSkipJump Attack (HSJA)** [58], πρόκειται για μια προηγμένη έκδοση της Boundary Attack που απαιτεί μόνο προβλέψεις κλάσης (ετικετών). Είναι και αυτή μια **decision-based** επίθεση η οποία εκτελείται πιο γρήγορα (χρειάζεται πολύ λιγότερες επερωτήσεις από το μοντέλο), κάτι που την κάνει πιο εύκολα εφαρμόσιμη σε πραγματικές συνθήκες, καθώς

πραγματικές ML υπηρεσίες περιορίζουν τον μέγιστο αριθμό αιτημάτων που μπορούν να γίνουν στο μοντέλο σε ένα χρονικό διάστημα, όπως π.χ. το Google cloud vision API που έχει όριο 1800 αιτημάτων το λεπτό, είτε για να μειώσουν τον φόρτο του συστήματος είτε για να μειώσουν τον κίνδυνο επιθέσεων.

Η HSJA πρόκειται για μια οικογένεια επαναληπτικών (**iterative**) επιθέσεων που μπορούν να λειτουργήσουν σαν **targeted** και σαν **untargeted** επιθέσεις και είναι βελτιστοποιημένες για είτε ℓ_2 ή ℓ_∞ μεγέθη διαταραχών. Σε κάθε επανάληψη γίνονται τρία βήματα: (i) εκτίμηση της κατεύθυνσης της κλίσης, (ii) αναζήτηση μεγέθους βήματος μέσω γεωμετρικής προόδου, (iii) αναζήτηση ορίων μέσω δυαδικής αναζήτησης.

Μια σύντομη περιγραφή της μεθοδολογίας που ακολουθείτε σε αυτήν τη μέθοδο είναι η εξής, η οποία φαίνεται και οπτικά στο σχήμα 3.13. Αρχικά, προτείνεται ένα πλαίσιο βελτιστοποίησης (optimization framework) με το οποίο παράγονται adversarial examples για targeted και untargeted επιθέσεις και όλα τα μεγέθη διαταραχών ($l_p, p \in \{0, 2, \infty\}$). Για την προσέγγιση του ορίου απόφασης χρησιμοποιείται αλγόριθμος δυαδικής αναζήτησης. Για την εύρεση της κατεύθυνσης της παραγώγου (gradient descent direction) κατά την οποία συγκλίνει το πρόβλημα βελτιστοποίησης, αλλά και για την επιλογή του κατάλληλου βήματος προς αυτήν την κατεύθυνση, με το ελάχιστο δυνατό σφάλμα, προτείνονται δυο άλλοι ξεχωριστοί αλγόριθμοι. Τέλος, γίνεται εφαρμογή γεωμετρικής προόδου για την επιλογή ενός επιτυχημένου adversarial example.



Σχήμα 3.13: *Οπτική εξήγηση της HSJA. x^*: αρχικό δείγμα, x_t adversarial example, $\tilde{x}_t$: ενδιάμεσο δείγμα. (a) Εύρεση ορίου απόφασης μέσω δυαδικής αναζήτησης και ανανέωση $\tilde{x}_t \to x_t$. (b) Υπολογισμός κλίσης στο όριο x_t. (c) Εφαρμογή γεωμετρικής προόδου και ανανέωση $x_t \to \tilde{x}_{t+1}$. (d) Εφαρμογή δυαδικής αναζήτησης και ανανέωση $\tilde{x}_{t+1} \to x_{t+1}$ [58].*

Η επίθεση αυτή είναι μια από τις καλύτερες decision-based επιθέσεις, χρειάζεται πολύ λιγότερα ερωτήματα για να επιτύχει το ίδιο αποτέλεσμα, δεν έχει hyperparameters, και μπορεί να χρησιμοποιηθεί σαν ένα πρώτο βήμα για την αξιολόγηση νέων αμυνών.

3.5 Παραδείγματα ανταγωνιστικών επιθέσεων από τον πραγματικό κόσμο

Τα adversarial examples δεν εφαρμόζονται μόνο σε ελεγχόμενα περιβάλλοντα, αλλά έχουν εφαρμογές και στον φυσικό κόσμο, όπου δεν είναι πάντα δεδομένος ο έλεγχος της εισόδου. Για παράδειγμα, σε ένα σύστημα αναγνώρισης οδικών πινακίδων (το οποίο μπορεί να είναι μέρος ενός γενικότερου AI συστήματος αυτόνομης οδήγησης), δεν υπάρχει η δυνατότητα προσθήκης διαταραχής, αν δε γίνει κατάληψη του συστήματος τροφοδοσίας εισόδου, όπου δίνει στο νευρωνικό δίκτυο την είσοδο, από τη/τις κάμερες που τραβάνε τις πινακίδες. Εκεί

Attack	Knowledge	Specificity	Perturbation	Scope	Learning
FGSM [1]	White-Box	Targeted	ℓ_∞	Image	One Shot
L-BFGS [31]	White-Box	Targeted	ℓ_∞	Image	One Shot
BIM [36]	White-Box	Untargeted	ℓ_∞	Image	Iterative
ILCM [36]	White-Box	Targeted	ℓ_∞	Image	Iterative
PGD [47]	White-Box	Targeted	ℓ_1, ℓ_∞	Image	Iterative
JSMA [49]	White-Box	Targeted	ℓ_0	Image	Iterative
C&W [26]	White-Box	Targeted, Untargeted	$\ell_0, \ell_2, \ell_\infty$	Image	Iterative
DeepFool [53]	White-Box	Untargeted	$\ell_1, \ell_2, \ell_\infty$	Image	Iterative
Universal [54]	White-Box	Untargeted	ℓ_2, ℓ_∞	Universal	Iterative
EAD [59]	White-Box	Targeted, Untargeted	ℓ_1	Image	Iterative
ZOO [56]	Black-Box	Targeted, Untargeted	ℓ_2	Image	Iterative
BA [57]	Black-Box	Targeted, Untargeted	ℓ_2	Image	Iterative
HSJA [58]	Black-Box	Targeted, Untargeted	ℓ_2, ℓ_∞	Image	Iterative
One-Pixel [60]	Black-Box	Targeted, Untargeted	ℓ_0	Image	Iterative
UPSET [61]	Black-Box	Targeted	ℓ_2, ℓ_∞	Universal	Iterative
ANGRI [61]	Black-Box	Targeted	ℓ_2, ℓ_∞	Image	Iterative
Houdini [62]	Black-Box	Targeted	ℓ_2, ℓ_∞	Image	Iterative
AdvGAN [63]	Gray-Box, Black-Box	Targeted	ℓ_2	Image	Iterative

Πίνακας 3.2: *Συγκριτικός πίνακας των state-of-the-art ανταγωνιστικών επιθέσεων (white-box, black-box) με τα κύρια χαρακτηριστικά τους*

θα πρέπει η κατάλληλη διαταραχή να είναι μέρος του περιβάλλοντος και της εισόδου που θα καταγραφεί από τους αισθητήρες εισόδου, όπως η κάμερα [46].

Παρακάτω αναλύουμε μερικά παραδείγματα ανταγωνιστικών επιθέσεων στον πραγματικό κόσμο από την ακαδημαϊκή βιβλιογραφία αλλά και από πραγματικές επιθέσεις που έχουν συμβεί.

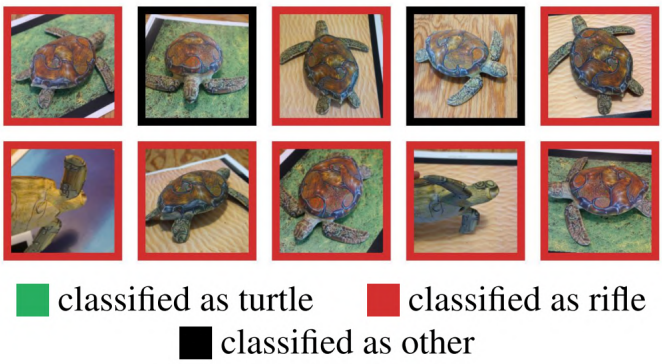
3.5.0.1 3D adversarial objects in the physical world

Ένα παράδειγμα adversarial example στον φυσικό κόσμο είναι αυτό του [64], στο οποίο εκτυπώνονται 3D adversarial φυσικά αντικείμενα, τα οποία επεκτείνουν τους αλγόριθμους παραγωγής adversarial examples από δισδιάστατες εικόνες σε αντικείμενα στον χώρο, για να προκαλέσουν τους ML ταξινομητές σε εσφαλμένη ταξινόμηση. Πρόκειται για μια από τις πρώτες επιδείξεις ύπαρξης 3D adversarial examples στον φυσικό κόσμο.

Πιο συγκεκριμένα, μερικά παραδείγματα τέτοιων εκτυπωμένων 3D adversarial examples που χρησιμοποιήθηκαν ήταν χελώνες, η οποίες αναγνωρίζονταν ως καραμπίνες αλλά και μπάλες του baseball, οι οποίες αναγνωρίζονταν ως καφές espresso. Συνολικά, χρησιμοποιήθηκαν 10 3D μοντέλα, και παράχθηκαν 200 adversarial examples για 20 διαφορετικές κλάσεις στόχους για το κάθε μοντέλο. Συνολικά για όλα τα μοντέλα, υπήρξε εσφαλμένη ταξινόμηση στην κλάση στόχο (**targeted adversarial attack**) με μέσο όρο ποσοστό 83.4%.

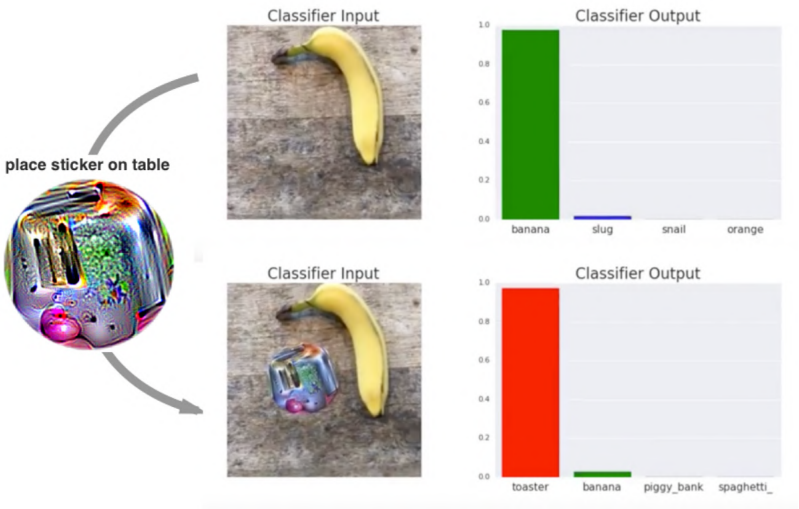
3.5.0.2 Adversarial patch: Banana as a Toaster

Ένα άλλο παράδειγμα adversarial example στον φυσικό κόσμο είναι αυτό του [65], στο οποίο χρησιμοποιούνται adversarial patches εικόνων τα οποία εκτυπώνονται υπό τη μορφή αυτοκόλλητου και μπορούν να τοποθετηθούν κοντά σε άλλα φυσικά αντικείμενα ώστε να προκαλέσουν τους ταξινομητές να τα αγνοήσουν και να διαλέξουν την επιλεγμένη κλάση στόχο



Σχήμα 3.14: *Τυχαίες πόζες μιας 3D εκτυπωμένης χελώνας που περιέχει adversarial perturbations, ταξινομώντας την ως καραμπίνα από κάθε οπτική γωνία. Στην περίπτωση χωρίς adversarial perturbations το μοντέλο ταξινομεί την χελώνα με ακρίβεια σχεδόν 100%, ενώ με την ύπαρξη τους καμία δεν ταξινομείται σωστά ως χελώνα (**πράσινο** πλαίσιο), παρά μόνο ως καραμπίνα (**κόκκινο** πλαίσιο) ή κάτι άλλο (**μαύρο** πλαίσιο) [64].*

(**targeted adversarial attack**). Αυτά τα adversarial patches συνήθως περιλαμβάνουν μεγάλα adversarial perturbations τα οποία μπορεί να έχουν μεγάλη επίπτωση σε μοντέλα τα οποία δεν είναι ανθεκτικά σε τόσο μεγάλες διαταραχές.



Σχήμα 3.15: *Παράδειγμα επίθεσης στον πραγματικό κόσμο σε ένα VGG16 ταξινομητή, χρησιμοποιώντας την τεχνική του adversarial physical patch που έχουν παραχθεί σε white-box σενάριο, σε μορφή ενός αυτοκόλλητου. Στην πρώτη περίπτωση το μοντέλο ταξινομεί την εικόνα ως μπανάνα με 97% βεβαιότητα. Στην κάτω περίπτωση όμως που έχει τοποθετηθεί το adversarial αυτοκόλλητο, ταξινομεί την εικόνα ως τοστιέρα με 99% βεβαιότητα [65].*

Πιο συγκεκριμένα, ένα τέτοια παράδειγμα αυτοκόλλητου που περιέχει adversarial patches τα οποία προέρχονται από εικόνα μιας τοστιέρας, όταν τοποθετηθεί κοντά σε άλλα αντικείμενα αναγκάζει τα μοντέλα να τα ταξινομήσουν σαν τοστιέρα και μάλιστα με μεγάλη ακρίβεια. Μάλιστα, όταν η εικόνα περιλαμβάνει λίγο περισσότερο από 10% attack perturbations που έχουν υπολογιστεί σε white-box σενάριο, έχει μεγαλύτερο από 80% Attack Success Rate (ASR), ενώ και στην περίπτωση του black-box σεναρίου χρειάζεται λίγο πάνω 50% attack perturbations για να έχει το ίδιο ποσοστό επιτυχίας η επίθεση.

3.5.0.3 Tay Poisoning

Ένα πραγματικό παράδειγμα ανταγωνιστικής επίθεσης, είναι αυτό του Twitter chatbot Tay[2], ένα chatbot προγραμματισμένο από τη Microsoft για να συμμετέχει σε συνομιλίες, να διασκεδάζει τους χρήστες και να μαθαίνει από αυτούς. Ενώ προηγούμενα chatbot, χρησιμοποιούσαν έτοιμα προγραμματισμένα σενάρια για να ανταποκρίνονται στις προτροπές (prompts) των χρηστών, το Tay χρησιμοποιούσε τεχνικές μηχανικής μάθησης που θα του επιτρέπουν να μαθαίνει από τις συζητήσεις. Όμως σε λιγότερο από 24 ώρες, μετά από συντονισμένες επιθέσεις από κακόβουλους χρήστες, χρησιμοποιούσαν υβριστική και προσβλητική γλώσσα για να μιλήσουν στο Tay, με αποτέλεσμα να αρχίσει το chatbot να δημιουργεί αντίστοιχο προσβλητικό και ρατσιστικό περιεχόμενο προς άλλους χρήστες, αναγκάζοντας τη Microsoft να το αποσύρει άμεσα, ζητώντας παράλληλα δημόσια συγνώμη.

Στην πραγματικότητα, πρόκειται για μια **poisoning επίθεση**, καθώς οι χρήστες κατάφεραν να μολύνουν τα δεδομένα εκπαίδευσης του μοντέλου, χρησιμοποιώντας τη φράση "repeat after me" ('επανάλαβε μετά από μένα'), η οποία ανάγκαζε το Tay bot να επαναλάβει οτιδήποτε γραφόταν μετά, δημιουργώντας έτσι ένα bias στο σύνολο δεδομένων, και με μεγάλο σύνολο τέτοιων δεδομένων να χρησιμοποιεί αντίστοιχα τέτοια φρασεολογία[3].

3.5.0.4 VirusTotal Poisoning

Ένα άλλο πραγματικό παράδειγμα **adversarial poisoning επίθεσης**, είναι αυτό το οποίο καταγράφηκε από την ομάδα ασφαλείας McAfee Advanced Threat Research[4], όπου παρατηρήθηκε μια αύξηση στις αναφορές μια συγκεκριμένης οικογένειας ransomware η οποία ήταν ασυνήθιστη. Η έρευνα αποκάλυψε ότι πολλά δείγματα της συγκεκριμένης οικογένειας ransomware είχαν υποβληθεί σε πολύ σύντομο χρονικό διάστημα στην πλατφόρμα του VirusTotal, η οποία πρόκειται για μια δημοφιλής πλατφόρμα κοινής χρήσης ιών και malware. Οι ερευνητές διαπίστωσαν ότι όλα τα δείγματα ήταν ισοδύναμα ως προς την ομοιότητα συμβολοσειράς (strings) και κώδικα σε ποσοστό 98% με 74%. Ανακάλυψαν ότι ένα εργαλείο που ονομάζεται metame[5] χρησιμοποιήθηκε για τη μετατροπή του αρχικού αρχείου σε μεταλλαγμένες παραλλαγές, οι οποίες μοιάζουν μεταξύ χωρίς απαραίτητα να είναι εκτελέσιμα, αλλά εξακολουθούν να ταξινομούνται ως μέρος της ίδιας οικογένειας ransomware.

Στην ουσία, τα διάφορα antivirus άρχισαν να ταξινομούν αρχεία τα οποία δεν ήταν καν εκτελέσιμα ως ransomware της συγκεκριμένης οικογένειας, αφού τα μεταλλαγμένα δείγματα είχαν δηλητηριάσει τα σύνολα δεδομένων των ML μοντέλων ταξινόμησης malware, τα οποία στηρίζονται στη βάση ιών του VirusTotal.

3.6 Άμυνες για ανταγωνιστικές επιθέσεις

Όπως έχει αναφερθεί ήδη, τα ML μοντέλα είναι ευάλωτα στις ανταγωνιστικές επιθέσεις, και ανάλογα με τον τρόπο χρήσης και πεδίο εφαρμογής του μοντέλου, αυτή η ευαλωτότητα μπορεί να έχει καταστροφικές συνέπειες. Αυτή όμως μπορεί να αντιμετωπιστεί χρησιμοποιώντας διάφορες τεχνικές αντιμετώπισης που έχουν αναπτυχθεί παράλληλα με τις επιθέσεις. Μια τέτοια τεχνική που έχουμε αναφερθεί είναι και το adversarial training, που στοχεύει

[2] https://en.wikipedia.org/wiki/Tay_(chatbot)
[3] https://atlas.mitre.org/studies/AML.CS0002
[4] https://atlas.mitre.org/studies/AML.CS0009
[5] https://github.com/a0rtega/metame

στην ισχυροποίηση του μοντέλου σε adversarial examples, εκπαιδεύοντας το με αυτά, αλλά υπάρχουν και άλλες άμυνες που δε βασίζονται μόνο στην εκπαίδευση του μοντέλου, αλλά στην τροποποίηση των δεδομένων ή στην ανίχνευση adversarial examples. Επίσης, είναι σημαντικό να αναφερθεί ξανά ότι ο σχεδιασμός κατάλληλων αντιμέτρων απέναντι σε ανταγωνιστικές εισόδους είναι ένα πρόβλημα βελτιστοποίησης που είναι μη γραμμικό και μη κυρτό και δεν υπάρχει μια στρατηγική άμυνα που να είναι αποτελεσματική για όλες τις επιθέσεις. Για αυτό πολλές άμυνες έχουν τους περιορισμούς τους και μπορούν να παρέχουν ασφάλεια έναντι επιθέσεων που ανήκουν σε ένα συγκεκριμένο μοντέλο απειλής [5].

Σε αυτήν την ενότητα γίνεται μια προσπάθεια συγκέντρωσης τον ποιο δημοφιλών τεχνικών άμυνας και αντιμετώπισης των ανταγωνιστικών επιθέσεων καθώς και μια προσπάθεια κατηγοριοποίησης τους.

3.6.1 Κατηγοριοποίηση Αμυνών

3.6.1.1 Ανά Στόχο Αντιμετώπισης

Ανάλογα με το κύριο στόχο αντιμετώπισης (**Objective**) των επιθέσεων, μπορούμε να κατηγοριοποιήσουμε τις άμυνες σε:

1. **Reactive Defenses**: Οι αντιδραστικές άμυνες (**reactive**) είναι άμυνες που στοχεύουν στο να αντιμετωπίσουν παρελθοντικές επιθέσεις. Σχεδιάζονται δηλαδή, με βάση προηγούμενες επιθέσεις στο μοντέλο καθώς και προσαρμόζονται ανάλογα με τις επιθέσεις και τη συμπεριφορά των επιτιθεμένων [27]. Συνήθως πρόκειται για τεχνικές ανίχνευσης των adversarial examples με χρήση άλλων νευρωνικών [7].
2. **Proactive Defenses**: Οι προληπτικές άμυνες (**proactive**) είναι άμυνες που στοχεύουν στο να αντιμετωπίσουν μελλοντικές επιθέσεις. Σχεδιάζονται δηλαδή, μοντελοποιώντας τον adversary και τις επιθέσεις του, ώστε να αξιολογηθεί τον αντίκτυπό τους και να αναπτυχθούν τα κατάλληλα αντίμετρα (countermeasures) [27]. Συνήθως πρόκειται για τεχνικές ενίσχυσης της ευρωστίας του νευρωνικού από adversarial examples [7]. Πιο αναλυτικά όμως αυτές τις άμυνες μπορούμε να τις κατηγοριοποιήσουμε σύμφωνα με τα εξής:
   (α′) **Security-by-Design Defenses**: Πρόκειται για βελτίωση της ασφάλειας ενός συστήματος, σχεδιάζοντας το εξ αρχής ώστε να είναι ασφαλές, κυρίως από **white-box επιθέσεις**.
   (β′) **Security-by-Obscurity Defenses**: Πρόκειται για βελτίωση της ασφάλειας ενός συστήματος, με απόκρυψη πληροφοριών από τους επιτιθέμενους ώστε να είναι ασφαλές, κυρίως από **gray-box/black-box επιθέσεις**.

Σχετικά με τις **reactive άμυνες**, η βελτίωση της ασφάλειας του συστήματος γίνεται ώστε το σύστημα να μπορέσει να προφυλαχτεί από γνωστές επιθέσεις. Αυτό γίνεται κυρίως με τεχνικές ανίχνευσης, όπως (i) *ανίχνευση των adversarial examples* με χρήση άλλων βοηθητικών (auxiliary) νευρωνικών που εκπαιδεύονται να αναγνωρίζουν τα ανταγωνιστικά δείγματα [66], [67], είτε ανιχνεύοντας μη φυσιολογικές εισόδους που βρίσκονται απομακρυσμένα από την κατανομή της εκπαίδευσης (*Out-of-Distribution Detection*) [68], είτε ανιχνεύοντας τις ανταγωνιστικές εισόδους με βάση των εγγενών ιδιοτήτων των DNN ταξινομητών [69], (ii) με μείωση των διαστάσεων της εισόδου για την ανίχνευση adversarial perturbations με τεχνικές όπως το *Feature Squeezing* [70] το οποίο αναλύεται στην ενότητα 3.6.2.4, (iii) και με

Reactive Defenses
Detection of Adversarial Examples
Out-of-Distribution Detection
Anomaly Detection
Feature Squeezing

Proactive Defenses	
Security-by-Design	***Security-by-Obscurity***
Adversarial Training	Randomization
Defense Distillation	Information Hiding
Classifier Ensemble	Random Noise Injection
	Detection of Probing Attacks

Πίνακας 3.3: *Κατηγοριοποίηση των τεχνικών άμυνας ανά objective [27].*

τεχνικές ανίχνευσης ανωμαλιών (*Anomaly Detection*) σε τριτογενείς παράγοντες όπως την καθυστέρησης πρόβλεψης του μοντέλου ή της χρήσης των πόρων του συστήματος (CPU, GPU, Memory) οι οποίες βασίζονται στο γεγονός ότι η δημιουργία και επεξεργασία των adversarial examples είναι μια χρονοβόρα διαδικασία που μπορεί να προκαλέσει απότομη αύξηση χρήσης των πόρων του συστήματος [7]. Το πρόβλημα με αυτές τις τεχνικές είναι ότι οι επιτιθέμενοι μπορούν πολλές φορές να αποφύγουν τα μοντέλα ανίχνευσης κατασκευάζοντας νέες συναρτήσεις απώλειας, ειδικά σε σενάριο white-box [71]. Επίσης, σε σχετικά περίπλοκα σύνολα δεδομένων, τα adversarial examples είναι πολύ πιο δύσκολο να ξεχωρίσουν από τις αρχικές εισόδους [27].

Σχετικά με τις **proactive άμυνες by-design**, η βελτίωση της ασφάλειας και της ευρωστίας του συστήματος γίνεται ώστε το σύστημα να μπορέσει να αντιδράσει σε μελλοντικές επιθέσεις. Αυτό γίνεται κυρίως με τεχνικές εκπαίδευσης οι οποίες ενισχύουν την ευρωστία των μοντέλων ενάντια στα adversarial examples, όπως (i) adversarial training [1] στην οποία έχουμε αναφερθεί εκτενώς στην ενότητα 3.3, (ii) και με την τεχνική του defense distillation [51] το οποίο αναλύεται στην ενότητα 3.6.2.3. Επίσης, μια άλλη τακτική είναι (iii) η ένωση δύο η περισσότερων ταξινομητών για τη δημιουργία ενός συνόλου (*Classifier Ensemble*), το οποίο μπορεί να προσφέρει προστασία από evasion επιθέσεις λόγο της υπόθεσης ότι κάθε μοντέλο αντισταθμίζει αμοιβαία τις αδυναμίες που μπορεί να έχει ένα άλλο μοντέλο κατά την ταξινόμηση μιας δεδομένης εισόδου [5]. Το πρόβλημα με αυτές τις τεχνικές είναι ότι πρόκειται κυρίως για ευριστικές λύσεις, οι οποίες δεν προσφέρουν εγγυήσεις για την ευρωστία, αλλά όταν πρόκειται για upper-bound τεχνικές που δίνουν εγγυήσεις, συνήθως δεν είναι επαρκείς ή είναι αρκετά ακριβές υπολογιστικά.

Σχετικά με τις **proactive άμυνες by-obscurity**, η απόκρυψης πληροφοριών για το σύστημα γίνεται με στόχο την αποτροπή των επιτιθεμένων από το να εξάγουν πληροφορίες για το μοντέλο και μέσω κατάλληλων επερωτήσεων ή μηχανισμών ανίχνευσης (probe attacks) να δημιουργήσουν υποκατάστατους μαθητές (surrogate learners) οι οποίοι προσομοιάζουν κατά προσέγγιση το αρχικό μοντέλο, και μπορούν να τους χρησιμοποιήσουν για να δημιουργήσουν τις κατάλληλες ανταγωνιστικές επιθέσεις. Αυτό μπορεί να αποφευχθεί με τεχνικές όπως (i) *η τυχαία συλλογή δεδομένων εκπαίδευσης* (συλλογή δεδομένων σε διαφορετικούς χρόνους και τοποθεσίες), (ii) η χρήση *μοντέλων που γίνονται δύσκολα reverse-engineered*, (iii) η *απαγόρευση πρόσβασης στα δεδομένα εκπαίδευσης του μοντέλου*, (iv) και η *προσθήκη τυχαίου*

θορύβου στην έξοδο του ταξινομητή [27]. Τέλος, για την αποτροπή ενός επιτιθέμενου από το να αποκτήσει πρόσβαση σε πληροφορίες ή εξόδους του μοντέλου μπορεί να χρησιμοποιηθούν (v) *τεχνικές ανίχνευσης των probe attacks*. Το πρόβλημα όμως με αυτές τις τεχνικές απόκρυψης είναι ότι πολλές φορές μπορεί να μην είναι χρήσιμες. Για παράδειγμα, τεχνικές που χρησιμοποιούν gradient masking οι οποίες αποκρύπτουν την κατεύθυνσης της κλίσης, η οποία χρησιμοποιείται για τη δημιουργία adversarial examples, έχει αποδειχθεί ότι μπορούν εύκολα να παρακαμφθούν με surrogate learners που πρόκειται για μια συχνή black-box τακτική, κάνοντας ουσιαστικά την άμυνα να μην έχει καμία ισχύ [72].

3.6.1.2 Ανά Προσέγγιση Αντιμετώπισης

Ανάλογα με την προσέγγιση (**Approach**) που ακολουθείτε για την αντιμετώπιση επιθέσεων, μπορούμε να κατηγοριοποιήσουμε τις άμυνες σε:

1. **Model hardening**: Οι τεχνικές θωράκισης του μοντέλου (**model hardening**) πρόκειται για τεχνικές που καταλήγουν σε έναν νέο μοντέλο με καλύτερες ιδιότητες ευρωστίας από τον αρχικό σε σχέση με ορισμένες δεδομένες μετρήσεις [42].
2. **Data preprocessing**: Οι τεχνικές προεπεξεργασίας δεδομένων (**data preprocessing**) πρόκειται για τεχνικές που επιτυγχάνουν υψηλότερη ευρωστία χρησιμοποιώντας μετασχηματισμούς των εισόδων και των ετικετών του ταξινομητή, αντίστοιχα, κατά το χρόνο δοκιμής ή/και εκπαίδευσης ώστε να φιλτράρουν τα δεδομένα από τον θόρυβο που τυχόν τους έχει προστεθεί ή και να τα απομακρύνουν αν δεν είναι δυνατόν να διορθωθούν [42].
3. **Runtime detection**: Οι τεχνικές ανίχνευσης κατά τον χρόνο εκτέλεσης (**runtime detection**) πρόκειται για τεχνικές ανίχνευσης adversarial examples κατά τη λειτουργία του ML συστήματος, επεκτείνοντας το αρχικό μοντέλο με έναν ανιχνευτή ο οποίος ελέγχει εάν μια δεδομένη είσοδος είναι adversarial ή όχι [42].

Σχετικά με τις **model hardening** τεχνικές, μπορούμε να κατατάξουμε εδώ όλες τις τεχνικές οι οποίες βελτιώνουν το ίδιο το μοντέλο που αναφέραμε στις **proactive άμυνες by-design**. Αυτές μπορούμε να τις διακρίνουμε σε (i) *τεχνικές ενίσχυσης των δεδομένων εκπαίδευσης* μέσω adversarial examples, όπως το *adversarial training*, (ii) *τεχνικές regularization κατά την διάρκεια της εκπαίδευσης* [73], και (iii) *τεχνικές που τροποποιούν τα στοιχεία της αρχιτεκτονικής των ταξινομητών* [74]. Συνήθως οι τεχνικές αυτές προτείνονται και έχουν καλύτερα αποτελέσματα στις **white-box επιθέσεις**.

Σχετικά με τις **data preprocessing** τεχνικές, μπορούμε να κατατάξουμε εδώ όλες τις τεχνικές προεπεξεργασίας εισόδων που αναφέραμε στις **reactive άμυνες**. Αυτές μπορούμε να τις διακρίνουμε σε (i) *τεχνικές που χρησιμοποιούν τυχαίους μετασχηματισμό*, ειδικά για εικόνες όπου μπορεί να γίνουν μετασχηματισμοί όπως περικοπές, αλλαγή ανάλυσης, μείωση του bit-depth, JPEG συμπίεση ή ελαχιστοποίηση της συνολικής διακύμανσης [75], (ii) *τεχνικές μείωσης των διαστάσεων της εισόδου* όπως το *Feature Squeezing* [70], (iii) τεχνικές οι οποίες χρησιμοποιούν ένα άλλο μοντέλο το οποίο εκπαιδεύεται για να μεταφέρει τα adversarial examples πιο κοντά στην πολλαπλότητα (manifold) των κανονικών δειγμάτων και να τα 'καθαρίζει' από τα adversarial perturbations, είτε με χρήση auto-encoders όπως *MagNet* [76] το οποίο αναλύεται στην ενότητα 3.6.2.8, είτε με χρήση GANs όπως το *Defense-GAN* [77] και *APE-GAN* [50] τα οποίο αναλύονται στην ενότητα 3.6.2.9. Συνήθως οι τεχνικές αυτές

προτείνονται και έχουν καλύτερα αποτελέσματα στις **black-box και gray-box επιθέσεις**.

Σχετικά με τις **runtime detection** τεχνικές, μπορούμε να κατατάξουμε εδώ όλες τις τεχνικές ανίχνευσης adversarial example που αναφέραμε στις **reactive άμυνες**. Συνήθως οι τεχνικές αυτές προτείνονται και έχουν καλύτερα αποτελέσματα στις **black-box και gray-box επιθέσεις**.

Οι τεχνικές θωράκισης συνήθως παρέχουν μεγαλύτερη ασφάλεια από τις άλλες μεθόδους, αλλά απαιτούν εκπαίδευση ή και αλλαγή του μοντέλου. Οι μέθοδοι προεπεξεργασίας και ανίχνευσης δεν απαιτούν εκπαίδευση και μπορούν να εφαρμοστούν σε πολλαπλά μοντέλα, αλλά μπορεί να είναι πιο εύκολο να δειχθούν ανεπαρκής για προστασία από adversarial examples.

3.6.1.3 Ανά Χρόνο Εφαρμογής

Ανάλογα με την φάση εφαρμογής (**Application Phase**) της κάθε άμυνας, μπορούμε να κατηγοριοποιήσουμε τις άμυνες σε:

1. **Άμυνες για training-time attacks**: Πρόκειται για τεχνικές που προσπαθούν να αντιμετωπίσουν poisoning επιθέσεις, δηλαδή που συμβαίνουν κατά την εκπαίδευση (**training**) του μοντέλου [4].
2. **Άμυνες για inference-time attacks**: Πρόκειται για τεχνικές που προσπαθούν να αντιμετωπίσουν evasion επιθέσεις, δηλαδή που συμβαίνουν κατά το τρέξιμο (**inference**) του μοντέλου [4].

Σχετικά με τις άμυνες για **training-time attacks**, οι περισσότερες τεχνικές βασίζονται στο γεγονός ότι τα poisoning δείγματα είναι συνήθως εκτός της αναμενόμενης κατανομής εισόδων (outliers). Οι τεχνικές που χρησιμοποιούνται για την αντιμετώπιση αυτών των δειγμάτων, μπορεί να είναι (i) *τεχνικές εξυγίανσης δεδομένων (data sanitization)* για να τα αντιμετωπίσουν τους outliers (δηλαδή, ανίχνευση και αφαίρεση επίθεσης), (ii) *τεχνικές randomization κατά τη διάρκεια της εκπαίδευσης* όπως τυχαία προσθήκη θορύβου, για να γίνει το μοντέλο λιγότερο ευαίσθητο σε μικρές διαταραχές [27], και (iii) *τεχνικές εκπαίδευσης του μοντέλου για την ενίσχυση του απέναντι σε adversarial examples*, όπως το *adversarial training*. Επίσης, μία άλλη σημαντική τακτική είναι (iv) *η επαλήθευση των πηγών των δεδομένων και των μοντέλων*, τα οποία πολλές φορές κυκλοφορούν ελεύθερα στο διαδίκτυο και είναι πιθανόν να περιέχουν δηλητηριασμένα δείγματα ή να είναι έδαφος για backdoor επιθέσεις.

Σχετικά με τις άμυνες για **inference-time attacks**, οι περισσότερες τεχνικές βασίζονται στην εκπαίδευση ή στην ανίχνευση των adversarial δειγμάτων. Οι τεχνικές που χρησιμοποιούνται για την αντιμετώπιση αυτών των δειγμάτων, μπορεί να είναι (i) *ευριστικές τεχνικές εκπαίδευσης του μοντέλου για την ενίσχυση του απέναντι σε adversarial examples* όπως το *adversarial training*, (ii) *τεχνικές εκπαίδευσης με εγγυήσεις για τα άνω όρια* (*Certified Defenses*), (iii) *τεχνικές ανίχνευσης adversarial example* και (iv) τα *Classifier Ensembles*, δηλαδή η χρήση πολλαπλών μοντέλων και συνδυασμός των εξόδων τους που αναφερθήκαμε στις **proactive άμυνες by-design** [27].

3.6.2 Τεχνικές Αμυνών

Σε αυτήν την ενότητα παρουσιάζουμε πιο αναλυτικά ενδεικτικά κάποιες από τις πιο δημοφιλείς ανταγωνιστικές άμυνες (**adversarial defenses**) ή κατηγορίες τεχνικών άμυνας από τη βιβλιογραφία. Κάποιες από αυτές τις τεχνικές δουλεύουν καλύτερα σε διαφορετικές συνθήκες και μοντέλα, κάποιες έχουν μεγαλύτερο θεωρητικό υπόβαθρο, ενώ κάποιες άλλες είναι πιο

εμπειρικές, όμως όλες έχουν του δικούς τους περιορισμούς. Στον πίνακα 3.4 μπορούμε να βρούμε συγκεντρωτικά όσες άμυνες αναλύουμε εδώ και στην προηγούμενη ταξινόμηση που κάναμε αλλά και άλλες state-of-the-art από τη βιβλιογραφία.

3.6.2.1 Adversarial Training

Το **adversarial training** [1] που παρουσιάσαμε στην ενότητα 3.3, πρόκειται για μια από τις πρώτες προτεινόμενες τεχνικές για την αντιμετώπιση των adversarial examples, η οποία στην ουσία χρησιμοποιεί τα παραγόμενα adversarial examples από κάποια μέθοδο (π.χ. FGSM) μαζί με καθαρά για να εκπαιδεύσει ένα νέο μοντέλο (model hardening) το οποίο θα είναι ανθεκτικό απέναντι σε adversarial examples.

Από όλες τις άμυνες που έχουν προταθεί στη βιβλιογραφία, το adversarial training είναι μια από τις πιο αξιόπιστες και πιο πολύ αξιολογημένες άμυνες [78], καθώς το πλεονέκτημα της είναι ότι δεν πρόκειται για κάποια τεχνική η οποία μπορεί να έχει κενά ασφαλείας, αλλά στην ουσία επεκτείνει την εκπαίδευση του μοντέλου, μαθαίνοντας το να μπορεί να λειτουργεί και σε δείγματα με διαταραχές, αλλάζοντας τα όρια απόφασης του μοντέλου. Αυτό όμως είναι και ένα από τα βασικά μειονεκτήματα της, καθώς είναι απαραίτητη η επανεκπαίδευσης του μοντέλου και για να είναι ανθεκτικό απέναντι σε πολλά και διαφορετικά είδη διαταραχών, θα πρέπει να εκπαιδευτεί με δείγματα από το κάθε είδος, αυξάνοντας τον χρόνο και την πολυπλοκότητα της εκπαίδευσης.

Στο πλαίσιο της εκπαίδευσης, υπάρχουν και άλλοι robust training μέθοδοι που εκπαιδεύουν το μοντέλο για να είναι ανθεκτικό απέναντι σε adversarial example, όπως οι **certified training** μέθοδοι, η οποίες χρησιμοποιούν μαθηματικές μεθόδους που αποδεικνύουν την ευρωστία του εκπαιδευμένου μοντέλου απέναντι σε ένα άνω όριο διαταραχών [79]. Πρόκειται για πιο ισχυρές εγγυήσεις ασφαλείας σε σχέση με το απλό adversarial training, αλλά συνήθως έχουν χαμηλότερες επιδόσεις σε κοινές επιθέσεις και αυξάνουν πολύ περισσότερο την πολυπλοκότητα της εκπαίδευσης.

Περισσότερη ανάλυση για το είδος και το μέγεθος της ευρωστίας που προσφέρει η κάθε ανταγωνιστική μέθοδος εκπαίδευσης καθώς και για τα πλεονεκτήματα και μειονεκτήματα τους, γίνεται στο κεφάλαιο 4.

3.6.2.2 Network Regularization

Το Network Regularization πρόκειται για ένα είδος αμυντικών τεχνικών, οι οποίες συνήθως εκπαιδεύουν ένα μοντέλο ενάντια (model hardening) στα adversarial examples προσθέτοντας ένα επιπλέον επίπεδο κανονικοποίησης (regularization) στην αρχική objective συνάρτηση [80].

Η **DeepDefense** [80], πρόκειται για μια τέτοια μέθοδο, η οποία σε αντίθεση με άλλες παρόμοιες τεχνικές που κάνουν προσεγγίσεις και βελτιστοποιούν με χαλαρά όρια, ενσωματώνει έναν regularizer βασισμένο σε διαταραχές μέσα στη objective συνάρτηση ενός ταξινομητή. Αυτό δίνει την ικανότητα στα νευρωνικά δίκτυα να μάθουν άμεσα από τις επιθέσεις, τιμωρώντας τα adversarial perturbations και ενθαρρύνοντας με σχετικά μεγάλες τιμές για τα σωστά ταξινομημένα δείγματα και μικρές για τα εσφαλμένα. Ο regularizer εφόσον ενσωματώνεται μέσα στο νευρωνικό δίκτυο, μπορεί να εκπαιδευτεί συνολικά, με αποδοτικό τρόπο σαν ένα αναδρομικό δίκτυο (RNN).

Μία άλλη μέθοδος χρησιμοποιεί **denoising autoencoders (DAEs)** [81] για προεπεξερ-

γασία των δειγμάτων, βοηθώντας στο να αφαιρεθεί σημαντικό μέγεθος του ανταγωνιστικού θορύβου (adversarial noise). Με χρήση αυτών των autoencoders δημιουργείται στην ουσία ένα Deep Contractive Network (DCN) το οποίο επιβάλει μια ποινή σε όλο το στρώμα ενός feed-forward νευρωνικού δικτύου, ελαχιστοποιώντας την διακύμανση των εξόδων σε σχέση με τις διαταραχές στην είσοδο, λειτουργώντας σαν έναν regularizer layer. Για αυτήν την ελαχιστοποίηση χρησιμοποιούνται οι ℓ_2 νόρμες των πινάκων Τζακόμπι σε κάθε στρώμα.

Τέλος, τα **Parseval Networks** [82], είναι μιας μορφής βαθιά νευρωνικά δίκτυα στα οποία υπάρχει μια μέθοδος regularization σε ένα στρώμα του δικτύου, η οποία είναι υπεύθυνη για τη συνολική μείωση της ευαισθησίας του μοντέλου σε μικρές διαταραχές, ελέγχοντας προσεκτικά τη συνολική Lipschitz σταθερά του μοντέλου. Η μέθοδος αυτή επιβάλει ένα περιορισμό στους πίνακες βαρών των γραμμικών και συνελικτικών στρωμάτων του δικτύου ώστε να είναι, έτσι ώστε ο μεγαλύτερος μοναδιαίος αριθμός κάθε πίνακα βατών να είναι μικρότερος από 1. Αυτή η μέθοδος βελτιώνει συνολικά τη γενίκευση του μοντέλου και το κάνει πιο εύρωστο απέναντι σε adversarial perturbations, αποδεδειγμένα εμπειρικά και θεωρητικά.

3.6.2.3 Defense Distillation

Η **Defense Distillation** [51], είναι μια **proactive** άμυνα η οποία βασίζεται στην τεχνική του distillation των νευρωνικών δικτύων [83], ένας μηχανικός που έχει σχεδιαστεί για να συμπιέζει μεγάλα μοντέλα σε μικρότερα διατηρώντας παράλληλα την ακρίβεια της πρόβλεψης. Δηλαδή, το μεγάλο μοντέλο επισημαίνει δεδομένα με πιθανότητες κλάσης, οι οποίες στη συνέχεια χρησιμοποιούνται για την εκπαίδευση του μικρού μοντέλου.

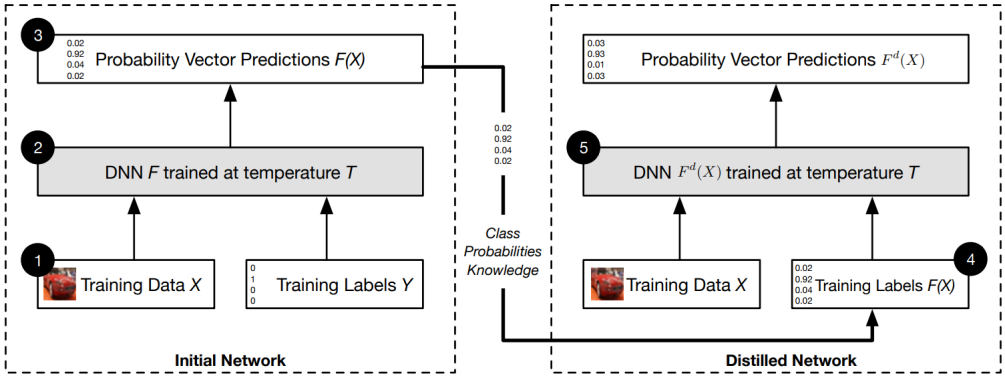
Σχετικά με την τεχνική του distillation, τα νευρωνικά δίκτυα τυπικά παράγουν πιθανότητες κλάσεων (class probabilities) χρησιμοποιώντας ένα softmax επίπεδο το οποίο, παίρνει σαν είσοδο ένα διάνυσμα εξόδου Z(x) που παράχθηκε από το τελευταίο κρυμμένο επίπεδο και ονομάζεται logits (z_i) και το μετατρέπει σε ένα διάνυσμα πιθανοτήτων $F(X)$ που αποτελεί την έξοδο του δικτύου και δηλώνει την πιθανότητα της κλάσης για κάθε είσοδο X, και κλάσεις με δείκτες $i \in 0..N-1$ [39]:

$$F(X) = \left[\frac{e^{z_i(X)/T}}{\sum_{l=0}^{N-1} e^{z_l(X/T)}} \right]_{i \in 0..N-1}$$

όπου T είναι μια παράμετρος που ονομάζεται Temperature, διανέμεται σε όλο το softmax επίπεδο και συνήθως έχει τιμή 1. Όσο μεγαλύτερη είναι η τιμή του T, παράγεται μια πιο δηλαδή τόσο πιο διακριτό γίνεται το μοίρασμα πιθανοτήτων (μόνο η σωστή κλάση θα έχει πιθανότητα 1 και οι υπόλοιπες 0).

Άρα, η defense distillation, επιτυγχάνεται με τη χρήση διαφορετικής τιμής της παραμέτρου T για το νευρωνικό δίκτυο που θέλουμε να γίνει εύρωστο με μόνο περιορισμό η τιμή της στο αρχικό δίκτυο να είναι μεγαλύτερη του ενός. Πιο συγκεκριμένα (βλ. σχήμα 3.16), ένας ταξινομητής εκπαιδεύεται σε δύο γύρος χρησιμοποιώντας μια παραλλαγή του distillation, το οποίο έχει ως αποτέλεσμα την εκπαίδευση ενός πιο ομαλού smooth δικτύου και μείωση του πλάτους των κλίσεων γύρω από τα σημεία εισόδου, καθιστώντας δύσκολο για τους επιτιθέμενους να δημιουργήσουν adversarial examples.

Η τεχνική αυτή παρουσιάζει πολύ καλά αποτελέσματα ενάντια σε αλγορίθμους που έχει δοκιμαστεί, μειώνοντας την πιθανότητα επιτυχίας τους από 95% σε 0,5%, μπορεί να εφαρμοστεί



Σχήμα 3.16: *Επισκόπηση του defense distillation μηχανισμού που βασίζεται σε μεταφορά γνώσης που περιέχεται στα διανύσματα πιθανοτήτων μέσω distillation: Αρχικά εκπαιδεύεται ένα αρχικό δίκτυο F σε δεδομένα X με softmax θερμοκρασία T. Στη συνέχεια, χρησιμοποιείται το διάνυσμα πιθανότητας $F(X)$, το οποίο περιλαμβάνει πρόσθετες γνώσεις σχετικά με τις κλάσεις σε σχέση με τις ετικέτες, που προβλέπεται από το δίκτυο F για την εκπαίδευση ενός distilled δικτύου F^d σε θερμοκρασία T στα ίδια δεδομένα X [56].*

σε οποιοδήποτε feed-forward νευρωνικό δίκτυο και απαιτεί μόνο ένα μόνο βήμα επανεκπαίδευσης (**model hardening** τεχνική), κάνοντας τη μία από τις μοναδικές άμυνες που παρέχουν ισχυρές εγγυήσεις ασφάλειας έναντι adversarial examples.

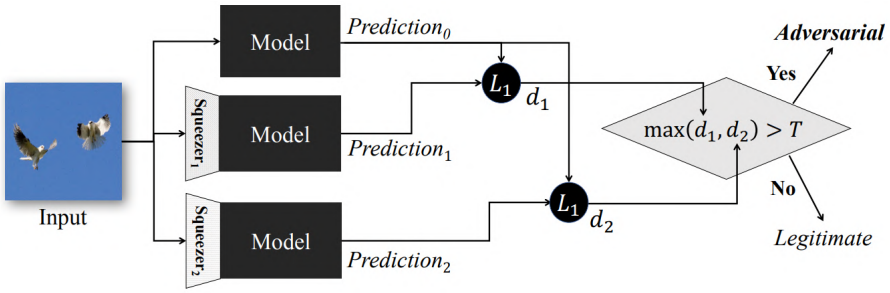
Όμως αργότερα, αποδείχθηκε ότι *αποτυγχάνει να αντισταθεί από ισχυρές επιθέσεις*, όπως η C&W επίθεση, αλλά και να *προστατεύσει επαρκώς από black-box επιθέσεις* με adversarial examples υψηλής εμπιστοσύνης που έχουν μεταφερθεί από άλλα ευάλωτα νευρωνικά δίκτυα [26].

3.6.2.4 Feature Squeezing

Το Feature Squeezing [70], είναι μια **reactive** αμυντική τεχνική, προεπεξεργασίας δεδομένων (**data preprocessing**), η οποία μπορεί να ενισχύσει τα DNN μοντέλα, ανιχνεύοντας adversarial examples. Η τεχνική αυτή μειώνει τον διαθέσιμο χώρο αναζήτησης σε έναν επιτιθέμενο, συγχωνεύοντας δείγματα που αντιστοιχούν σε πολλά διαφορετικά διανύσματα χαρακτηριστικών στον αρχικό χώρο σε ένα ενιαίο δείγμα. Συγκρίνοντας την πρόβλεψη ενός DNN μοντέλου στο αρχικό δείγμα και στο συμπιεσμένο (squeezed) δείγμα, η feature squeezing τεχνική, ανιχνεύει τα adversarial examples με υψηλή ακρίβεια και ελάχιστα false-positives (βλ. σχήμα 3.17).

Η ιδέα προέκυψε από την παρατήρηση ότι οι χώροι χαρακτηριστικών εισόδου είναι συνήθως άσκοπα μεγάλοι και παρέχουν ένα τεράστιο χώρο σε επιτιθέμενους να κατασκευάσουν adversarial perturbations και ως εκ τούτου προτάθηκε η συμπίεση χαρακτηριστικών, ώστε να γίνεται εστίαση μόνο στο ουσιαστικό μέρος των δεδομένων εισόδου. Αυτή η μέθοδος συμπίεσης χαρακτηριστικών (μείωση των διαστάσεων της εισόδου) μπορεί να επιτευχθεί με δύο στρατηγικές: (i) με **μείωση του color bit depth** σε κάθε pixel, και (ii) με εξομάλυνση του χώρου (**spatial smoothing**). Για την πρώτη στρατηγική, ο στόχος είναι η εξάλειψη μικρών διαταραχών καλύπτοντας διάφορα pixel και η συμπίεση που χρησιμοποιείται στην περίπτωση των έγχρωμων εικόνων είναι από 24-bit color depth (3×8-bit για κάθε κανάλι χρώματος: red, green, blue) σε 8-bit (1 bit ανά κανάλι), κάνοντας έτσι τον ανταγωνιστικό θόρυβο πιο

αντιληπτό καθώς μειώνεται το βάθος του bit. Για τη δεύτερη στρατηγική, ο στόχος είναι η εξάλειψη μεγάλων διαταραχών καλύπτοντας ορισμένα pixel, καθώς η μέθοδος αυτή μετακινεί ένα φίλτρο σε μια αρχική εικόνα και τροποποιεί την τιμή του κεντρικού pixel στη διάμεσο των τιμών των pixel στο φίλτρο.



Σχήμα 3.17: *Επισκόπηση του Feature Squeezing μηχανισμού για ανίχνευση adversarial examples. Το μοντέλο αξιολογείται τόσο στην αρχική όσο και στην feature squeezed είσοδο. Αν η διαφορά μεταξύ της πρόβλεψης του μοντέλου σε μια συμπιεσμένη είσοδο και στην αρχική υπερβαίνει ένα συγκεκριμένο threshold, η είσοδος προσδιορίζεται ως adversarial [70].*

Οι δύο αυτές στρατηγικές, είναι αρκετά φθηνές και συμπληρωματικές με άλλες άμυνες, οπότε και μπορεί η feature squeezing τεχνική να συνδυαστεί με άλλες άμυνες για να επιτευχθούν ακόμα καλύτερα ποσοστά ανίχνευσης adversarial examples. Όμως και αυτή η τεχνική μετά από μεταγενέστερη έρευνα *φάνηκε να είναι λιγότερο αποτελεσματική από την αρχική αξιολόγηση* [84].

3.6.2.5 Label Smoothing

Το **Label Smoothing** [85], πρόκειται για μία **proactive** πιο απλοποιημένη εναλλακτική του **defense distillation** η οποία ενισχύει το robusstness ενάντια σε adversarial examples που έχουν δημιουργηθεί με τη FGSM μέθοδο. Η μέθοδος αυτή αντικαθιστά τα hard labels μιας κλάσης (ένα διάνυσμα όπου το μόνο μη μηδενικό στοιχείο είναι ο σωστός δείκτης κλάσης), με soft labels (σε κάθε κλάση εκχωρείται μία τιμή κοντά στο $1/N$ για ένα πρόβλημα N-κλάσεων).

Πιο συγκεκριμένα, σε ένα πρόβλημα ταξινόμησης k κλάσεων, των οποίων το διάνυσμα που αντιστοιχεί στο σωστό label έχει τιμές 1 στη σωστή κλάση και 0 στις υπόλοιπες $k-1$ κλάσεις. Κατά το label smoothing γίνεται αντικατάσταση αυτού του one-hot encoding του διανύσματος πρόβλεψης και η σωστή κλάση έχει τιμή 0.9 και οι υπόλοιπες έχουν τιμή $1/10k$.

Το κίνητρο πίσω από αυτή την προσέγγιση είναι ότι μπορεί να βοηθήσει με τη μείωση των κλίσεων που ένας επιτιθέμενος θα μπορούσε να εκμεταλλευτεί στην κατασκευή adversarial δειγμάτων [42].

Ωστόσο, αυτή η μέθοδος βρέθηκε ότι *δεν μπορεί να προσφέρει ικανοποιητική προστασία* έναντι σε πιο ακριβές υπολογιστικές επιθέσεις, όπως Jacobian-based επαναληπτικές επιθέσεις [4]. Πιο συγκεκριμένα, ο μηχανισμός στον οποίο βασίζονται η αμυντικές τεχνικές (π.χ. defense distillation, label smoothing) που εξομαλύνουν τις εξόδους των μοντέλων σε πολύ μικρές γειτονιές των δεδομένων εκπαίδευσης, αποτυγχάνουν να εγγυηθούν την ακεραιότητα της λειτουργίας τους. *Είναι σχεδόν βέβαιο ότι θα έχουν περιορισμένη επιτυχία*, λόγω της δυνατότητας μεταφοράς adversarial example, γιατί οποιαδήποτε άμυνα πειράζει ευριστικά τον τρόπο που δημιουργούνται τα adversarial examples (π.χ. με gradient masking) και δεν α-

ντιμετωπίζει το υποκείμενο πρόβλημα των λανθασμένων προβλέψεων του μοντέλου σε τέτοια δείγματα, θα μπορεί να αποφευχθεί χρησιμοποιώντας black-box transferred-based επιθέσεις.

3.6.2.6 Spatial Smoothing

Το **Spatial Smoothing** [70], πρόκειται για μία **data preprocessing** τεχνική άμυνας που είχε προταθεί μαζί με το **feature squeezing**, είναι σχεδιασμένη συγκεκριμένα για χρήση σε εικόνες και θα μπορούσε να καταταχθεί στη γενικότερη κατηγορία αμυνών που εκτελούν **input transformations**. Σκοπός της είναι να μειώσει (filter) τον ανταγωνιστικό θόρυβο που υπάρχει στις εικόνες εισόδου, μέσω διάφορων τεχνικών. Συγκεκριμένα, οι δύο μέθοδοι που προτείνονται είναι:

1. Local Spatial Smoothing: Οι τιμές των γειτονικών pixel χρησιμοποιούνται για να γίνει πιο ομαλό (smooth) το κάθε pixel της εικόνας.
2. Non-Local Spatial Smoothing: Εφαρμογή smoothing σε πολύ μεγαλύτερο μέρος pixel της εικόνας. Αφού βρεθούν αρκετά όμοια pixel τότε η τιμή τους αντικαθίσταται από στη διάμεσο των τιμών αυτών.

3.6.2.7 Input Transformations

Η **Input Transformations** (ή και **Adversarial Transformations**) πρόκειται για μια κατηγορία **data preprocessing** τεχνικών, οι οποίες εφαρμόζουν μετασχηματισμούς, πειράζοντας τα χαρακτηριστικά των εισόδων είτε για να μειώσουν την επίπτωση που έχουν τα adversarial perturbations είτε για να χρησιμοποιηθούν σε επαύξηση του συνόλου δεδομένων εκπαίδευσης (data augmentation) για εκπαίδευση ενός πιο εύρωστου μοντέλου.

Μια από αυτές είναι και η **JPEG Compression** [86], η οποία μπορεί να χρησιμοποιηθεί ως τεχνική προεπεξεργασίας σε ένα βήμα της διαδικασίας της ταξινόμησης για να αντιμετωπίσει adversarial attacks και να μειώσει δραματικά την επίδρασής τους. Η JPEG συμπίεση έχει εφαρμογή σε εικόνες, και έχει την ικανότητα να αφαιρεί στοιχεία σήματος υψηλής συχνότητας μέσα σε τετράγωνα block μια εικόνας, το οποίο ισοδυναμεί με επιλεκτικό θάμπωμα της εικόνας, βοηθώντας στην αφαίρεση των πρόσθετων διαταραχών [42].

Μια άλλη τεχνική είναι η **Thermometer Encoding** [87], η οποία εφαρμόζει έναν πολύ απλό μετασχηματισμό, ο οποίος αυξάνει σημαντικά την ευρωστία του δικτύου. Πρόκειται για μια κωδικοποίηση του κάθε χαρακτηριστικού της εισόδου ως ένα δυαδικό διάνυσμα σταθερού μεγέθους. Ο τομέας της εισόδου αρχικά διαιρείται ομοιόμορφα σε b διακριτούς κάδους (buckets), όπου το b αναπαριστά των αριθμό των bit που θα χρησιμοποιηθούν για την κωδικοποίηση του κάθε χαρακτηριστικού. Η κωδικοποιημένη τιμή ανά χαρακτηριστικό αντιστοιχεί σε έναν αριθμό άσσων (1) ίσων με τον δείκτη του κάδου που περιέχει την αρχική τιμή. Οι άσσοι (1) γεμίζουν από το πίσω μέρος του διανύσματος, και καθώς η αναπαράσταση είναι σταθερού μεγέθους, το υπόλοιπο διάνυσμα είναι γεμάτο με μηδενικά (0).

Δύο άλλες προτεινόμενες τεχνικές οι οποίες είναι μη διαφορίσιμες (non-differentiable) και εγγενώς τυχαίες, κάτι που τις κάνει δύσκολες για έναν επιτιθέμενο για να τις παρακάμψει, είναι οι εξής. Η **Total Variance Minimization (TVM)** [75], πρόκειται για μια τεχνική προεπεξεργασίας που ελαχιστοποιεί τη συνολική διακύμανση της εικόνας εισόδου, στην οποία επιλέγεται τυχαία ένα μικρό σύνολο pixel και γίνεται αναδόμηση της απλούστερης εικόνας που είναι συνεπής με τα επιλεγμένα pixel, έτσι η ανακατασκευασμένη εικόνα δεν περιέχει τα adversarial perturbations επειδή αυτές οι διαταραχές τείνουν να είναι μικρές και εντοπισμένες.

Η **Image Quilting** [75], είναι μια μη παραμετρική μέθοδος η οποία συνθέτει εικόνες, συνδυάζοντας μαζί κομμάτια μικρών εικόνων (patches) τα οποία έχουν προέλθει από ένα σετ δεδομένων από patches καθαρών εικόνων, το οποίο πρέπει να έχει δημιουργηθεί από πριν. Τα patches που συνθέτουν την τελική εικόνα επιλέγονται βάση των K κοντινότερων γειτόνων (στον χώρο των pixel)των patches της αντίστοιχης ανταγωνιστικής εικόνας, όπου ένας εκ των γειτόνων επιλέγεται τυχαία με βάση ομοιόμορφης κατανομής. Αυτό έχει σαν αποτέλεσμα μια τελική εικόνα η οποία περιέχει όσο το δυνατόν περισσότερα pixels τα οποία δεν έχουν τροποποιηθεί από την ανταγωνιστική έκδοση της.

Μια ακόμα τεχνική που αναλύουμε είναι η **Gaussian Data Augmentation (GDA)** [74], η οποία είναι μια τυπική τεχνική αύξησης δεδομένων στην όραση υπολογιστών (computer vision) που έχει επίσης χρησιμοποιηθεί για τη βελτίωση της ευρωστίας ενός μοντέλου σε ανταγωνιστικές επιθέσεις. Η τεχνική αυτή προσθέτει θόρυβο Gauss σε καθαρά δείγματα του αρχικού συνόλου δεδομένων, πάνω στα οποία θα εκπαιδευτεί ένα μοντέλο ταξινόμησης. Η εφαρμογή του θορύβου μπορεί να γίνει με δύο τρόπους: (i) με αύξηση του συνόλου δεδομένων εκπαίδευσης data augmentation με τα αρχικά δείγματα με τον θόρυβο, ή (ii) με αντικατάσταση των αρχικών δειγμάτων με τα θορυβώδη δείγματα, χωρίς αύξηση του συνόλου δεδομένων. Η συγκεκριμένη μέθοδος θυμίζει αρκετά αυτή του adversarial training η οποία σημειώνει επιτυχία ενάντια σε white-box επιθέσεις, αλλά αυτή η μέθοδος σημειώνει καλύτερη επιτυχία ενάντια σε black-box επιθέσεις καθώς ένα σημαντικό πλεονέκτημα αυτής της άμυνας είναι η ανεξαρτησία της από τη στρατηγική επίθεσης [42].

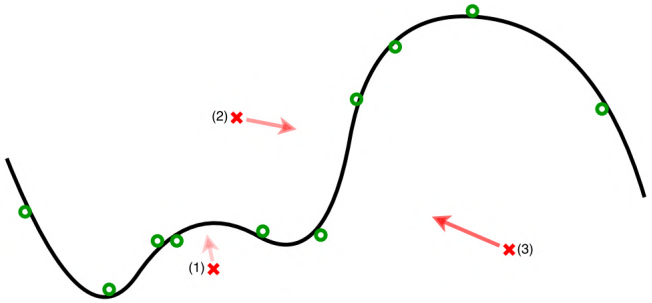
Συνολικά, πολλές φορές οι τυχαίοι μετασχηματισμοί (π.χ. περικοπή εικόνας) και οι μη διαφορίσιμοι (non-differentiable) μετασχηματισμοί (π.χ. total variance minimization) μπορεί να είναι πιο αποτελεσματικές τεχνικές άμυνες ενάντια σε adversarial examples, από ότι ανταγωνιστικοί μετασχηματισμοί που εφαρμόσουν μετασχηματισμούς συγκεκριμένα για την ανακατασκευή των adversarial examples σε καθαρά δείγματα [75].

3.6.2.8 MagNet

Το **MagNet** [76], είναι μια μη ντετερμινιστική και **reactive** αρχιτεκτονική για την προστασία των νευρωνικών δικτύων ενάντια σε adversarial examples, το οποίο δεν τροποποιεί το μοντέλο ούτε γνωρίζει τη διαδικασία παραγωγής των adversarial examples και αποτελείτε από δύο επίπεδα υλοποιημένα με χρήση autoencoders: (i) ένα επίπεδο ανίχνευσης (detector) το οποίο απορρίπτει adversarial δείγματα που περιέχουν μεγάλες διαταραχές και τα οποία απέχουν αρκετά από το όριο απόφασης, (ii) και ένα επίπεδο αναμόρφωσης (reformer) το οποίο αναμορφώνει τις εικόνες που προέρχονται από το επίπεδο της ανίχνευσης για να αφαιρέσει τυχόν υπαρχόντων διαταραχών τα οποία εξακολουθούν να υπάρχουν. Η αρχιτεκτονική αυτή λειτουργεί σαν 'μαγνήτης', προσελκύοντας τα adversarial example που διέφυγαν από το **detection** επίπεδο στις περιοχές του ορίου απόφασης που αντιστοιχούν στις σωστές κλάσεις τους [5].

Η αρχιτεκτονική αυτή βασίζεται στους εξής δύο λόγους που τα adversarial examples οδηγούν τα μοντέλα να κάνουν λάθος αποφάσεις: (i) είτε γιατί το δείγμα είναι πολύ μακριά από τον χώρο των κανονικών δειγμάτων (manifold), αλλά το δίκτυο δεν έχει τη δυνατότητα να το απορρίψει, (ii) είτε γιατί το δείγμα είναι πολύ κοντά στον χώρο των κανονικών δειγμάτων αλλά το μοντέλο δε γενικεύει καλά για δείγματα αυτή της περιοχής (manifold). Το

MagNet προσπαθεί να μάθει να διακρίνει μεταξύ κανονικών και ανταγωνιστικών δειγμάτων προσεγγίζοντας το manifold των κανονικών δειγμάτων, και να ανακατασκευάσει τα adversarial examples μετακινώντας τα προς το manifold, κάτι που είναι αποτελεσματικό για τη σωστή ταξινόμηση των ανταγωνιστικών παραδειγμάτων με μικρές διαταραχές (βλ. σχήμα 3.18).



Σχήμα 3.18: *Αναπαράσταση λειτουργίας τους detector και reformer σε ένα 2D χώρο. Το manifold των κανονικών δειγμάτων αναπαριστάται με την καμπυλωτή γραμμή και τα κανονικά και ανταγωνιστικά δείγματα με πράσινο και κόκκινο χρώμα αντίστοιχα. Ο detector μετρά το κόστος ανακατασκευής και απορρίπτει τα δείγματα με μεγάλο κόστος (2 και 3), ενώ ο reformer βρίσκει ένα δείγμα πάνω στο manifold των κανονικών δειγμάτων για την προσέγγιση του αρχικού (1) [76].*

Δεδομένου ότι δε βασίζεται σε καμία διαδικασία για τη δημιουργία adversarial examples, έχει πολύ σημαντική δύναμη γενίκευσης και εμπειρικά έχει δειχθεί ότι το ότι το MagNet είναι αποτελεσματικό έναντι στις περισσότερες προηγμένες επιθέσεις σε black-box και gray-box σενάρια, ενώ διατηρεί το ποσοστό false positives σε κανονικά δείγματα πολύ χαμηλό [76].

3.6.2.9 GANs

Από την πρώτη εμφάνιση των adversarial examples είχε επινοηθεί η ιδέα χρήσης της παραγωγικής (generative) εκπαίδευσης για την προστασία από τέτοια δείγματα [1]. Ωστόσο, αρκετά αργότερα παρουσιάστηκαν τεχνικές που χρησιμοποιούν Generative Adversarial Networks (GANs) για την προστασία των ML μοντέλων από adversarial examples. Έχουν προταθεί διάφορες αμυντικές τεχνικές οι οποίες ανήκουν κυρίως στην **data preprocessing** κατηγορία αμυνών, καθώς εκτελούν μετατροπές στις εικόνες εισόδου πριν σταλούν σε έναν ταξινομητή. Για αυτήν την κατηγορία αμυνών ενδεικτικά παρουσιάζουμε δύο δημοφιλείς τακτικές με χρήση GANs.

Το **Adversarial Perturbation Elimination GAN (APE-GAN)** [50], πρόκειται για ένα GAN εκπαιδευμένο για την αφαίρεση των adversarial perturbations από εικόνες, εκπαιδεύοντας ταυτόχρονα έναν generator για την παραγωγή καθαρών εικόνων και έναν discriminator για τη διάκριση μεταξύ των καθαρών και ανταγωνιστικών εικόνων, ενισχύοντας συνολικά την ευρωστία του συστήματος. Το σύστημα αυτό είναι σε θέση να προστατεύσει άλλα μοντέλα από τις κοινές ανταγωνιστικές επιθέσεις, χωρίς να χρειάζεται να γνωρίζει καμία λεπτομέρεια για την αρχιτεκτονική ή τις παραμέτρους αυτών [55].

Το **Defense-GAN** [77], είναι ένα framework άμυνας βασισμένο σε GANs, το οποίο εκπαιδεύεται να μαθαίνει την κατανομή των κανονικών εικόνων. Κατά το inference, για κάθε εικόνα εισόδου βρίσκει μια κοντινή έξοδο, που δεν περιέχει adversarial perturbations, την οποία στη συνέχεια τροφοδοτεί στον αρχικό ταξινομητή. Η συγκεκριμένη μέθοδος μπορεί να

χρησιμοποιηθεί με οποιοδήποτε μοντέλο ταξινόμησης και δεν τροποποιεί καθόλου τον ταξινομητή ή τη διαδικασία εκπαίδευσης. Επίσης, μπορεί να χρησιμοποιηθεί ως άμυνα ενάντιων οποιασδήποτε επίθεσης, καθώς δεν προϋποθέτει γνώση της διαδικασίας για τη δημιουργία των adversarial examples και έχει δειχθεί εμπειρικά ότι είναι αρκετά αποτελεσματικό ενάντια σε πλήθος διαφορετικών επιθέσεων.

3.6.2.10 Ensemble Defenses

Όπως έχουμε ήδη αναφερθεί οι **Ensemble** άμυνες (ή **Classifier Ensembles**, ή **Ensemble Models**) είναι άμυνες οι οποίες ενώνουν δύο η περισσότερους ταξινομητές (ή τεχνικές αμυνών) για τη δημιουργία ενός συνόλου, το οποίο θα μπορεί να εφαρμόζει τεχνικές (σε σειρά ή παράλληλα) για να προσφέρει βελτιωμένη ευρωστία ενάντια σε adversarial examples με μεγαλύτερη αποτελεσματικότητα [88]. Οι περισσότερες τεχνικές βασίζονται στη θεώρηση ότι το κάθε μοντέλο που χρησιμοποιείται αντισταθμίζει αμοιβαία τις αδυναμίες που μπορεί να έχει ένα άλλο μοντέλο κατά την ταξινόμηση μιας δεδομένης εισόδου [5].

Όμως, ένας προσαρμοστικός (adaptive) επιτιθέμενος, ο οποίος έχει γνώση της άμυνας που χρησιμοποιείται, μπορεί να κατασκευάσει adversarial perturbations με χαμηλές παραμορφώσεις τα οποία να διατηρούνται και μετά από πολλά επίπεδα αμυνών, *οπότε όσες άμυνες ή μοντέλα και να συνδυαστούν, αν από μόνα τους προσφέρουν αδύναμη άμυνα, δεν πρόκειται σαν σύνολο να παρέχουν επαρκή προστασία ενάντια στα adversarial examples* [84].

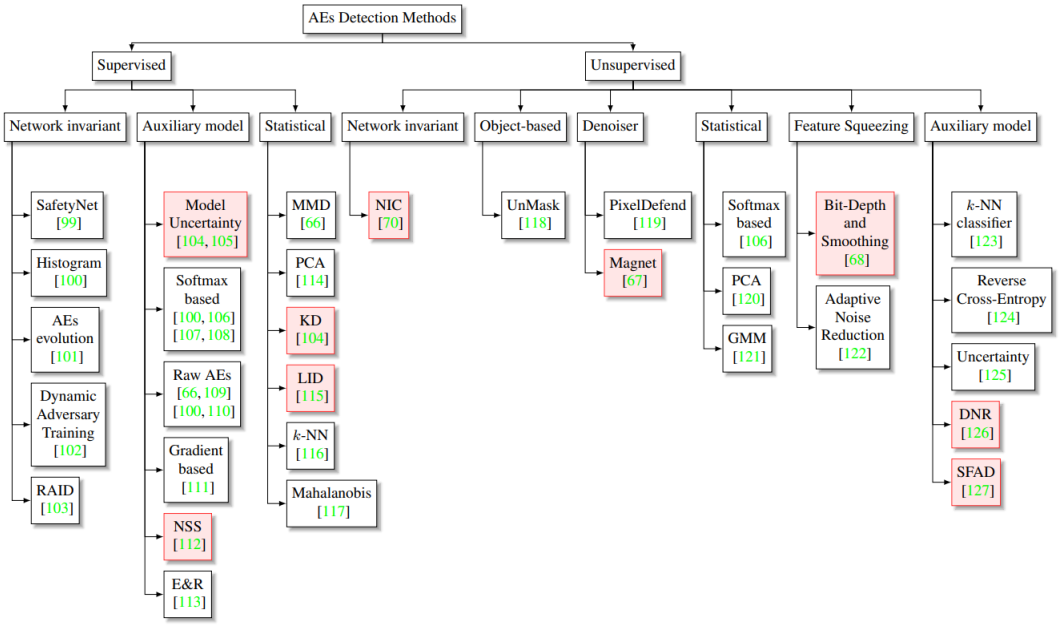
Μια μέθοδος ensemble defense (άμυνας συνόλου), είναι η **Random Self-Ensemble (RSE)** [89], η οποία προσθέτει τυχαίο θόρυβο στα επίπεδα ενός νευρωνικού δικτύου για να αποτρέψει τις ισχυρές επιθέσεις που βασίζονται σε κλίσεις, και υπολογίζει τον μέσο όρο των προβλέψεων που προκύπτουν από τα μοντέλα πάνω σε τυχαίους θορύβους, συνδυάζοντας στην ουσία την πρόβλεψη πάνω σε τυχαίους θορύβους, για να σταθεροποιήσει την απόδοση. Μια άλλη μέθοδος, προτείνει **Adaptive Diversity Promoting (ADP) regularizers** [90], δηλαδή ένα προσαρμοστικό ρυθμιστή προώθησης της ποικιλομορφίας, ο οποίος χρησιμοποιεί μια ποικιλία πολλαπλών μεμονωμένων μοντέλων για να προκύψει μια μη μαξιμαλιστική πρόβλεψη, ενθαρρύνοντας έτσι τη διαφορετικότητα και οδηγώντας σε συνολικά καλύτερη ευρωστία, καθιστώντας δύσκολο στα adversarial examples να μεταφερθούν μεταξύ των μεμονωμένων μελών. Επίσης, το **MagNet** (βλ. 3.6.2.8) που περιγράψαμε προηγουμένως μπορεί να θεωρηθεί ως μια ensemble άμυνα, καθώς χρησιμοποιεί δύο διαφορετικούς autoencoders για την ανίχνευση και αναμόρφωσης της εισόδου, οπότε υπάρχουν δύο επίπεδα στα οποία μπορεί να αντιμετωπιστεί η επίθεση με adversarial examples.

Το **PixelDefend** [91], είναι ένα άλλο χαρακτηριστικό παράδειγμα ensemble άμυνας, στο οποίο ένας adversarial ανιχνευτής και ένας αναμορφωτής εισόδου, είναι ενσωματωμένοι για να περιορίσουν τα adversarial examples. Πρόκειται για μια μέθοδος που μπορεί να εφαρμοστεί σε οποιονδήποτε ταξινομητή και μπορεί να προσφέρει προστασία σε ένα πλήθος ανταγωνιστικών επιθέσεων. Βασίζεται στην υπόθεση ότι τα adversarial examples βρίσκονται κυρίως στις περιοχές χαμηλής πιθανότητας της κατανομής της εκπαίδευσης, ανεξάρτητα των τύπων επιθέσεων και μοντέλων, οπότε και προσπαθεί από το πρώτο στάδιο να μετακινήσει τα ανταγωνιστικά δείγματα προς την κατανομή των δεδομένων εκπαίδευσης.

3.6.2.11 Detection - Auxiliary Detection Models

Οι τεχνικές ανίχνευσης (**Detection**) πρόκειται για μια κατηγορία αμυνών, οι οποίες επεκτείνουν το αρχικό μοντέλο, προσθέτοντάς ένα επίπεδο ή ένα επιπλέον βοηθητικό μοντέλο ανίχνευσης (**Auxiliary Detection Model - ADM**) adversarial examples, η οποία κατά τη λειτουργία του αρχικού μοντέλου, αναγνωρίζει αν κάθε είσοδος είναι adversarial ή όχι και αποφασίζει αν θα προωθήσει την είσοδο στο αρχικό μοντέλο ή όχι [42]. Αυτές οι άμυνες στις ουσία λειτουργούν σαν φίλτρα για το αρχικό μοντέλο, και συνήθως χρησιμοποιούν απλούς binary ML ταξινομητές, για να μην επιφέρουν μεγάλο υπολογιστικό κόστος κατά την χρήση τους ή μεγάλη χρονική καθυστέρηση στη συνολική πρόβλεψη ή ταξινόμηση του μοντέλου [5]. Οι περισσότερες τεχνικές ανίχνευσης εκπαιδεύουν τους ταξινομητές να ξεχωρίζουν την αναπαράσταση των χαρακτηριστικών των adversarial δειγμάτων από τα κανονικά [55], ενώ άλλες τεχνικές εφαρμόζουν **outlier detection**, δηλαδή εισάγουν μια ακραία κλάση outlier κατά την εκπαίδευση του μοντέλου, ώστε ύστερα το μοντέλο να μαθαίνει να ανιχνεύει τα adversarial examples ως ακραία δείγματα (outliers) [92].

Στη βιβλιογραφία εμφανίζονται πάρα πολλές τεχνικές για adversarial detection, οι οποίες μπορούν να κατηγοριοποιηθούν κυρίως σε κατηγορίες ανάλογα με τον τρόπο εκπαίδευσης τους: (i) σε **supervised** ανίχνευση αν χρησιμοποιήθηκαν adversarial examples κατά την εκπαίδευση του ανιχνευτή και σε (ii) **unsupervised** ανίχνευση αν χρησιμοποιήθηκαν μόνο κανονικά δείγματα [93]. Η κάθε μια από αυτές τις κατηγορίες μπορεί να χωριστεί σε περαιτέρω κατηγορίες, ανάλογα με την τεχνική που χρησιμοποιεί, π.χ. για τις supervised τεχνικές υπάρχουν μέθοδοι που είναι **gradient-based** ή **softmax/logits-based**, ενώ π.χ. για τις unsupervised τεχνικές υπάρχουν μέθοδοι που χρησιμοποιούν **denoisers** ή βασίζονται σε στατιστικές ιδιότητες των δειγμάτων (**statistical**). Παρακάτω αναφέρουμε ενδεικτικά κάποια παραδείγματα adversarial detection από τη βιβλιογραφία, που το καθένα χρησιμοποιεί διαφορετική τεχνική.



Σχήμα 3.19: *Κατηγοριοποίηση των μεθόδων για adversarial detection [93].*

Ένα από τα πρώτα μοντέλα adversarial detection που χρησιμοποίησε binary ταξινομητή,

είναι η **Dynamic Adversary Training (DAT)** [67], η οποία είναι μια supervised μέθοδος που ενισχύει τους ανιχνευτές-ταξινομητές, με data augmentation στον προεκπαιδευμένο ταξινομητή σε ένα συγκεκριμένη επίπεδο εξόδου. Η μέθοδος αυτή λαμβάνει την έξοδο του επιπέδου αυτού για καθαρά δείγματα και δημιουργεί εκείνη την ώρα adversarial examplesγια την εκπαίδευση του binary ταξινομητή.

Μια άλλη supervised μέθοδος βασίζεται σε ανίχνευση **Out-of-Distribution** [68], δηλαδή ανίχνευση των adversarial δειγμάτων συγκρίνοντάς την κατανομή των χαρακτηριστικών τους και βρίσκοντας αν είναι μακριά από την κατανομή των καθαρών δειγμάτων. Η συγκεκριμένη τεχνική μπορεί να εφαρμοστεί σε οποιονδήποτε προεκπαιδευμένο softmax νευρωνικό ταξινομητή. Χρησιμοποιεί διακριτική ανάλυση Gauss (Gauss Discriminant Analysis) για την παραγωγή βαθμολογιών εμπιστοσύνης με χρήση της Mahalanobis απόστασης. Η τεχνική αυτή δουλεύει εξίσου καλά για την ανίχνευση των out-of-distribution δειγμάτων και adversarial examples [93].

Μια supervised μέθοδος η οποία δε χρησιμοποιεί binary ταξινομητή, είναι το **SafetyNet** [94], το οποίο χρησιμοποιεί έναν binary RBF-kernel Support Vector Machines (SVMs) ταξινομητή. Το SafetyNet στηρίζεται στην υπόθεση ότι οι ανταγωνιστικές επιθέσεις λειτουργούν παράγοντας διαφορετικά μοτίβα ενεργοποίησης σε ReLUs τελευταίων σταδίων από αυτές που παράγονται από καθαρά παραδείγματα. Οπότε κβαντίζει το τελευταίο επίπεδο ενεργοποίησης ReLU και δημιουργεί τον binary SVM RBF ταξινομητή, ο οποίος είναι υπεύθυνος για την ανίχνευση των adversarial examples.

Μια supervised μέθοδος ανίχνευσης είναι η **Intrinsic-Defender (I-Defender)** [69], η οποία προσπαθεί να συλλάβει τις εγγενείς ιδιότητες ενός DNN ταξινομητή και να τις χρησιμοποιεί για να ανιχνεύσει ανταγωνιστικές εισόδους. Οι εγγενείς ιδιότητες που χρησιμοποιούνται είναι οι κατανομές εξόδου των κρυμμένων επιπέδων του DNN ταξινομητή για καθαρά δείγματα. Επίσης, χρησιμοποιεί Gaussian mixture models (GMM) για να προσεγγίσει την εγγενή κατανομή των κρυφών καταστάσεων της κάθε κλάσης, και ο ανιχνευτής ανακοινώνει για κάθε δείγμα εισόδου αν είναι adversarial, εφόσον η πιθανότητα της κρυφής κατανομής είναι μικρότερη από το όριο της προβλεπομένης κλάσης.

Επίσης, το **MagNet** (βλ. 3.6.2.8), αλλά και το **PixelDefend** (βλ. 3.6.2.10) που περιγράψαμε προηγουμένως περιέχουν ένα μέρος ανίχνευσης, καθώς το πρώτο επίπεδο τους είναι υπεύθυνο για την ανίχνευση adversarial examples.

Όμως, έρευνα έχει δείξει ότι πολλές τεχνικές ανίχνευσης μπορούν να παρακαμφθούν από επιτιθέμενους, δημιουργώντας καινούριες συναρτήσεις απώλειας για την παραγωγή των adversarial examples, ακόμα και αν δεν είναι γνωστή η τεχνική άμυνας που χρησιμοποιείται [71]. Η συγκεκριμένη έρευνα, υποστηρίζει ότι τα adversarial examples είναι πολύ πιο δύσκολο να ανιχνευθούν από ότι υποστηρίζεται και οι ιδιότητες τους που θεωρούνται ξεχωριστές, στην πραγματικότητα δεν είναι. Ενώ σε απλά σύνολα δεδομένων και μικρές παραμορφώσεις είναι πιο εύκολη η ανίχνευση τους, σε πολύπλοκα σύνολα δεδομένων *τα adversarial examples είναι δυσδιάκριτα από τις αρχικές κανονικές εικόνες.*

Defense	Objective	Approach	Details
Adversarial Training [1]	Proactive	Model Hardening	Empirical Robustness
Certified Training [95]	Proactive	Model Hardening	Provable Robustness
DeepDefense [80]	Proactive	Model Hardening	Network Regularization
DAEs [81]	Proactive	Model Hardening	Network Regularization
Parseval Networks [82]	Proactive	Model Hardening	Network Regularization
Defense Distillation [51]	Proactive	Model Hardening	Train a robust distilled model
Label Smoothing [85]	Proactive	Model Hardening	Reduce gradients
Feature Squeezing [70]	Reactive	Preprocessing	Reduce input data dimensionality
Spatial Smoothing [70]	Reactive	Preprocessing	Filter out adversarial noise
Feature Denoising [96]	Reactive	Preprocessing	Suppress adversarial noise
JPEG Compression [86]	Reactive	Preprocessing	Removes adversarial noise
Thermometer Encoding [87]	Reactive	Preprocessing	Fixed-size binary vector encoding
TVM [75]	Reactive	Preprocessing	Image reconstruction from set of pixels
Image Quilting [75]	Reactive	Preprocessing	Image synthesis from patches
GDA [74]	Reactive	Preprocessing	Gaussian noise addition
MagNet [76]	Reactive	Preprocessing	Detect and reform adversarial examples
APE-GAN [50]	Reactive	Preprocessing	GANs
Defense-GAN [77]	Reactive	Preprocessing	GANs
RSE [89]	Reactive	Preprocessing	Ensemble random noise
ADP [90]	Reactive	Preprocessing	Ensemble with diversity
PixelDefend [91]	Reactive	Preprocessing	Ensemble & Detection
DAT [67]	Reactive	Detection	Binary ADM
Out-of-Distribution [68]	Reactive	Detection	Outlier Detection
SafetyNet [94]	Reactive	Detection	Binary SVM RBF ADM
I-Defender [69]	Reactive	Detection	Unsupervised Detection

Πίνακας 3.4: *Συγκριτικός πίνακας των state-of-the-art ανταγωνιστικών αμυνών με τα κύρια χαρακτηριστικά τους*

Κεφάλαιο 4

Ανάλυση Ευρωστίας Μοντέλων Μηχανικής Μάθησης και Εργαλεία

Στο κεφάλαιο αυτό ορίζεται η έννοια των εύρωστων μοντέλων μηχανικής μάθησης (robust ML models) αλλά και γενικότερα των συστημάτων τεχνητής νοημοσύνης και αναλύονται οι μέθοδοι και οι τρόποι αξιολόγησης της ευρωστίας αυτών.

4.1 Εύρωστα Μοντέλα Μηχανικής Μάθησης

Στο προηγούμενο κεφάλαιο παρουσιάσαμε τα adversarial examples, εισόδους με μικρές αλλά συγκεκριμένες διαταραχές με σκοπό την πρόκληση αστοχίας σε ML συστήματα, επιθέσεις που εκμεταλλεύονται αυτή την ευαλωτότητα των ML συστημάτων για την πρόκληση δυσλειτουργιών, αλλά και αμυντικές τεχνικές που μπορούν να χρησιμοποιηθούν για να περιοριστούν αυτά τα κενά ασφαλείας. Όταν ένα ML σύστημα αντιστέκεται σε ανταγωνιστικές εισόδους και μπορεί να διατηρήσει την ακρίβεια και την επίδοση του, τότε λέμε ότι είναι **εύρωστο** (**robust**). Όλες αμυντικές τεχνικές προσπαθούν στην ουσία να βελτιώσουν την ευρωστία (robustness) αυτών των συστημάτων χρησιμοποιώντας τις διάφορες μεθόδους που αναλύσαμε στην ενότητα 3.6. Για παράδειγμα, για έναν ταξινομητή εικόνων, ο στόχος μιας άμυνας είναι να μπορέσει να ταξινομήσει μια ανταγωνιστική εικόνα σωστά με μικρή απώλεια απόδοσης σε σχέση με την αντίστοιχη 'καθαρή' εικόνα [97].

Εφόσον τα ML συστήματα, ειδικότερα τα βαθιά νευρωνικά δίκτυα, έχουν γίνει ολοένα και πιο αποτελεσματικά και χρησιμοποιούνται σε όλο και περισσότερους κρίσιμους τομείς είναι πολύ σημαντική η **αξιολόγηση της ευρωστίας** αυτών των συστημάτων, για την ασφάλεια αλλά και την επέκταση χρήσης και σε άλλους κρίσιμους τομείς. Για παράδειγμα, εάν χρησιμοποιούμε νευρωνικά δίκτυα σε αυτόνομα αυτοκίνητα, ανταγωνιστικές επιθέσεις θα μπορούσαν να επιτρέψουν σε έναν εισβολέα να αναγκάσει το αυτοκίνητο να προβεί σε ανεπιθύμητες ενέργειες [26].

Ενώ οι ανταγωνιστικές επιθέσεις ενάντια στα ML συστήματα έχουν ερευνηθεί αρκετά τα τελευταία χρόνια, η έρευνα και ανάπτυξη των ανταγωνιστικών αμυνών δεν έχει την ίδια πρόοδο, καθώς πάρα πολλές αμυντικές τεχνικές που προτείνονται, μετά από αξιολογήσεις αποδεικνύεται ότι δεν προσφέρουν πραγματική ασφάλεια ή ότι δεν προσφέρουν τα ποσοστά ευρωστίας που υπόσχονται [78]. Αντίστοιχα, ο τομέας της αξιολόγησης της ευρωστίας αυτών των συστημάτων είναι ακόμα ένα ανοιχτό πεδίο έρευνας και λείπουν συστηματικοί τρόποι κατανόησης και καταγραφής τις προόδου που γίνεται, αλλά υπάρχουν τρόποι για να αξιολογηθεί η ευρωστία και να μπορεί να είναι συγκρίσιμη [98]. Για τη σωστή αξιολόγηση, πρέπει να καθοριστούν οι κατάλληλες συνθήκες, όπως π.χ. οι στόχοι μιας άμυνας, οι ικανότητες και η γνώση του επιτιθέμενου, και τα αποτελέσματα μιας αξιολόγησης θα πρέπει να είναι επανα-

λαμβανόμενα και να μην εξαρτώνται από περιβαλλοντικές ή μη ελεγχόμενες συνθήκες. Ένα τέτοιος τρόπος αξιολόγησης μπορεί να γίνει με κάποιες καλές ορισμένες μετρήσεις (**metrics**) οι οποίες μπορούν να χρησιμοποιηθούν σε οποιαδήποτε περίπτωση ανεξάρτητα την αμυντική τεχνική που χρησιμοποιείται (βλ. ενότητα 4.4). Πρέπει όμως να υπάρχουν και **προσαρμοστικές** (adaptive) αξιολογήσεις, που θα δημιουργούνται για κάθε αμυντική τεχνική για να βρεθούν οι αδυναμίες της και ένα εμπειρικό άνω όριο ευρωστίας [28].

Αρχικά, για τη σωστή αξιολόγηση της ευρωστίας γίνεται ο διαχωρισμός του σε 2 διακριτά είδη, ανάλογα και με την αμυντική τεχνική που χρησιμοποιείται, στις **εμπειρικές** (empirical) άμυνες οι οποίες δεν μπορούν να προσφέρουν εγγύηση για τον βαθμό ευρωστίας που παρέχουν, και στις **πιστοποιημένες** (certified) άμυνες οι οποίες παρέχουν εγγυήσεις μέχρι κάποιο όριο για την ευρωστία που παρέχουν. Στην ενότητα 4.3 παρουσιάζουμε τις διαφορές τους αναλυτικά καθώς και τα προτερήματα και μειονεκτήματα που παρουσιάζουν.

Επίσης, είναι σημαντικό πέρα από τη μεμονωμένη αξιολόγηση του κάθε ML συστήματος, να μπορούν να υπάρχουν τρόποι για τη συγκριτική αξιολόγηση της ευρωστίας, κάτι το οποίο επιτυγχάνεται και στις δύο περιπτώσεις με τα κατάλληλα **benchmarks** (βλ. ενότητα 4.6), τα οποία αντικατοπτρίζουν την ευρωστία των συστημάτων κάτω από τις ίδιες συνθήκες.

Η ενίσχυση της ευρωστίας των ML συστημάτων δεν έρχεται χωρίς κόστος όμως. Υπάρχουν κάποιοι συμβιβασμοί (**trade-offs**) που πρέπει να ληφθούν υπόψιν, όταν αυξάνουμε την ευρωστία, όπως η αύξηση του υπολογιστικού κόστους και χρόνου της εκπαίδευσης, αλλά ακόμα και η επίδραση που έχει στη 'καθαρή' ακρίβεια του συστήματος, δηλαδή σε δείγματα που δεν είναι adversarial (βλ. ενότητα 4.7). Το τελευταίο ειδικά, είναι κάτι το οποίο μπορεί να έχει σημαντικό αντίκτυπο σε κρίσιμους τομείς στους οποίες η ακρίβεια είναι πολύ σημαντικός παράγοντας, και η πιθανή πτώση της είναι ανεπιθύμητη. Οπότε πρέπει να ληφθούν υπόψιν άλλες τεχνικές που θα περιορίζουν ή θα ανιχνεύουν τις ανταγωνιστικές επιθέσεις, χωρίς όμως να επηρεάζουν την ακρίβεια του βασικού συστήματος [99].

Για τη βοήθεια στην αξιολόγηση της ευρωστίας των ML συστημάτων, έχουν αναπτυχθεί **εργαλεία και βιβλιοθήκες** οι οποίες παρέχουν τη δυνατότητα στους χρήστες, να εκπαιδεύσουν εύρωστα μοντέλα ή να προσφέρουν έτοιμες συναρτήσεις και μηχανισμούς για την ενίσχυση της ευρωστίας αλλά και να μπορέσουν να αξιολογήσουν την ευρωστία με τις κατάλληλες επιθέσεις και μετρήσεις, για κάθε πιθανή ML εργασία σε εφαρμογές εικόνων, βίντεο, ήχου, ή κείμενου (βλ. ενότητα: 4.8).

4.2 Ορολογία Ευρωστίας Συστημάτων

Αρχικά ορίζουμε κάποιες έννοιες και μαθηματικά σύμβολα για να μπορούμε να τα χρησιμοποιούμε στις επόμενες ενότητες, χωρίς να χρειάζεται να επαναλάβουμε τον ορισμό τους. Στη βιβλιογραφία θα συναντήσουμε τους παρακάτω όρους με διαφορετικές ονομασίες ή σύμβολα κάθε φορά, αλλά για αυτήν την εργασία επιλέγουμε τους παρακάτω.

4.2.1 Καθαρή Ακρίβεια

Ως καθαρή ακρίβεια (clean, standard ή conventional accuracy) ορίζεται η τυπική ακρίβεια δοκιμής σε κανονικά δεδομένα, χωρίς δηλαδή δείγματα με adversarial perturbations [100], [98]. Είναι ένα μέτρο του πόσο καλά το μοντέλο μπορεί να ταξινομήσει σωστά τα δείγματα εισόδου από το σύνολο δεδομένων στο οποίο εκπαιδεύτηκε. Με άλλα λόγια, η καθαρή ακρίβεια είναι η

ακρίβεια του μοντέλου όταν δοκιμάζεται σε δεδομένα που δεν περιέχουν σκόπιμα adversarial perturbations.

Ποσοτικά μπορούμε να την ορίσουμε ως η αναλογία σωστών προβλέψεων ή ταξινομήσεων πάνω στα κανονικά δεδομένα:

$$clean\ accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ all\ clean\ samples}$$

Συνοπτικά, η **Καθαρή Ακρίβεια (Clean Accuracy)** αξιολογεί την απόδοση του μοντέλου σε τυπικά, καθαρά δεδομένα.

4.2.2 Εύρωστη Ακρίβεια

Η εύρωστη ακρίβεια μετράει την απόδοση ενός μοντέλου όταν υποβάλλεται σε adversarial examples παραδείγματα κατά τη διάρκεια της δοκιμής. Αξιολογεί πόσο καλά το μοντέλο γενικεύει σε perturbed δεδομένα και διατηρεί ακριβείς προβλέψεις παρά την παρουσία ανταγωνιστικών επιθέσεων. Πρόκειται για μια κρίσιμη μέτρηση για την αξιολόγηση της αποτελεσματικότητας της ανταγωνιστικής εκπαίδευσης, ως προς την ενίσχυση της ικανότητας του μοντέλου να χειρίζεται ανταγωνιστικές εισόδους [98].

Ποσοτικά μπορούμε να την ορίσουμε ως η αναλογία σωστών προβλέψεων ή ταξινομήσεων πάνω στα adversarial examples:

$$robust\ accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ all\ perturbed\ samples}$$

Συνοπτικά, η **Εύρωστη Ακρίβεια (Robust Accuracy)** μετράει την απόδοσή του μοντέλου σε περίπτωση σκόπιμα κατασκευασμένων adversarial examples.

4.2.3 Ρυθμίσεις Διαταραχών

Έχουμε ήδη αναφέρει κάποιες έννοιες και στις προηγούμενες ενότητες για τα adversarial examples 3.2.2, και το adversarial training 3.3.2, αλλά επαναλαμβάνουμε τον ορισμό κάποιων από αυτών εδώ για να είναι ξεκάθαρη η ερμηνεία τους, όσο αναφορά τη χρήση τους σε αυτό το κεφαλαίο.

Ορίζουμε ως Δ το σύνολο όλων των πιθανών διαταραχών ($\Delta \subseteq \mathbb{R}^d$), και ως δ το διάνυσμα που αναπαριστά μία διαταραχή ($\delta \in \mathbb{R}^d$).

Ορίζουμε ως ℓ_p-σφαίρα (ή αλλιώς ℓ_p-φραγμένο) ένα σύνολο διαταραχών, όπου αυτό το σύνολο περιέχει διαταραχές όπου η νόρμα (μέγεθος) τους δεν ξεπερνάει ένα συγκεκριμένο όριο (threshold) ϵ:

$$\Delta_p = \left\{ \delta \in \mathbb{R}^d, \|\delta\|_p \leq \epsilon \right\}$$

Οι πιο συνήθεις τιμές που χρησιμοποιούνται στη βιβλιογραφία για τα σύνολα διαταραχών είναι: ℓ_1, ℓ_2 και ℓ_∞. Όσο αναφορά για τα μεγέθη, για παράδειγμα, η ℓ_∞ νόρμα ενός διανύσματος z ορίζεται ως εξής:

$$\|z\|_\infty = \max_i |z_i|$$

Στην ουσία για τα ℓ_∞ σύνολα διαταραχών, η διαταραχή επιτρέπεται να έχει μέγεθος από

$[-\epsilon, \epsilon]$ σε καθένα από τα μέρη του διανύσματος. Στις περιπτώσεις των εικόνων με διαταραχές, για μικρά μεγέθη ϵ στο κάθε pixel της εικόνας θα προστεθεί μια μικρή τιμή, οι οποίες στο σύνολο θα δημιουργήσουν μια εικόνα που δε θα ξεχωρίζει με το ανθρώπινο μάτι από την αρχική [38].

Γενικότερα το σύνολο (ℓ_p, ϵ) καθορίζει την κάθε επίθεση, και συνήθως οι συγκρίσεις για ευρωστία μεταξύ διαφορετικών μοντέλων ή τεχνικών έχουν νόημα μόνο για τις ίδιες τιμές (ℓ_p, ϵ), π.χ. ℓ_∞ και $\epsilon = 0.1$.

4.3 Είδη Ευρωστίας

Σε αυτήν την ενότητα, διακρίνουμε τα διαφορετικά είδη ευρωστίας που μπορούν να προσφέρουν οι ανταγωνιστικές άμυνες που συναντήσαμε στο προηγούμενο κεφάλαιο 3.6. Γενικά, τα μοντέλα άμυνας έναντι των adversarial examples μπορεί να είναι κατηγοριοποιηθούν ευρέως σε δύο κατηγορίες: εμπειρικές (**empirical**) και πιστοποιημένες (**certified**) άμυνες. Οι εμπειρικές άμυνες μπορούν να προσφέρουν μόνο εμπειρική ευρωστία (**Empirical Robustness**) ενάντια σε ανταγωνιστικές επιθέσεις, χωρίς συνήθως να παρέχεται καμία εγγύηση πιστοποίησης [101]. Υπάρχουν όμως και άλλες τεχνικές, τις οποίες δεν αναφέραμε στο προηγούμενο κεφάλαιο, οι οποίες μπορούν να εκπαιδεύσουν τα νευρωνικά δίκτυα παρέχοντας εγγυημένη ευρωστία (**Certified Robustness**), όπως π.χ. η τεχνική του Randomized Smoothing [79]. Στις παρακάτω ενότητες γίνεται παραπάνω ανάλυση μεταξύ αυτών των διαφορετικών ειδών ευρωστίας, των πλεονεκτημάτων αλλά και των μειονεκτημάτων που έχει το κάθε είδος.

4.3.1 Εμπειρική Ευρωστία

Στην ενότητα αυτή παρουσιάζεται η έννοια της εμπειρικής ευρωστίας (Empirical Robustness). Όταν ένα μοντέλο είναι εμπειρικά εύρωστο, σημαίνει ότι είναι εύρωστο από κάποιες υπάρχουσες επιθέσεις για συγκεκριμένες διαταραχές και χωρίς εγγυήσεις ή ορισμένες συγκεκριμένες προσεγγίσεις επαλήθευσης. Επίσης, δεν είναι εγγυημένη η ασφάλεια του από προσαρμοστικές (adaptive) επιθέσεις, δηλαδή επιθέσεις που σχεδιάζονται αποκλειστικά για το συγκεκριμένο μοντέλο γνωρίζοντας την αρχιτεκτονική και τις παραμέτρους του [102]. Πρόκειται για best-effort robustness, δηλαδή για προσπάθεια επίτευξης της βέλτιστης ευρωστίας και αυτά τα μοντέλα μπορεί να είναι ευάλωτα σε πιο περίπλοκες επιθέσεις [103].

Όπως είχαμε ορίσει και στην ενότητα 3.3.2, ο υπολογισμός του Empirical Robustness γίνεται λύνοντας το πρόβλημα της εξωτερικής ελαχιστοποίησης με χρήση των κάτω ορίων. Υπάρχουν πολλοί τρόποι για να λυθεί η συγκεκριμένη ελαχιστοποίηση, αλλά οι περισσότερες χρησιμοποιούν μεθόδους τοπικής αναζήτησης (π.χ. Projected Gradient Descent ή PGD [47]) ή ευριστικές μεθόδους [79]. Για παράδειγμα, η αμυντική τεχνική του adversarial training, προσφέρει την καλύτερες εμπειρική ευρωστία εναντίων γνωστών adversarial επιθέσεων, αλλά χωρίς καμία εγγύηση [101]. Πολλές άμυνες μάλιστα έχουν καταρριφθεί για αυτόν τον λόγο από μεταγενέστερη έρευνα, όπου έχουν βρεθεί νέες επιθέσεις που τις 'σπάνε' [71].

4.3.2 Αποδεδειγμένη Ευρωστία

Στην ενότητα αυτή παρουσιάζεται η έννοια της αποδεδειγμένης ευρωστίας (Certified Robustness). Όταν ένα μοντέλο είναι αποδεδειγμένα εύρωστο σημαίνει ότι υπάρχει ένα κατώτερο όριο ακρίβειας όπου το μοντέλο είναι εγγυημένα εύρωστο από οποιεσδήποτε επιθέσεις και αντιπάλους υπό ορισμένους περιορισμούς, π.χ., για ℓ_∞-φραγμένες επιθέσεις [102].

Ο διαρκής ανταγωνισμός μεταξύ επιτιθεμένων και αμυνομένων δημιούργησε τις συνθήκες για μελέτες σχετικά με τις αποδεδειγμένα εύρωστες προσεγγίσεις για τα βαθιά νευρωνικά δίκτυα, οι οποίες περιλαμβάνουν τόσο την επαλήθευση ευρωστίας (**Robustness Verification**) όσο και τις προσεγγίσεις εύρωστης εκπαίδευσης (**Robust Training**). Οι Robustness Verification προσεγγίσεις στοχεύουν στην αξιολόγηση της ευρωστίας του μοντέλου παρέχοντας ένα θεωρητικά πιστοποιημένο κατώτερο όριο ευρωστίας υπό ορισμένους περιορισμούς διαταραχών, ενώ οι αντίστοιχες Robust Training προσεγγίσεις στοχεύουν στην εκπαίδευση των μοντέλων για τη βελτίωση αυτού του κατώτερου ορίου [102].

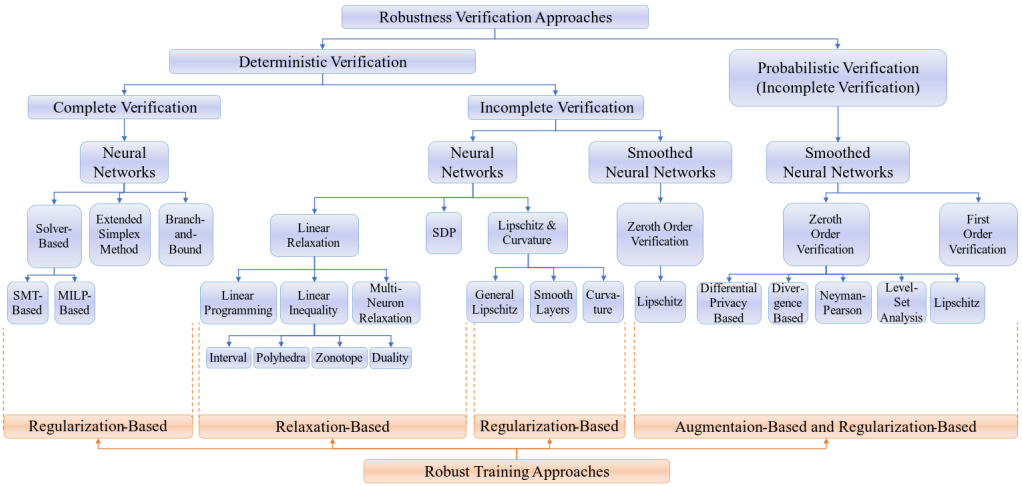
Όπως είχαμε ορίσει και στην ενότητα 3.3.2, ο υπολογισμός του Certified Robustness γίνεται λύνοντας το πρόβλημα της εξωτερικής ελαχιστοποίησης με χρήση των άνω ορίων. Μια αρχική προσέγγιση είναι ο υπολογισμός της διαταραχής στη χειρότερη περίπτωση χρησιμοποιώντας διακριτή βελτιστοποίηση (discrete optimization), όπως έχει προταθεί από τον αλγόριθμο Reluplex [104] ο οποίος βασίζεται στη θεωρία του satisfiability modulo theories (SMT) που χρησιμοποιείται στην πληροφορική για να προσδιορίσει εάν ένας τύπος ή πρόγραμμα ικανοποιεί τις συνθήκες και είναι ορθό. Όμως, αυτήν τη στιγμή αυτές οι ακριβείς προσεγγίσεις μπορεί να χρειαστούν αρκετές ώρες ή περισσότερο για να υπολογιστεί η απώλεια για ένα μόνο παράδειγμα, ακόμη και για μικρά δίκτυα με μερικές εκατοντάδες κρυφές μονάδες, οπότε και η εκπαίδευση ενός ολόκληρου δικτύου καθίσταται ανέφικτη [95]. Πιο συγκεκριμένα, αυτές οι τεχνικές δεν μπορούν να κλιμακωθούν σε μεγάλα νευρωνικά δίκτυα (π.χ. ResNet50) ή σε σύνολα δεδομένων μεγάλων διαστάσεων όπως το ImageNET [101], παρά μόνο σε μικρότερα όπως το MNIST [105].

Για την αποφυγή αυτών των ακριβών υπολογισμών, έχουν προταθεί πολλοί άλλοι τρόποι για να λυθεί η συγκεκριμένη ελαχιστοποίηση με πιο χαλαρές εγγυήσεις, όπως με χρήση *semidefinite relaxations* [95], *linear relaxations* [106], ή τεχνικών όπως *randomized smoothing* [79], και *BOX relaxations* [107], οι οποίες είναι κυρίως πιθανολογικές (probabilistic) και παρέχουν εγγυήσεις για certified robustness με μεγάλη πιθανότητα. Στο διάγραμμα 4.1 φαίνεται με πορτοκαλί χρώμα μια πλήρης ταξινόμηση των τεχνικών που υπάρχουν για robust training με εγγυήσεις.

Στο διάγραμμα 4.1 φαίνεται επίσης με μπλε χρώμα μια πλήρης ταξινόμηση των τεχνικών που υπάρχουν για robust verification. Υπάρχουν, δύο κυρίως τρόποι για να προσεγγίσουμε το robust verification των νευρωνικών δικτύων, με **Deterministic Verification** ή με **Probabilistic Verification** μεθόδους. Οι ντετερμινιστικοί μέθοδοι επαλήθευσης, για κάθε είσοδο που δεν είναι ανθεκτική ενάντια σε μια επίθεση, παρέχουν εγγυημένα ότι η έξοδος θα είναι *μη επαληθευμένη*, ενώ οι πιθανολογικές μέθοδοι επαλήθευσης, αντίστοιχα παρέχουν εγγυημένα ότι η έξοδος θα είναι *μη επαληθευμένη* με μία πιθανότητα (π.χ. 99.9%) όπου η τυχαιότητα είναι ανεξάρτητη της εισόδου [102].

Για τις Deterministic Verification μεθόδους, υπάρχουν, δύο κυρίως τρόποι για να προσεγγίσουμε το robust verification των νευρωνικών δικτύων, είτε μέσω **Exact Verification** (ή αλλιώς **Complete Verification**) είτε μέσω **Relaxed Verification** (ή αλλιώς **Incomplete Verification**) [102].

Οι τεχνικές ακριβών επαληθεύσεων (**exact verification**) προσπαθούν να απαντήσουν πραγματικά στο ερώτημα εάν υπάρχει ή όχι ένα adversarial perturbation για κάθε σημείο δοκιμής ενός μοντέλου. Με άλλα λόγια, η έξοδος τους είναι είτε *NAI*, οπότε δεν υπάρχει



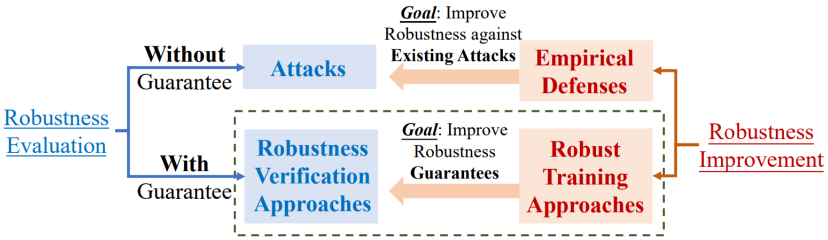
Σχήμα 4.1: *Ταξινόμηση certified robustness προσεγγίσεων. Τα μπλε πλαίσια δείχνουν την ταξινόμηση των robust verification ενώ τα πορτοκαλί πλαίσια δείχνουν την ταξινόμηση των robust training προσεγγίσεων. Οι κάθετες διακεκομμένες γραμμές δείχνουν την κατάλληλη επαλήθευση για τις αντίστοιχες προσεγγίσεις εκπαίδευσης [102].*

adversarial perturbation, είτε *ΟΧΙ*, οπότε υπάρχει adversarial perturbation [105].

Οι τεχνικές χαλαρών επαληθεύσεων (**relaxed verification**) επιτρέπουν μια εναλλακτική έξοδο: *ΙΣΩΣ*, την οποία ο αλγόριθμος πιστοποίησης είναι πάντα σίγουρο ότι θα παράγει. Ωστόσο, εάν η έξοδος είναι *ΝΑΙ*, τότε είναι αλήθεια (με κάποιες προϋποθέσεις) ότι δεν υπάρχουν adversarial perturbation. Και φυσικά, ο αλγόριθμος δεν πρέπει να βγάζει *ΙΣΩΣ* ή *ΟΧΙ* πολύ συχνά [105].

4.3.3 Σύγκριση Εμπειρικής και Αποδεδειγμένης Ευρωστίας

Οι άμυνες και οι τεχνικές ενίσχυσης της ευρωστίας των νευρωνικών δικτύων σε adversarial examples όπως αναφέραμε μπορούν να προσφέρουν είτε empirical είτε certified ευρωστία. Και στις δύο περιπτώσεις, για να αξιολογηθεί η ευρωστία απαιτείται η ίδια μεθοδολογία, η οποία φαίνεται και στο σχήμα 4.2. Αρχικά χρειάζεται μια μέθοδος η οποία θα ενισχύσει την ευρωστία είτε ενάντια σε κάποιες επιθέσεις στην περίπτωση των empirical αμυνών, είτε θα προσφέρει εγγυήσεις ευρωστίας (robustness guarantees) στην περίπτωση των robust training μεθόδων. Έπειτα χρειάζεται κάποιος τρόπος για την αξιολόγηση αυτών των τεχνικών, είτε μέσω νέων adaptive επιθέσεων στην περίπτωση των empirical αμυνών, είτε μέσω robustness verification προσεγγίσεων στην περίπτωση των robust training μεθόδων.



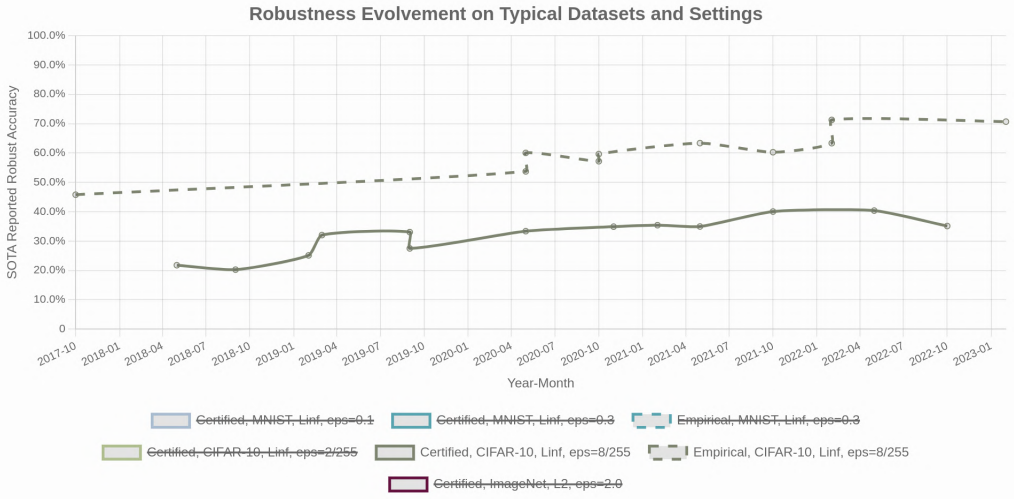
Σχήμα 4.2: *Σχηματική αναπαράσταση των δύο προσεγγίσεων για την αξιολόγηση και τη βελτίωση της ευρωστίας των νευρωνικών δικτύων ενάντια σε ανταγωνιστικές επιθέσεις [102]*

Σχετικά όμως με τη σύγκριση των δύο μεθόδων, οι μέθοδοι που προσφέρουν certified

robustness, έχουν αρχικά το πλεονέκτημα της **εγγυημένης ευρωστίας**, όμως συνήθως έρχονται αντιμέτωπες με το **αυξημένο υπολογιστικό κόστος και τον χρόνο** που χρειάζεται για την εκπαίδευση του νευρωνικού, ανάλογα με το μέγεθος και την πολυπλοκότητα του μοντέλου και των διαστάσεων των δειγμάτων εισόδου. Καθώς η πολυπλοκότητα του μοντέλου και του χώρου εισόδου αυξάνεται, η εύρεση αυστηρών ορίων και εγγυήσεων γίνεται όλο και πιο δύσκολη [101].

Ένα άλλο μειονέκτημα που παρουσιάζουν οι μέθοδοι που προσφέρουν certified robustness, έχει να κάνει και με την **αυξημένο σφάλμα** (robust error) που εμφανίζουν σε σχέση με τις empirical μεθόδους, και αντίστοιχα το μειωμένη ακρίβεια. Για παράδειγμα, σε ένα σύνολο δεδομένων όπως το CIFAR-10, τα πιο γνωστά εύρωστα μοντέλα που μπορούν να χειριστούν μια διαταραχή ($\ell_\infty, epsilon = 8/255 = 0.031$) τιμών χρώματος έχουν (empirical) robust error της τάξης του 55%, ενώ τα καλύτερα αποδεδειγμένα εύρωστα μοντέλα έχουν robust error άνω του 70% [38].

Στην πιο παρακάτω ενότητα 4.6, παρουσιάζουμε αναλυτικά τα δύο επικρατέστερα benchmarks με τις state-of-the-art τεχνικές για empirical robustness μέσω του RobustBench [98] και για certified robustness μέσω του SoK: Certified Robustness for Deep Neural Networks (ή για συντομία SoK Benchmark) [102]. Σε αυτά τα benchmarks φαίνονται οι επιδόσεις των τεχνικών και μοντέλων μέσω του robust accuracy που πετυχαίνουν ανάλογα το επιλεγμένο σύνολο δεδομένων και μέγεθος διαταραχής, ώστε να μπορούν να συγκριθούν μεταξύ τους αποτελεσματικά.



Σχήμα 4.3: *Διάγραμμα εξέλιξης ευρωστίας από δημοσιευμένες μεθόδους δοκιμασμένες σε συγκεκριμένα dataset και συνθήκες. Ο x-άξονας αναπαριστά τον χρόνο χωρισμένο ανά τρίμηνο, ενώ ο y-άξονας αναπαριστά το robust accuracy της κάθε δημοσιευμένης μεθόδου [102].*

Με μία όμως απλή σύγκριση μόνο από το Leaderboard του SoK Benchmark (βλ. 4.3), μπορούμε να δούμε συγκριτικά το robust accuracy από το 2017 έως και το 2023, για συγκεκριμένες συνθήκες (ℓ_∞, $\epsilon = 8/255$) και σύνολο δεδομένων (CIFAR-10), για διάφορες δημοσιευμένες μεθόδους για empirical robustness (διακεκομμένη γραμμή) και για certified robustness (συνεχόμενη γραμμή). Το πρώτο συμπέρασμα που φαίνεται είναι ότι και για τις δύο μεθόδους, υπάρχει συνεχόμενη αύξηση του robust accuracy ανά τα χρόνια, ξεκινώντας από 45.80% και φτάνοντας έως και το 71.29% για τις empirical μεθόδους, και ξεκινώντας από

21.78% και φτάνοντας έως και το 40.39% για τις certified μεθόδους. Το δεύτερο συμπέρασμα που φαίνεται επίσης, είναι ότι το robust accuracy των empirical μεθόδων είναι πάντα υψηλότερο από των certified μεθόδων, με τη διαφορά να αγγίζει το 30.90% για τις καλύτερες δημοσιευμένες τεχνικές και από τις δύο μεθόδους.

Συνοψίζοντας, οι empirical μέθοδοι ενώ δεν προσφέρουν εγγυήσεις, στην πράξη αποδεικνύουν ότι δημιουργούν πρακτικά καλύτερα εύρωστα μοντέλα από τις certified μεθόδους, αυτήν τη χρονική στιγμή. Δηλαδή, παρέχουν εμπειρικά καλύτερο clean accuracy αλλά και robust accuracy για τις καλύτερες διαθέσιμες επιθέσεις [38]. Οπότε είναι σημαντικό να ληφθούν και οι δύο προσεγγίσεις υπόψη κατά τον σχεδιασμό εύρωστων μοντέλων και να επιλεχθούν αυτές που ικανοποιούν κάθε φορά τις ζητούμενες απαιτήσεις.

4.4 Τυποποιημένες Μετρήσεις Ευρωστίας

Στην ενότητα αυτή παρουσιάζουμε μερικές τυποποιημένες μετρήσεις (**robustness metrics**) ευρωστίας που ποσοτικοποιούν το ποσό της διαταραχής (perturbation) που απαιτείται για να προκληθεί μια εσφαλμένη ταξινόμηση (πιο γενικά: η ευαισθησία των εξόδων του μοντέλου σε σχέση με τις αλλαγές στις εισόδους τους) [42].

Αυτή τη στιγμή στη βιβλιογραφία δεν υπάρχουν πολλές μετρήσεις οι οποίες να δουλεύουν καθολικά για όλες τις επιθέσεις, είδη μοντέλων, ή και μεγέθη διαταραχών, και να μπορούν να αξιολογήσουν την ευρωστία μιας τεχνικής άμυνας είτε συνολικά των μοντέλων. Μερικές όμως από αυτές που είναι διαθέσιμες παρουσιάζονται παρακάτω.

4.4.1 Adversarial Attack Success Rate (ASR)

Ως ποσοστό επιτυχίας adverarial επίθεσης (Adversarial Attack Success Rate - ASR) ορίζεται το ποσοστό των adverarial παραδειγμάτων για τα οποία το μοντέλο παρείχε εσφαλμένη απάντηση. Για τη σωστή αξιολόγηση των gradient-based επιθέσεων, συνήθως συγκρίνεται το ASR υπό το ίδιο μέγεθος διαταραχής μεταξύ διαφορετικών επιθέσεων [108]. Όσο πιο χαμηλό είναι αυτό το ποσοστό, τόσο πιο εύρωστο είναι και το μοντέλο που αξιολογείται.

4.4.2 Empirical Robustness

Ως εμπειρική ευρωστία (Empirical Robustness) ορίζεται η μέση ελάχιστη διαταραχή που πρέπει να εισάγει ένας επιτιθέμενος για μια επιτυχημένη επίθεση. Αυτή η μέτρηση αξιολογεί την ευρωστία ενός ταξινομητή σε σχέση με έναν συγκεκριμένο σύνολο δεδομένων επίθεσης και δοκιμής, επομένως δεν μπορεί να αποτελεί ένα μέτρο σύγκρισης μεταξύ διαφορετικών επιθέσεων ή μεγεθών διαταραχής [53].

Αν θεωρήσουμε έναν εκπαιδευμένο ταξινομητή $C(x)$, μια μη στοχευμένη επίθεση $\rho(x)$ και ένα σύνολο δειγμάτων δεδομένων δοκιμής $X = (x_1, \ldots, x_n)$, έστω I το υποσύνολο των δεικτών $i \in 1, \ldots, n$ για το οποίο $C(\rho(x_i)) \neq C(x_i)$, δηλαδή για το οποίο η επίθεση ήταν επιτυχής. Τότε μπορούμε να ορίσουμε την Empirical Robustness (ER) ως εξής [42]:

$$ER(C, \rho, X) = \frac{1}{|I|} \sum_{i \in I} \frac{\|\rho(x_i) - x_i\|_p}{\|x_i\|_p}$$

όπου το p είναι η νόρμα που χρησιμοποιείται στη δημιουργία των adversarial δειγμάτων (εάν ισχύει) και η συνήθης προεπιλεγμένη τιμή είναι $p = 2$.

4.4.3 Local Loss Sensitivity

Η τοπική ευαισθησία απώλειας (Local Loss Sensitivity) μετρά τη μεγαλύτερη διακύμανση μιας συνάρτησης υπό μια μικρή αλλαγή στο η είσοδος του. Όσο μικρότερη είναι η τιμή, τόσο πιο ομαλή είναι η συνάρτηση. Πιο λεπτομερώς, στοχεύει στην ποσοτικοποίηση της ομαλότητας ενός μοντέλου εκτιμώντας την Lipschitz του σταθερά συνέχειας. Πρόκειται για μια μέτρηση ανεξάρτητη της επιλεγμένης επίθεσης, που μπορεί να προσφέρει μια ικανοποιητική εικόνα για τις ιδιότητες ενός μοντέλου [109].

Αν θεωρήσουμε έναν εκπαιδευμένο ταξινομητή $C(x)$ και ένα σύνολο δειγμάτων δεδομένων δοκιμής $X = (x_1, \ldots, x_n)$, τότε μπορούμε να ορίσουμε την Local Loss Sensitivity (LLS) ως εξής [42]:

$$LLS(C, X, y) = \frac{1}{n} \sum_{i=1}^{n} \|\nabla \mathcal{L}(x_i, y_i)\|_2$$

Στην ουσία η Local Loss Sensitivity υπολογίζει τη μέση ευαισθησία της συνάρτησης απώλειας του μοντέλου σε σχέση με τις αλλαγές στις εισόδους [42].

4.4.4 CLEVER score

Η μέτρηση Cross Lipschitz Extreme Value for Network Robustness (CLEVER) εκτιμά, για μια δεδομένη είσοδο x και ℓ_p νόρμα, ένα κατώτερο όριο γ για την ελάχιστη διαταραχή που απαιτείται για να αλλαχθεί η ταξινόμηση του x, δηλαδή $\|x - x'\|_p < \gamma$ υποδηλώνει $C(x) = C(x')$ [110]. Δεδομένου ότι γενικά δεν υπάρχει έκφραση κλειστής μορφής ή άνω όρια για τη σταθερά Lipschitz, ο αλγόριθμος CLEVER χρησιμοποιεί μια εκτίμηση που βασίζεται στη θεωρία ακραίων τιμών [42].

Στην ουσία ο αλγόριθμος CLEVER υπολογίζει ένα κατώτερο όριο για την ελάχιστη διαταραχή που απαιτείται για την αλλαγή της ταξινόμησης. Η βαθμολογία CLEVER μπορεί να υπολογιστεί για μια στοχευμένη ή και μη στοχευμένη επίθεση.

4.5 Αξιολόγηση Ευρωστίας

Στην ενότητα αυτή παρουσιάζεται η θεωρητική μεθοδολογία και οι προτάσεις για τη σωστή αξιολόγηση της ευρωστίας των ML μοντέλων και αμυνών, σύμφωνα με συστάσεις από την πρόσφατη βιβλιογραφία, καθώς το πεδίο της ανταγωνιστικής ευρωστίας είναι αρκετά πρόσφατο και ταχύτητα εξελισσόμενο.

4.5.1 Μεθοδολογία Αξιολόγησης Ευρωστίας

Κάνοντας μια σύνοψη όλων αυτών που έχουμε αναφέρει ήδη για την ευρωστία των νευρωνικών δικτύων, αναφέραμε αρχικά για τον εμπειρικό έλεγχο της ευρωστίας με διάφορες καλές επιθέσεις όπως FGSM [1], JSMA [49], DeepFool [53] και Carlini & Wagner [26], οι οποίες όμως δεν μπορούν να επιβεβαιώσουν την ευρωστία τους. Στη συνέχεια, αναφερθήκαμε στην επιβεβαίωση της ευρωστίας με ολοκληρωμένες και ακριβείς μεθόδους όπως το Reluplex [104] οι οποίες όμως δεν μπορούν να κλιμακωθούν σε σύνολα δεδομένων μεγαλύτερα του MNIST. Ύστερα, αναφερθήκαμε στην επιβεβαίωση με πιο χαλαρές εγγυήσεις ή με πιθανολογικές μεθόδους όπως με χρήση Randomized Smoothing [79] για την αποφυγή του υπολογιστικού κόστους και χρόνου. Τέλος, αναφερθήκαμε στις μετρήσεις που μπορούμε να χρησιμοποιήσου-

με για να έχουμε έναν πιο αντικειμενικό τρόπο να συγκρίνουμε τεχνικές μεταξύ τους, όπως Local Loss Sensitivity [109] και CLEVER score [110].

Σε αυτήν την ενότητα θα αναφέρουμε κάποιες κατευθυντήριες γραμμές και προτάσεις που επικρατούν στη βιβλιογραφία, για την πιο σωστή αξιολόγηση της ευρωστίας των μοντέλων, αλλά και συχνά λάθη τα οποία γίνονται.

4.5.1.1 Ορισμός του μοντέλου απειλής

Αρχικά, είναι σημαντικό να καθοριστεί το κατάλληλο μοντέλο απειλής (threat model) κάτω από το οποίο αξιολογούμε την ευρωστία ενός μοντέλου [26], καθώς αυτό αλλάζει ανάλογα με το είδος και το τις ικανότητες του επιτιθέμενου. Δηλαδή πρέπει να καθοριστεί η **γνώση του επιτιθέμενου** (white-box, black-box), η **επιρροή του επιτιθέμενου** (evasion, poisoning) όπως και οι άλλες ιδιότητες που μπορεί να έχει ένα επιτιθέμενος που αναφέραμε στην ενότητα 3.4.1, καθώς και το **μέγεθος του της επίθεσης** (ℓ_p, ϵ).

4.5.1.2 Αρχές αξιολογήσεων αμυνών

Εφόσον έχουμε μια adversarial άμυνα ή γενικότερα τεχνική ενίσχυσης της ευρωστίας ενός μοντέλου, όπως αυτές που παρουσιάσαμε στην ενότητα 3.6, μπορούμε να την αξιολογήσουμε αρχικά με βάση τις παρακάτω τρεις αρχές, σύμφωνα με το άρθρο [78] στο οποίο γίνεται για πρώτη φορά μια αναλυτική ανασκόπησή των συνθηκών και μεθόδων για την αξιολόγηση της ευρωστίας:

1. **Άμυνα ενάντια σε αντίπαλο που επιτίθεται στο σύστημα**. Είναι σημαντικό να ληφθούν υπόψη και να μελετηθούν οι αντίπαλοι που μπορεί να υπάρχουν για το κάθε ML σύστημα και να γίνει το σωστό μοντέλο απειλής που αναφέραμε παραπάνω. Για παράδειγμα, διαφορετικές ικανότητες και στόχους έχει ένας επιτιθέμενος εναντίων ενός συστήματος αυτόνομης οδήγησης που θέλει να προκαλέσει σφάλμα στο ML σύστημα που αναγνωρίζει τα οδικά σήματα [49], από έναν επιτιθέμενο εναντίων ενός ML ad-blocker που θέλει να αναγνωρίσει μια διαφήμιση ως κανονικό περιεχόμενο [111].
2. **Έλεγχος ευρωστίας στη χειρότερη περίπτωση**. Σε ένα πραγματικό περιβάλλον υπάρχει εγγενής τυχαιότητα η οποία δεν μπορεί να προβλεφθεί. Αναλύοντας όμως την ευρωστία στη χειρότερη δυνατή περίπτωση (worst-case), δηλαδή θεωρώντας έναν ισχυρό αντίπαλο ο οποίος έχει κάθε γνώση για το σύστημα, αν αποτυγχάνει να το φέρει σε δυσλειτουργία, μπορούμε να κάνουμε υποθέσεις για την καλή λειτουργία του και στο πραγματικό περιβάλλον. Σε σχέση με προσεγγίσεις τυχαίων δοκιμών (random testing), αυτή η προσέγγιση μπορεί να ξεχωρίσει ένα σύστημα το οποίο μπορεί να προκαλέσει σφάλμα μία φορά σε δισεκατομμύρια προσπάθειες, από ένα σύστημα το οποία δε θα προκαλέσει ποτέ.
3. **Μέτρηση προόδου της τεχνητής νοημοσύνης σε σχέση με τις ανθρώπινες ικανότητες**. Για να διασφαλιστεί η πρόοδος της τεχνητής νοημοσύνης, είναι σημαντικό να μπορεί να γίνει κατανοητό γιατί οι ML αλγόριθμοι και μοντέλα αποτυγχάνουν σε συγκεκριμένες ρυθμίσεις. Στη βιβλιογραφία, υπάρχουν πολλά παραδείγματα ML μεθόδων που έχουν πολύ μικρό χάσμα απόδοσης σε σχέση με τους ανθρώπους, όπως μοντέλα που είναι εκπαιδευμένα να παίζουν Go ή σκάκι [112]. Όμως, στο κομμάτι του adversarial robustness το χάσμα είναι τεράστιο, καθώς μικρές αλλαγές στην είσοδο οι οποίες δεν προκαλούν καμία διαφορά στη συμπεριφορά των ανθρώπων, μπορούν να

επηρεάσουν δραματικά την επίδοση των ML συστημάτων. Για αυτό και το adversarial robustness μπορεί να είναι ένα μέτρο της ML προόδου που είναι άμεσα εξαρτημένη ως προς την απόδοση.

4.5.1.3 Προτάσεις για αξιολόγηση

Σε αυτήν την ενότητα θα αναφερθούμε σε κάποιες προτάσεις για την πιο έγκυρη αξιολόγηση της ευρωστίας των αμυνών ή μοντέλων. Αυτές οι προτάσεις, σύμφωνα με τους συγγραφείς του [78], προέρχονται από τα πιο συνηθισμένα λάθη που έχουν γίνει σε αξιολογήσεις αμυνών, και προσφέρονται ώστε να αποφευχθούν τα ίδια λάθη σε μελλοντικές αξιολογήσεις:

1. **Εκτέλεση ισχυρών προσαρμοστικών (adaptive) επιθέσεων για να δοθεί ένα άνω όριο ευρωστίας**. Οι επιθέσεις πρέπει να έχουν πλήρη γνώση τις άμυνες και του μοντέλου για να μπορέσουν να βρεθούν οι αδυναμίες και το όριο προστασίας που παρέχουν. Πρέπει να υπάρξει εστίαση στις πιο ισχυρές επιθέσεις και όχι σε γενικές ή αδύναμες επιθέσεις, για να μπορέσει να δοθεί ένα αντιπροσωπευτικό άνω όριο.
2. **Δημοσίευση των εκπαιδευμένων μοντέλων και του πηγαίου κώδικα**. Εφόσον είναι δυνατόν, είναι σημαντικό να δημοσιεύεται και ο κώδικας ή το εκπαιδευμένο μοντέλο μαζί με κάποιο άρθρο, ώστε να είναι πιο εύκολη η διαδικασία εξωτερικών αξιολογήσεων από άλλους ερευνητές ή μηχανικούς, καθώς δε χρειάζεται να ξαναγίνει υλοποίηση των τεχνικών, κάτι που προσθέτει επιπλέον κόπο, χρόνο και είναι πιθανό σε σφάλματα κατά την υλοποίηση.
3. **Αναφορά της καθαρής ακρίβειας του μοντέλου όταν δε δέχεται επιθέσεις**. Είναι σημαντικό μαζί με το robust accuracy που αναφέρεται πάντα, να εμφανίζεται και το clean accuracy, καθώς ένα πολύ εύρωστο μοντέλο να χάνει σημαντικό μέρος της κανονικής του ακρίβειας.
4. **Διεξαγωγή βασικών δοκιμών (sanity tests)**. Για να μπορέσει να επαληθευτεί το robust accuracy, πρέπει να γίνου δοκιμές σε διάφορες συνθήκες, π.χ. δοκιμές με τυχαίο θόρυβο. Πρέπει να δοκιμαστούν διαφορετικές υπερπαράμετροι τις επίθεσης και να επιλεχθούν αυτές με την καλύτερη επίδοση.
5. **Εκτέλεση ενός ποικίλου συνόλου επιθέσεων**. Για την καλύτερη επιβεβαίωση, είναι σημαντικό να δοκιμαστούν πολλές και διαφορετικές μεταξύ τους επιθέσεις. Δε θα πρέπει να γίνεται τυχαία εφαρμογή πολλών επιθέσεων οι οποίες μπορεί να έχουν παρόμοιες προσεγγίσεις, γιατί δε θα προσφέρουν ένα ποιοτικό αποτέλεσμα.
6. **Εκτέλεση επιθέσεων με δυνατότητα μεταφοράς (transferability attack) μέσω ενός υποκατάστατου μοντέλου**. Επειδή τα adversarial examples έχουν συχνά τη δυνατότητα να μεταφερθούν σε διαφορετικά μοντέλα, είναι σημαντικό να επιλεχθεί ή κατασκευαστεί ένα υποκατάστατο μοντέλο (substitute model) από το αρχικό, και να εκλεχθεί αν οι επιθέσεις στο αρχικό μοντέλο (white-box) έχουν την ίδια επίπτωση και στο υποκατάστατο (black-box).
7. **Διερεύνηση εάν είναι δυνατή η χρήση προσεγγίσεων αποδεδειγμένης ευρωστίας**. Για να δοθεί ένα κάτω όριο ευρωστίας είναι σημαντικό να ελεγχθεί αν είναι δυνατόν να γίνουν χρήση robust verification τεχνικών.
8. **Χρήση στοχευμένων και μη στοχευμένων επιθέσεων**. Για την πιο σωστή

αξιολόγηση της ευρωστίας, είναι σημαντικό να ελεγχθούν και τα δύο σενάριο επιθέσεων, καθώς είναι πολύ πιο εύκολο να επιτύχει μια μη στοχευμένη επίθεση. Αλλιώς θα πρέπει να δηλώνεται ρητά ποιο σενάριο επίθεσης εκτελέστηκε.

9. **Διερεύνηση εφαρμογής άμυνας και σε άλλους τομείς**. Συνήθως οι περισσότερες τεχνικές άμυνας επικεντρώνονται στους τομείς επεξεργασίας εικόνας, όμως είναι σημαντικό να ελεγχθεί εάν το θεωρητικό υπόβαθρο αυτής έχει και εφαρμογή σε άλλους τομείς, για να είναι πραγματικά αποτελεσματική.

4.6 Συγκριτικές Αξιολογήσεις Ευρωστίας

Σε αυτήν την ενότητα, θα γίνει μια αναφορά των state-of-the-art τεχνικών για empirical και certified robustness σε διάφορα σενάρια, οι επιδόσεις των οποίων έχουν προκύψει από τα επικρατέστερες συγκριτικές αξιολογήσεις (**benchmarks**) στον τομέα αυτόν, για να υπάρχει και μια ποσοτική αίσθηση των καλύτερων τεχνικών και των επιδόσεων τους.

4.6.1 AutoAttack

4.6.1.1 Περιγραφή

Στον τομέα του empirical robustness, μια από τις πρώτες προσπάθειες συγκεντρωτικής δοκιμής και αξιολόγησης αμυντικών τεχνικών είναι το **AutoAttack** [113]. Σε αυτήν την εργασία, επιλέχθηκαν περισσότερα από 50 μοντέλα από δημοσιευμένα άρθρα στα κορυφαία επιστημονικά συνέδρια τεχνητής νοημοσύνης και όρασης υπολογιστών, εκτελέστηκαν ένα σύνολο διαφορετικών επιθέσεων (untargeted, targeted) σε διαφορετικά σύνολα δεδομένων και συνθήκες (ℓ_p, ϵ) κάθε φορά. Στην αξιολόγηση αυτή, για τις white-box επιθέσεις, οι οποίες εκτελέστηκαν πάνω από μια φορά, όλα τα μοντέλα απέδωσαν μικρότερο robust accuracy από το δημοσιευμένο, και μάλιστα πάνω από 13 από αυτά είχαν παραπάνω από 10% διαφορά.

Το AutoAttack μπορεί να μην αποτελεί την απόλυτη ανταγωνιστική επίθεση που είναι αρκετή για να αξιολογηθεί ένα οποιαδήποτε μοντέλο για την ευρωστία του, θεωρείται όμως η ελάχιστη δυνατή αξιολόγηση που μπορεί να γίνει σε οποιοδήποτε νέο μοντέλο ή προτεινόμενη άμυνα, καθώς έχει καλή επίδοση και μπορεί να δώσει μια καλή πρώτη αξιολόγηση του empirical robustness.

4.6.1.2 Μοντέλο Απειλής

Το σύνολο των επιθέσεων που χρησιμοποιήθηκαν αποτελούνται από μια ποικιλία white-box και black-box επιθέσεων, οι οποίες είναι αποδοτικές και ελεύθερες από παραμέτρους, κάτι που καθιστά το AutoAttack ένα αξιόπιστο, γρήγορο και αυτόματο τρόπο αξιολόγησης ευρωστίας. Πιο συγκεκριμένα, οι δύο επιθέσεις βασίζονται σε βελτιωμένη έκδοση της Projected Gradient Descent (PGD) επίθεσης [47], με όνομα Auto-PGD (**APGD**), η οποία παραλλαγή έχει βελτιωμένο step size και είναι budget-aware, δηλαδή δε γνωρίζει αν η απώλεια μειώνεται σε κάθε επανάληψη. Αυτές οι δύο επιθέσεις (APG_{CE}, $APGD_{DLR}$) είναι απλά δύο παραλλαγές της ίδιας τεχνικής απλά με διαφορετική συνάρτηση κόστους. Οι άλλες δυο white-box επιθέσεις βασίζονται πάνω στην Fast Adaptive Boundary (**FAB**) επίθεση [114], και χρησιμοποιείται μια στοχευμένη (FAB^T) και μια μη στοχευμένη έκδοση (FAB) τους. Η black-box επίθεση που χρησιμοποιείται είναι η **Square Attack** [115].

Σε αυτές τις αξιολογήσεις, τα μοντέλα δοκιμάστηκαν, σε ℓ_2 και ℓ_∞ μοντέλα επιθέσεων,

και στα MNIST, CIFAR-10, CIFAR-100, ImageNet datasets. Επίσης, γίνεται ο διαχωρισμός μεταξύ ντετερμινιστικών (ύπαρξη threshold ϵ) και randomized αμυνών (ύπαρξη στοχαστικού μέρους) στα αποτελέσματα.

Ενδεικτικά παρουσιάζουμε τα αποτελέσματα της αξιολόγησης ευρωστίας για τις *ντετερμινιστικές* τεχνικές, στις παρακάτω συνθήκες:

- CIFAR-10, ℓ_∞, $\epsilon = 8/255$
- CIFAR-10, ℓ_∞, $\epsilon = 0.031$
- CIFAR-100, ℓ_∞, $\epsilon = 8/255$
- MNIST, ℓ_∞, $\epsilon = 0.3$
- ImageNet, ℓ_∞, $\epsilon = 4/255$
- CIFAR-10, ℓ_2, $\epsilon = 0.5$

4.6.1.3 Αποτελέσματα

Στους πίνακες αποτελεσμάτων 4.4, 4.5 ξεκινώντας από την πρώτη στήλη, βλέπουμε το όνομα της δημοσίευσής για κάθε μοντέλο, το clean accuracy, τα αποτελέσματα robust accuracy για κάθε δοκιμασμένη επίθεση και τη στήλη με τα AutoAttack (AA) αποτελέσματα, που πρόκειται για ένα συνδυαστικό ποσοστό robust accuracy όλων των υπόλοιπων επιθέσεων. Οι τελευταίες δύο στήλες δείχνουν την αρχικά δημοσιευμένη robust accuracy του κάθε μοντέλου και τέλος τη διαφορά μεταξύ αυτής και της AA.

#	paper	clean	APGD_{CE}	$\text{APGD}^{\text{T}}_{\text{DLR}}$	FAB^{T}	Square	AA	reported	reduct.
CIFAR-10 - l_∞ - $\epsilon = 8/255$									
1	(Carmon et al., 2019)	89.69	61.74	**59.54**	60.12	66.63	59.53	62.5	-2.97
2	(Alayrac et al., 2019)	86.46	60.17	**56.27**	56.81	66.37	56.03	56.30	-0.27
3	(Hendrycks et al., 2019)	87.11	57.23	**54.94**	55.27	61.99	54.92	57.4	-2.48
4	(Rice et al., 2020)	85.34	57.00	**53.43**	53.83	61.37	53.42	58	-4.58
5	(Qin et al., 2019)	86.28	55.70	52.85	53.28	60.01	52.84	52.81	0.03
6	(Engstrom et al., 2019)	87.03	51.72	**49.32**	49.81	58.12	49.25	53.29	-4.04
7	(Kumari et al., 2019)	87.80	51.80	**49.15**	49.54	58.20	49.12	53.04	-3.92
8	(Mao et al., 2019)	86.21	49.65	**47.44**	47.91	56.98	47.41	50.03	-2.62
9	(Zhang et al., 2019a)	87.20	46.15	**44.85**	45.39	55.08	44.83	47.98	-3.15
10	(Madry et al., 2018)	87.14	44.75	**44.28**	44.75	53.10	44.04	47.04	-3.00
11	(Pang et al., 2020)	80.89	57.07	**43.50**	44.06	49.73	43.48	55.0	-11.52
12	(Wong et al., 2020)	83.34	45.90	**43.22**	43.74	53.32	43.21	46.06	-2.85
13	(Shafahi et al., 2019)	86.11	43.66	**41.64**	43.44	51.95	41.47	46.19	-4.72
14	(Ding et al., 2020)	84.36	50.12	**41.74**	42.47	55.53	41.44	47.18	-5.74
15	(Moosavi-Dezfooli et al., 2019)	83.11	41.72	**38.50**	38.97	47.69	38.50	41.4	-2.90
16	(Zhang & Wang, 2019)	89.98	64.42	**37.29**	38.48	59.12	36.64	60.6	-23.96
17	(Zhang & Xu, 2020)	90.25	71.40	**37.54**	38.99	66.88	36.45	68.7	-32.25
18	(Jang et al., 2019)	78.91	37.76	**34.96**	35.50	44.33	34.95	37.40	-2.45
19	(Kim & Wang, 2020)	91.51	56.64	35.93	**35.41**	61.30	34.22	57.23	-23.01
20	(Moosavi-Dezfooli et al., 2019)	80.41	36.65	**33.70**	34.08	43.46	33.70	36.3	-2.60
21	(Wang & Zhang, 2019)	92.80	59.09	33.61	**31.19**	64.22	29.35	58.6	-29.25
22	(Wang & Zhang, 2019)	92.82	69.62	29.73	**29.10**	66.77	26.93	66.9	-39.97
23	(Mustafa et al., 2019)	89.16	8.16	1.13	**0.71**	33.91	0.28	32.32	-32.04
24	(Chan et al., 2020)	93.79	2.06	**0.53**	58.13	71.43	0.26	15.5	-15.24
25	(Pang et al., 2020)	93.52	89.48	**0.00**	**0.00**	35.82	0.00	31.4	-31.40

Σχήμα 4.4: *Συγκριτικός πίνακας αξιολόγησης adversarial robustness με το AutoAttack για το CIFAR-10 dataset και διαταραχή* ($\ell_\infty, \epsilon = 8/255$) *[113].*

Οι πρώτες παρατηρήσεις που μπορούμε να κάνουμε είναι σχετικά με τις επιδόσεις των επιθέσεων μεταξύ τους. Όπως ήδη γνωρίζαμε, οι black-box επιθέσεις είναι πιο εύκολο να αντιμετωπιστούν από τις white-box και για αυτό στην Square Attack εμφανίζονται συγκριτικά καλύτερες τιμές robust accuracy, έως και 10% διαφορά με τις άλλες επιθέσεις. Η εξαίρεση στη διαφορά αυτήν είναι μόνο στα πιο απλά σύνολα δεδομένων (όπως το MNIST), στα οποία είναι πιο εύκολο να επιτευχθεί υψηλό robust accuracy σε όλες τις περιπτώσεις, οπότε τα αποτελέσματα μεταξύ των επιθέσεων είναι αρκετά υψηλά και κοντά μεταξύ τους. Στις υπόλοι-

πες white-box επιθέσεις υπάρχουν παρόμοιες επιδόσεις μεταξύ των μοντέλων, εκτός από τα τελευταία σε κάποιες περιπτώσεις, τα οποία αποτυγχάνουν πλήρως να έχουν ένα καλό robust accuracy.

Η επόμενη παρατήρηση που μπορούμε να κάνουμε είναι σχετικά με τη διαφορά μεταξύ clean και robust accuracy. Όπως φαίνεται και από τα αποτελέσματα, υπάρχει μια διαφορά άνω του 30% στις περισσότερες περιπτώσεις. Αυτό είναι ένα καθολικό φαινόμενο του adversarial robustness, όπου συνήθως η ακρίβεια ευρωστίας είναι πολύ μικρότερη από ότι η ακρίβεια στα καθαρά δείγματα. Η εξαίρεση πάλι στη διαφορά αυτήν είναι μόνο στα πιο απλά σύνολα δεδομένων (όπως το MNIST), στα οποία είναι πιο εύκολο να επιτευχθεί υψηλό robust accuracy, οπότε και να πλησιάσει το clean accuracy.

Επίσης, όπως φαίνεται και από την τελευταία στήλη του πίνακα, σε όλες σχεδόν τις περιπτώσεις η AA έχει μικρότερο robust accuracy από το δημοσιευμένο, έως και 10% σε πολλές περιπτώσεις. Αυτό δείχνει ότι πρόκειται για μια αξιολόγηση που μπορεί να προσφέρει πιο αξιόπιστα εμπειρικά νούμερα, καθώς η τιμή της καθορίζεται από διαφορετικές επιθέσεις.

CIFAR-10 - l_∞ - $\epsilon = 0.031$									
1	(Zhang et al., 2019b)	84.92	55.28	**53.10**	53.45	59.43	53.08	56.43	-3.35
2	(Atzmon et al., 2019)	81.30	79.67	41.16	**40.73**	47.99	40.22	43.17	-2.95
3	(Xiao et al., 2020)	79.28	39.99	32.34	79.28	**20.44**	18.50	52.4	-33.90
CIFAR-100 - l_∞ - $\epsilon = 8/255$									
1	(Hendrycks et al., 2019)	59.23	33.02	**28.48**	28.74	34.26	28.42	33.5	-5.08
2	(Rice et al., 2020)	53.83	20.57	**18.98**	19.24	23.57	18.95	28.1	-9.15
MNIST - l_∞ - $\epsilon = 0.3$									
1	(Zhang et al., 2020)	98.38	95.32	94.88	96.84	**93.97**	93.96	96.38	-2.42
2	(Gowal et al., 2019)	98.34	94.79	93.93	97.03	**92.88**	92.83	93.88	-1.05
3	(Zhang et al., 2019b)	99.48	93.60	93.58	94.62	**92.97**	92.81	95.60	-2.79
4	(Ding et al., 2020)	98.95	94.58	94.62	95.37	**91.42**	91.40	92.59	-1.19
5	(Atzmon et al., 2019)	99.35	99.10	94.16	95.26	**90.86**	90.85	97.35	-6.50
6	(Madry et al., 2018)	98.53	90.57	90.57	93.69	**88.56**	88.50	89.62	-1.12
7	(Jang et al., 2019)	98.47	94.05	93.56	94.74	**88.00**	87.99	94.61	-6.62
8	(Wong et al., 2020)	98.50	86.68	86.34	88.28	**83.07**	82.93	88.77	-5.84
9	(Taghanaki et al., 2019)	98.86	30.50	**0.00**	0.01	**0.00**	0.00	64.25	-64.25
ImageNet - l_∞ - $\epsilon = 4/255$									
1	(Engstrom et al., 2019)	63.4	31.0	**27.7**	28.4	46.8	27.6	33.38	-5.78
CIFAR-10 - l_2 - $\epsilon = 0.5$									
1	(Augustin et al., 2020)	91.08	74.70	**72.91**	73.18	83.10	72.91	73.27	-0.36
2	(Engstrom et al., 2019)	90.83	69.62	**69.24**	69.46	80.92	69.24	70.11	-0.87
3	(Rice et al., 2020)	88.67	68.58	**67.68**	67.97	79.01	67.68	71.6	-3.92
4	(Rony et al., 2019)	89.05	66.59	**66.44**	66.74	78.05	66.44	67.6	-1.16
5	(Ding et al., 2020)	88.02	66.21	**66.09**	66.33	76.99	66.09	66.18	-0.09

Σχήμα 4.5: *Συγκριτικός πίνακας αξιολόγησης adversarial robustness με το AutoAttack για τα CIFAR-10, CIFAR-100, MNIST, ImageNet datasets και διαταραχή* $(\ell_\infty, \epsilon = 0.031)$, $(\ell_\infty, \epsilon = 8/255)$, $(\ell_\infty, \epsilon = 0.3)$, $(\ell_\infty, \epsilon = 4/255)$, $(\ell_2, \epsilon = 0.5)$ *[113].*

Όσον αφορά τη σύγκριση μεταξύ των διαφορετικών σύνολων δεδομένων, όπως περιμέναμε στα πιο 'εύκολα' όπως το MNIST το robust accuracy είναι πολύ υψηλή (τάξης του 80%-90%), ενώ στα σύνολα δεδομένων μεγαλύτερων διαστάσεων και δειγμάτων όπως το ImageNet, το robust accuracy είναι πολύ χαμηλή (τάξης του 30%).

Όσον αφορά τη σύγκριση μεταξύ διαφορετικών διαταραχών, αν παρατηρήσουμε για το CIFAR-10 τις περιπτώσεις με $(\ell_\infty, \epsilon = 8/255)$ και $(\ell_2, \epsilon = 0.5)$, θα δούμε ότι στη 2η περίπτωση έχουμε πολύ καλύτερες τιμές robust accuracy. Πιο συγκεκριμένα, για το μοντέλο του άρθρου *Ding et al. 2020* στην πρώτη περίπτωση (νούμερο 14 στον πίνακα 4.4) έχει τιμή AA 41.44%, ενώ στη δεύτερη περίπτωση (νούμερο 5 στον πίνακα 4.5) έχει τιμή AA 66.09%, μία διαφορά της τάξης του 25%.

Αυτοί οι βαθμολογικοί πίνακες με τα μοντέλα με τις καλύτερες επιδόσεις δε διατηρούνται πλέον από τους αρχικούς συγγραφείς, όμως η πιο ενημερωμένη κατάταξη που χρησιμοποιεί παρόμοιο σύστημα αξιολόγησης και διατηρείται ακόμα είναι το RobustBench 4.6.2 που αναλύουμε παρακάτω.

4.6.2 RobustBench

4.6.2.1 Περιγραφή

Στον τομέα του empirical robustness, το **RobustBench** [98] αποτελεί ουσιαστικά τη συνέχεια του AutoAttack, καθώς χρησιμοποιεί τις ίδιες επιθέσεις για να αξιολογήσει το robustness των μοντέλων, επεκτείνοντας όμως τα σύνολα δεδομένων και μοντέλα απειλής που χρησιμοποιεί και αξιολογώντας πάνω από 120 μοντέλα έως αυτήν τη στιγμή. Ο στόχος είναι να παραμείνει μια ενημερωμένη λίστα (υπό την μορφή leaderboard) με τις καλύτερες τεχνικές και μοντέλα από θέμα robust accuracy και για αυτόν τον λόγο, οι συγγραφείς έχουν μετατρέψει το συγκεκριμένο benchmark σε μία βιβλιοθήκη που είναι ανοιχτού κώδικα[1] και μπορεί ο καθένας να χρησιμοποιήσει για να αξιολογήσει το μοντέλου του, και ενθαρρύνεται μάλιστα η ανεξάρτητη αξιολόγηση και δημοσίευση των αποτελεσμάτων για τον εμπλουτισμό των leaderboards.

Στο RobustBench έχουν γίνει επεκτάσεις πάνω στο AutoAttack, χρησιμοποιώντας ισχυρότερες white-box και black-box επιθέσεις για να γίνει και μια worst-case αξιολόγηση και ελέγχοντας αν μια αξιολόγηση είναι ρεαλιστική ή όχι και ανάλογα χρησιμοποιώντας adaptive επιθέσεις για την καλύτερη αξιολόγηση. Επίσης, έχει γίνει επέκταση και στα είδη διαταραχών που χρησιμοποιούνται, προσθέτοντας πέρα από τα ℓ_∞, ℓ_2 και Common Image Corruptions [116], το οποίο αντιπροσωπεύει ένα μεγάλο σημαντικό είδος διαταραχών, οι οποίες όπως και με τα adversarial perturbations δε θα έπρεπε να αλλάζουν τις αποφάσεις των μοντέλων.

4.6.2.2 Μοντέλο Απειλής

Για την αξιολόγηση των μοντέλων χρησιμοποιούνται κυρίως οι 4 επιθέσεις που είδαμε και στο AutoAttack, λόγω τις ποικιλίας τους και της ευκολίας χρήση (δε χρειάζεται ρύθμιση των υπερπαραμέτρων).

Για τη χρήση του συγκεκριμένου benchmark πρέπει τα μοντέλα ή οι αμυντικές τεχνικές να ικανοποιούν κάποιες συνθήκες για να μπορέσει να βγει ένα αξιόλογο συμπέρασμα. Αυτές είναι ότι τα μοντέλα: (i) δεν πρέπει να έχουν μηδενικές κλίσεις σε σχέση με τις εισόδους, καθώς οι περισσότερες επιθέσεις είναι gradient-based, (ii) πρέπει να έχουν ένα πλήρως ντετερμινιστικό forward pass, δηλαδή να μην περιέχουν στοχαστικά στοιχεία, καθώς ενώ μπορεί να αυξάνουν την ευρωστία κάνουν την τυποποίηση της αξιολόγησης πιο δύσκολη, (iii) δεν πρέπει να έχουν βρόχο βελτιστοποίησης, καθώς κάνουν τη διαδικασία του backpropagation εξαιρετικά ακριβή, άρα και την αξιολόγηση. Συχνά, οι άμυνες που παραβιάζουν αυτές τις 3 αρχές κάνουν μόνο πιο δύσκολες τις επιθέσεις με κλίση (gradient-based), αλλά δε βελτιώνουν ουσιαστικά την ευρωστία [78] εκτός από εκείνες που μπορούν να παρουσιάσουν certified robustness [79].

Σε αυτές τις αξιολογήσεις, τα μοντέλα δοκιμάζονται όπως αναφέραμε, σε ℓ_∞, ℓ_2 και Common Image Corruptions (Common) μοντέλα απειλών και στα CIFAR-10, CIFAR-100, ImageNet datasets.

[1] https://github.com/RobustBench/robustbench

Ενδεικτικά παρουσιάζουμε τα αποτελέσματα της αξιολόγησης ευρωστίας στις παρακάτω συνθήκες:

- CIFAR-10, ℓ_∞, $\epsilon = 8/255$
- CIFAR-10, ℓ_2, $\epsilon = 0.5$
- CIFAR-100, ℓ_∞, $\epsilon = 8/255$
- ImageNet, ℓ_∞, $\epsilon = 4/255$

4.6.2.3 Αποτελέσματα

Στους πίνακες αποτελεσμάτων 4.6, 4.7 ξεκινώντας από την πρώτη στήλη, βλέπουμε τον αύξων αριθμό κατάταξης στον πίνακα, το όνομα της δημοσίευσής για κάθε μοντέλο, το clean accuracy, τα αποτελέσματα robust accuracy που προκύπτει συνδυαστικά από όλες τις επιθέσεις, αν χρησιμοποιήθηκαν επιπλέον δεδομένα για την εύρωστη εκπαίδευση του μοντέλου, η αρχιτεκτονική του μοντέλου, το συνέδριο στο οποίο δημοσιεύτηκε πρώτη φορά, και ένα τέλος ένα ID για το κάθε εκπαιδευμένο μοντέλο που υπάρχει διαθέσιμο στο Model Zoo[2] το οποίο είναι επίσης διαθέσιμο, για την εύκολη χρήση προεκπαιδευμένων robust μοντέλων.

Leaderboard for the ℓ_∞-threat model, CIFAR-10.

	Model	Clean	Robust	Extra data	Architecture	Venue	Model Zoo ID
1	Rebuffi et al. [111]	92.23	66.56	Y	WRN-70-16	arXiv, Mar 2021	Rebuffi2021Fixing_70_16_cutmix_extra
2	Gowal et al. [50]	91.10	65.87	Y	WRN-70-16	arXiv, Oct 2020	Gowal2020Uncovering_70_16_extra
3	Rebuffi et al. [111]	88.50	64.58	N	WRN-106-16	arXiv, Mar 2021	Rebuffi2021Fixing_106_16_cutmix_ddpm
4	Rebuffi et al. [111]	88.54	64.20	N	WRN-70-16	arXiv, Mar 2021	Rebuffi2021Fixing_70_16_cutmix_ddpm
5	Rade and Moosavi-Dezfooli [107]	91.47	62.83	Y	WRN-34-10	OpenReview, Jun 2021	Rade2021Helper_extra
6	Gowal et al. [50]	89.48	62.76	Y	WRN-28-10	arXiv, Oct 2020	Gowal2020Uncovering_28_10_extra
7	Rade and Moosavi-Dezfooli [107]	88.16	60.97	N	WRN-28-10	OpenReview, Jun 2021	Rade2021Helper_ddpm
8	Rebuffi et al. [111]	87.33	60.73	N	WRN-28-10	arXiv, Mar 2021	Rebuffi2021Fixing_28_10_cutmix_ddpm
9	Wu et al. [154]	87.67	60.65	Y	WRN-34-15	arXiv, Oct 2020	N/A
10	Sridhar et al. [133]	86.53	60.41	Y	WRN-34-15	arXiv, Jun 2021	Sridhar2021Robust_34_15

Σχήμα 4.6: *Συγκριτικός πίνακας αξιολόγησης adversarial robustness με τα 10 καλύτερα μοντέλα από το RobustBench για το CIFAR-10 dataset και διαταραχή* ($\ell_\infty, \epsilon = 8/255$) *[98].*

Αρχικά βλέπουμε ότι τα συμπεράσματα που βγάλαμε και στην περίπτωση του AutoAttack ισχύουν και εδώ πέρα. Η διαφορά μεταξύ clean και robust accuracy είναι εξίσου μεγάλη, της τάξεως του 20% με 30%, ανάλογα με το σύνολο δεδομένων και τη διαταραχή. Επίσης, πάλι στα πιο 'εύκολα' σύνολα δεδομένων παρουσιάζεται μεγαλύτερο robust accuracy, π.χ. για το μοντέλο του άρθρου *Gowal et al.* στο CIFAR-10 (νούμερο 2 στον πίνακα 4.6) έχει τιμή robust accuracy 66.56%, ενώ στο CIFAR-100 (νούμερο 1 στον πίνακα 4.7) έχει τιμή 36.88%, μία διαφορά της τάξης του 30%. Όπως επίσης, τα ίδια συμπεράσματα ισχύουν και για τη διαφορά μεταξύ των διαταραχών ανά μοντέλο απειλής.

Έχει ιδιαίτερο ενδιαφέρον να αναφέρουμε τις **καλύτερες επιδόσεις empirical robustness** που έχουν παρουσιαστεί στο RobustBench από κάθε σύνολο δεδομένων και σε κάθε μοντέλο απειλής που εξετάζουμε, έως τη δεδομένη χρονική στιγμή, και άρα σύμφωνα με τις τελευταίες προσθήκες που έχουν γίνει στην ιστοσελίδα τους[3]:

- CIFAR-10, ℓ_∞, $\epsilon = 8/255$: Robust Accuracy 71.07%
- CIFAR-10, ℓ_2, $\epsilon = 0.5$: Robust Accuracy 84.97%
- CIFAR-100, ℓ_∞, $\epsilon = 8/255$: Robust Accuracy 42.67%
- ImageNet, ℓ_∞, $\epsilon = 4/255$: Robust Accuracy 59.56%

[2] https://github.com/RobustBench/robustbench/tree/master/robustbench/model_zoo
[3] https://robustbench.github.io

Leaderboard for the ℓ_2-threat model, CIFAR-10.

	Model	Clean	Robust	Extra data	Architecture	Venue	Model Zoo ID
1	Rebuffi et al. [111]	95.74	82.32	Y	WRN-70-16	arXiv, Mar 2021	Rebuffi2021Fixing_70_16_cutmix_extra
2	Gowal et al. [50]	94.74	80.53	Y	WRN-70-16	arXiv, Oct 2020	Gowal2020Uncovering_extra
3	Rebuffi et al. [111]	92.41	80.42	N	WRN-70-16	arXiv, Mar 2021	Rebuffi2021Fixing_70_16_cutmix_ddpm
4	Rebuffi et al. [111]	91.79	78.80	N	WRN-28-10	arXiv, Mar 2021	Rebuffi2021Fixing_28_10_cutmix_ddpm
5	Augustin et al. [7]	93.96	78.79	Y	WRN-34-10	ECCV 2020	Augustin2020Adversarial_34_10_extra
6	Augustin et al. [7]	92.23	76.25	Y	WRN-34-10	ECCV 2020	Augustin2020Adversarial_34_10
7	Rade and Moosavi-Dezfooli [107]	90.57	76.15	N	PreActRN-18	OpenReview, Jun 2021	Rade2021Helper_R18_ddpm
8	Sehwag et al. [123]	90.31	76.12	N	WRN-34-10	arXiv, Apr 2021	Sehwag2021Proxy
9	Rebuffi et al. [111]	90.33	75.86	N	PreActRN-18	arXiv, Mar 2021	Rebuffi2021Fixing_R18_cutmix_ddpm
10	Gowal et al. [50]	90.90	74.50	N	WRN-70-16	arXiv, Oct 2020	Gowal2020Uncovering

Leaderboard for the ℓ_∞-threat model, CIFAR-100.

	Model	Clean	Robust	Extra data	Architecture	Venue	Model Zoo ID
1	Gowal et al. [50]	69.15	36.88	Y	WRN-70-16	arXiv, Oct 2020	Gowal2020Uncovering_extra
2	Rebuffi et al. [111]	63.56	34.64	N	WRN-70-16	arXiv, Mar 2021	Rebuffi2021Fixing_70_16_cutmix_ddpm
3	Rebuffi et al. [111]	62.41	32.06	N	WRN-28-10	arXiv, Mar 2021	Rebuffi2021Fixing_28_10_cutmix_ddpm
4	Cui et al. [30]	62.55	30.20	N	WRN-34-20	ICCV 2021	Cui2020Learnable_34_20_LBGAT6
5	Gowal et al. [50]	60.86	30.03	N	WRN-70-16	arXiv, Oct 2020	Gowal2020Uncovering
6	Cui et al. [30]	60.64	29.33	N	WRN-34-10	ICCV 2021	Cui2020Learnable_34_10_LBGAT6
7	Rade and Moosavi-Dezfooli [107]	61.50	28.88	N	PreActRN-18	OpenReview, Jun 2021	Rade2021Helper_R18_ddpm
8	Wu et al. [155]	60.38	28.86	N	WRN-34-10	NeurIPS 2020	Wu2020Adversarial
9	Rebuffi et al. [111]	56.87	28.50	N	PreActRN-18	arXiv, Mar 2021	Rebuffi2021Fixing_R18_ddpm
10	Hendrycks et al. [60]	59.23	28.42	Y	WRN-28-10	ICML 2019	Hendrycks2019Using

Leaderboard for the ℓ_∞-threat model, ImageNet.

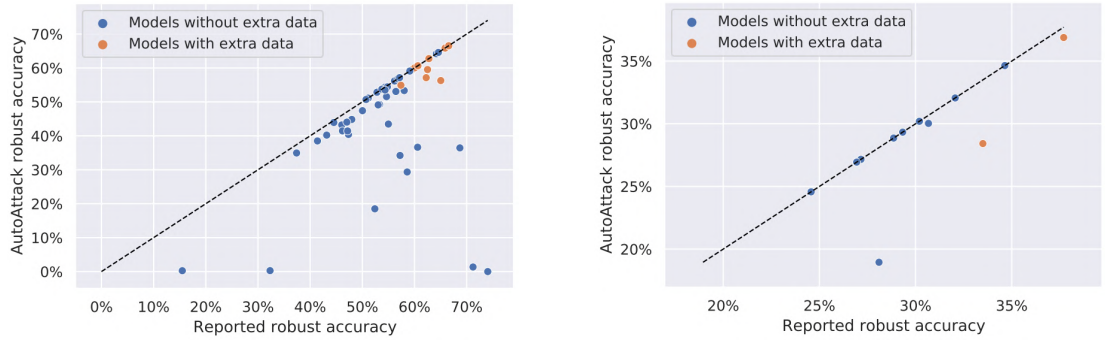
	Model	Clean	Robust	Extra data	Architecture	Venue	Model Zoo ID
1	Salman et al. [117]	68.46	38.14	N	WRN-50-2	NeurIPS 2020	Salman2020Do_50_2
2	Salman et al. [117]	64.02	34.96	N	RN-50	NeurIPS 2020	Salman2020Do_R50
3	Engstrom et al. [37]	62.56	29.22	N	RN-50	GitHub, Oct 2019	Engstrom2019Robustness
4	Wong et al. [153]	55.62	26.24	N	RN-18	ICLR 2020	Wong2020Fast
5	Salman et al. [117]	52.92	25.32	N	RN-50	NeurIPS 2020	Salman2020Do_R18
6	Standard_R50	76.52	0.0	N	RN-50	N/A	Standard_R50

Σχήμα 4.7: *Συγκριτικός πίνακας αξιολόγησης adversarial robustness με τα έως 10 καλύτερα μοντέλα από το RobustBench για τα CIFAR-10, CIFAR-100, ImageNet datasets και διαταραχή* $(\ell_2, \epsilon = 0.5)$, $(\ell_\infty, \epsilon = 8/255)$, $(\ell_\infty, \epsilon = 4/255)$ *[98].*

Από όλα τα συγκεντρωμένα αποτελέσματα που έχουν παραχθεί από το RobustBench, έχουν δημιουργηθεί και κάποια συγκριτικά διαγράμματα που παρουσιάζουν αρκετό ενδιαφέρον. Στα διαγράμματα 4.8 παρουσιάζεται το δημοσιευμένο robust accuracy σε σχέση με την υπολογισμένη από το AutoAttack, στο 4.8α′ για όλα τα αποτελέσματα από το CIFAR-10 και με $(\ell_\infty, \epsilon = 8/255)$ και στο 4.8β′ για όλα τα αποτελέσματα από το CIFAR-100 και με $(\ell_\infty, \epsilon = 8/255)$. Στα διαγράμματα διαχωρίζονται με πορτοκαλί χρώμα τα μοντέλα που χρειάστηκαν επιπλέον δεδομένα για την εύρωστη εκπαίδευση ενώ με μπλε χρώμα αυτά που δε χρειάστηκαν. Η διαγώνια διακεκομμένη γραμμή δείχνει όλα τα στοιχεία που έχουν ακριβώς το ίδιο ποσοστό και για τη δημοσιευμένη και για το υπολογισμένο robust accuracy.

Στο CIFAR-10, τα περισσότερα μοντέλα φαίνεται ότι είναι τοποθετημένα κοντά στις δύο τιμές, αλλά όλα βρίσκονται κάτω από τη γραμμή, δηλαδή η δημοσιευμένη τιμή είναι μεγαλύτερη από την υπολογισμένη από το AutoAttack, άρα και υπερεκτιμημένη. Επίσης, βλέπουμε και κάποια μοντέλα τα οποία έχουν πολύ μεγάλη διαφορά στις τιμές αυτές, είτε γιατί η AutoAttack δεν μπορούσε να ολοκληρωθεί σωστά είτε γιατί οι άμυνες που αξιολογήθηκαν είχαν σοβαρά σφάλματα. Στο CIFAR-100 ενώ ισχύουν τα ίδια δεν είναι τόσο έντονες οι διαφορές, δηλαδή δεν υπάρχει τόσο μεγάλη υπερεκτίμηση και κυρίως δεν υπάρχει κανένα με μηδενικό robust accuracy. Και στις δύο περιπτώσεις όμως, φαίνεται ότι όσα μοντέλα έχουν χρησιμοποιήσει περισσότερα δεδομένα έχουν πολύ μεγαλύτερες τιμές robust accuracy και πλησιάζουν πιο πολύ τις δημοσιευμένες.

Στα διαγράμματα 4.9 παρουσιάζεται η clear (ή standard) accuracy σε σχέση με την robust accuracy υπολογισμένη από το AutoAttack, στο 4.9α′ για τις ίδιες συνθήκες με τα

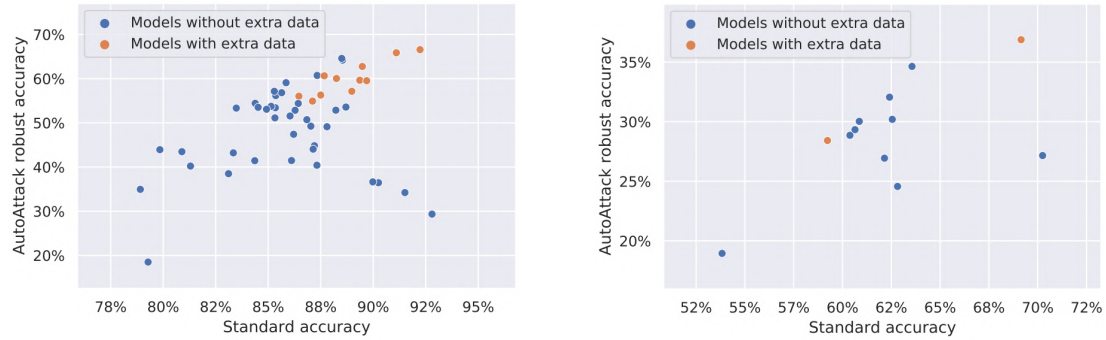


(α′) *CIFAR-10 και* $(\ell_\infty, \epsilon = 8/255)$ (β′) *CIFAR-100 και* $(\ell_\infty, \epsilon = 8/255)$

Σχήμα 4.8: *Διάγραμμα robust accuracy (x-άξονας) σύμφωνα με το AutoAttack και δημοσιευμένου accuracy (y-άξονας) [98].*

προηγούμενα.

Εδώ παρατηρούμε αυτό που ήδη έχουμε αναφέρει και αναλύουμε περισσότερα στην παρακάτω ενότητα 4.7 για τους συμβιβασμούς ευρωστίας, ότι το robust accuracy του μοντέλου μειώνεται σε σχέση με την κανονική του ακρίβεια. Στο CIFAR-10 είναι πολύ πιο έντονη αυτή η διαφορά, και φαίνεται ότι τα περισσότερα δεδομένα βοηθάνε στο να καλυφθεί αυτό το κενό. Στο CIFAR-100 υπάρχουν πολύ λιγότερα δείγματα, αλλά επίσης παρατηρούμε το ίδιο μεγάλο κενό μεταξύ των δύο αυτών τιμών.



(α′) *CIFAR-10 και* $(\ell_\infty, \epsilon = 8/255)$ (β′) *CIFAR-100 και* $(\ell_\infty, \epsilon = 8/255)$

Σχήμα 4.9: *Διάγραμμα robust accuracy (x-άξονας) σύμφωνα με το AutoAttack και standard accuracy (y-άξονας) [98].*

Τα παραπάνω leaderboards με τις καλύτερες επιδόσεις που αναφέραμε, αφορούν τα δημοσιευμένα αποτελέσματα, με τελευταία ανανέωση το 2021. Στην ιστοσελίδα του RobustBench[4]*, βρίσκονται οι ενημερωμένοι πίνακες κατάταξης με νέα μοντέλα από το 2022 και 2023 που έχουν αυξήσει τις προηγούμενες μέγιστες τιμές robust accuracy.*

4.6.3 SoK: Certified Robustness for Deep Neural Networks

4.6.3.1 Περιγραφή

Στον τομέα του certified robustness, το SoK: Certified Robustness for Deep Neural Networks (ή για συντομία SoK Benchmark) [102], όπως αναφέραμε και στην παραπάνω ενότητα

[4] https://robustbench.github.io

4.3.3, πρόκειται για ένα benchmark το οποίο αξιολογεί μοντέλα και αμυντικές τεχνικές ως προς την επιβεβαιωμένη ευρωστία τους. Το SoK Benchmark πρόκειται για ένα Toolbox (πλέον VeriGauge[5]), ανοιχτού κώδικα, που στοχεύει στο να παρέχει δίκαιες συγκρίσεις για certified robustness αλλά και robust verification προσεγγίσεις.

Ξεχωριστά από το benchmark, υπάρχει και ένα leaderboard, με τις state-of-the-art δημοσιευμένες τεχνικές και μοντέλα με certified robustness. Το SoK Leaderboard αντικατοπτρίζει κυρίως την πρόοδο που έχει επιτευχθεί στα μοντέλα με επιβεβαιωμένη ευρωστία, με στόχο να διατηρηθεί και να επεκταθεί με νέα μοντέλα που θα εμφανιστούν στο μέλλον με καλύτερες επιδόσεις από τα υπάρχοντα. Για αυτό και δίνεται η δυνατότητα στον καθένα που δημοσιεύει ένα νέο μοντέλο η τεχνική και μπορεί να αξιολογήσει το certified robustness αυτού, να δημοσιεύσει τα αποτελέσματα για τον εμπλουτισμό της ιστοσελίδας των leaderboards[6].

4.6.3.2 Μοντέλο Απειλής

Σχετικά με το **benchmark**, για την επαλήθευση της ευρωστίας, οι τεχνικές χωρίζονται σε ντετερμινιστικές (**Deterministic**) προσεγγίσεις, όπου συγκρίνεται η certified robustness ακρίβεια τους πάνω σε διαφορετικά μοντέλα και κλίμακες, και σε πιθανολογικές (**Probabilistic**) προσεγγίσεις, όπου συγκρίνεται η καλύτερη certified robustness ακρίβεια που επιτυγχάνουν από κοινού. Για τις πιθανολογικές προσεγγίσεις, συμπεριλαμβάνονται υπόψιν μόνο όσες έχουν certification confidence $\geq 99.9\%$.

Για την αξιολόγηση του certified robustness, χρησιμοποιείται η certified accuracy, η οποία υπολογίζεται ως το κλάσμα των δειγμάτων δοκιμής που είναι επιβεβαιωμένα εύρωστο εναντίων των ορισμένων (ℓ_p, ϵ) διαταραχών:

$$certified\ accuracy = \frac{\#\ of\ samples\ verified\ to\ be\ robust}{\#\ of\ all\ evaluated\ samples}$$

Για την αξιολόγηση των Deterministic Verification προσεγγίσεων, χρησιμοποιούνται 17 διαφορετικές τεχνικές Complete και Incomplete επιβεβαίωσης, 3 πλήρως συνδεδεμένα νευρωνικά δίκτυα (FCNNa, FCNNb, FCNNc) και 4 συνελικτικά νευρωνικά δίκτυα (CNNa, CNNb, CNNc, CNNd), 2 datasets (MNIST, CIFAR-10) και επιθέσεις ℓ_∞.

Για την αξιολόγηση των Probabilistic Verification προσεγγίσεων, χρησιμοποιούνται 4 Verification και 5 Robust Training προσεγγίσεις, 3 ResNet μοντέλα (ResNet-110, Wide ResNet 40-2, ResNet-50), 2 datasets (CIFAR-10, ImageNet) και επιθέσεις ℓ_∞, ℓ_1, ℓ_2.

Σχετικά με το **leaderboard**, παρουσιάζουμε ενδεικτικά τις καλύτερες τεχνικές και μοντέλα από θέμα certified robustness σε διάφορα σύνολα δεδομένων (MNIST, CIFAR-10, ImageNet) και είδη διαταραχών (ℓ_∞, ℓ_1, ℓ_2 για διάφορες τιμές ϵ threshold). Πιο συγκεκριμένα παρουσιάζουμε τα αποτελέσματα των leaderboards στις παρακάτω συνθήκες (επιλεγμένες ώστε να είναι κοντά στις επιλεγμένες για toRobustBench):

- CIFAR-10, ℓ_∞, $\epsilon = 8/255$
- CIFAR-10, ℓ_2, $\epsilon = 0.25$
- ImageNet, ℓ_∞, $\epsilon = 1/255$

[5] https://github.com/AI-secure/VeriGauge
[6] https://sokcertifiedrobustness.github.io

4.6.3.3 Αποτελέσματα

Από τα αποτελέσματα του **benchmark**, στην περίπτωση των Deterministic Verification προσεγγίσεων (βλ. πίνακα αποτελεσμάτων 4.10), προκύπτει αρχικά ότι οι στα μικρά μοντέλα (π.χ. FCNNAa, FCNNb, οι complete verification προσεγγίσεις μπορούν να επαληθεύσουν αποτελεσματικά την ευρωστία, επομένως αποτελούν την καλύτερη επιλογή. Σε μεγαλύτερα όμως μοντέλα (π.χ. CNNb, CNNc, CNNd), αποδίδουν σχεδόν μηδενική certified ακρίβεια, οπότε οι incomplete verification προσεγγίσεις προσφέρονται για καλύτερα αποτελέσματα, όπως για παράδειγμα τεχνικές linear relaxation. Τεχνικές που βασίζονται στη SDP (Semidefinite Programming) [117] μέθοδο παίρνουν πάρα πολύ χρόνο και δεν καταφέρνουν να είναι πρακτικά χρήσιμες σε κανένα μοντέλο.

Στην περίπτωση των Probabilistic Verification προσεγγίσεων (βλ. πίνακα αποτελεσμάτων 4.11), προκύπτει ότι το adversarial training, επιτυγχάνει τις καλύτερες επιδόσεις certified ακρίβειας, μεταξύ των robust training τεχνικών. Επίσης, για ℓ_1 και ℓ_2 επιθέσεις, οι προσεγγίσεις που βασίζονται στη Neyman-Pearson (Randomized Smoothing) [79] μέθοδο, επιτυγχάνουν τις υψηλότερες τιμές certified robustness.

Verification Approach					FCNNA		FCNNB		FCNNC		CNNA		CNNB		CNNC		CNND	
Category				Name	adv	cadv	adv	cadv	adv	cadv	adv	cadv	adv	cadv	adv	cadv	adv	cadv
Complete	Solver-Based			BOUNDED MILP [46]	**19%**	**27%**	1%	**25%**	0%	0%	0%	**34%**	0%	**36%**	0%	0%	0%	0%
	Branch-and-Bound			AI^2 [37]	**19%**	**27%**	**7%**	23%	0%	22%	**8%**	**34%**	0%	20%	0%	14%	0%	0%
Incomplete	Linear Relaxtion	Linear Programming		LP-FULL [39], [40]	15%	**27%**	6%	**25%**	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Linear Inequality	Interval	IBP [61]	0%	**27%**	0%	**25%**	0%	**30%**	0%	**34%**	0%	35%	0%	**38%**	0%	**28%**
			Polyhedra	FAST-LIN [39]	15%	25%	4%	18%	0%	19%	3%	26%	0%	15%	0%	7%	0%	0%
				CROWN [38]	15%	**27%**	6%	20%	0%	22%	**8%**	33%	**1%**	20%	0%	0%	0%	0%
				CNN-CERT [101]	15%	**27%**	5%	20%	0%	0%	7%	33%	0%	20%	0%	0%	0%	0%
				CROWN-IBP [112]	9%	**27%**	0%	22%	0%	28%	0%	**34%**	0%	31%	0%	32%	0%	25%
				DEEPPOLY [64]	15%	**27%**	6%	20%	0%	22%	**8%**	33%	**1%**	20%	0%	7%	0%	0%
				REFINEZONO [25]	0%	**27%**	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
			Duality	WK [27], [71]	15%	25%	4%	18%	0%	19%	3%	26%	0%	15%	0%	7%	0%	5%
		Multi-Neuron Relaxation		K-RELU [24]	15%	**27%**	2%	23%	0%	0%	0%	32%	0%	0%	0%	0%	0%	0%
	SDP			SDPVERIFY [79]	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
				LMIVERIFY [77]	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Lipschitz	General Lipschitz		OP-NORM [11], [83]	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
				FASTLIP [39]	12%	**27%**	0%	17%	0%	17%	0%	24%	0%	0%	0%	0%	0%	0%
				RECURJAC [84]	14%	**27%**	2%	17%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Accuracy under PGD (Upper Bound of Robust Accuracy)					22%	28%	23%	26%	19%	34%	34%	34%	33%	39%	36%	40%	41%	31%
Clean Accuracy					33%	31%	37%	30%	26%	39%	44%	46%	53%	48%	52%	46%	66%	46%

Σχήμα 4.10: *Αποτελέσματα SoK benchmark για Deterministic Verification προσεγγίσεις στο CIFAR-10 dataset και διαταραχές* ($\ell_\infty, \epsilon = 8/255$) *[102]. Με adv σημειώνεται το empirical robust accuracy και με cadv το certified robust accuracy. Με* ***bold*** *σημειώνονται οι καλύτερες επιδόσεις σε κάθε κατηγορία.*

Adversary	Model Structure	Verification Approach	Robust Training	Smooth Dist.	Certified Robust Accuracy under Perturbation Radius ϵ					
				$\epsilon =$	0.25	0.50	0.75	1.00	1.25	1.50
ℓ_2	Wide ResNet 40-2	Differential Privacy Based [92]	Data Augmentation [22], [28]	Gaussian	34.2%	14.8%	6.8%	2.2%	0.0%	0.0%
		Neyman-Pearson [22], [28], [95], [96]			**68.8%**	**46.8%**	**36.0%**	**25.4%**	**19.8%**	**15.6%**
		f-Divergence [94]			62.2%	41.8%	27.2%	19.2%	14.2%	11.4%
	ResNet-110	Neyman-Pearson [22], [28], [95], [96]	Data Augmentation [22], [28]		61.2%	43.2%	32.0%	22.4%	17.2%	14.0%
			Adversarial Training [98]		73.0%	57.8%	48.2%	**37.2%**	33.6%	**28.2%**
			Adversarial + Pretraining [98], [144]		**81.8%**	**62.6%**	**52.4%**	37.2%	**34.0%**	30.2%
			MACER [149]		68.8%	52.6%	40.4%	33.0%	27.8%	25.0%
			ADRE [227]		68.0%	50.2%	37.8%	30.2%	23.0%	17.0%
				$\epsilon =$	0.5	1.0	1.5	2.0	3.0	4.0
ℓ_1	Wide ResNet 40-2	Differential Privacy Based [92]	Data Augmentation [22], [28]	Laplace	43.0%	20.8%	12.2%	7.2%	1.4%	0.0%
		Rényi Divergence [93]			58.2%	39.4%	27.0%	16.8%	9.2%	4.0%
		Neyman-Pearson [22], [28], [95], [96]			58.4%	39.6%	27.0%	17.2%	9.2%	4.2%
				Uniform	**69.2%**	**56.6%**	**48.0%**	**39.4%**	**26.0%**	**20.4%**
				$\epsilon =$	1/255	2/255	4/255	8/255		
ℓ_∞	Wide ResNet 40-2	Neyman-Pearson [22], [28], [95], [96]	Data Augmentation [22], [28]	Gaussian	71.4%	52.0%	29.0%	12.8%		
			Adversarial Training [98]		**83.2%**	**65.0%**	**49.6%**	**25.4%**		

Σχήμα 4.11: *Αποτελέσματα SoK benchmark για Probabilistic Verification προσεγγίσεις στο CIFAR-10 dataset και διαταραχές* $\ell_\infty, \ell_1, \ell_2$ *για διάφορες τιμές* ϵ *[102]. Με* ***bold*** *σημειώνονται οι καλύτερες επιδόσεις σε κάθε κατηγορία.*

Από τα αποτελέσματα του **leaderboard**, όπου φαίνονται οι καλύτερες δημοσιευμένες τεχνικές για διάφορες συνθήκες, μπορούμε να βγάλουμε κάποια συμπεράσματα για το πεδίο αυτό

και την εξέλιξη του. Στην περίπτωση των ℓ_∞ επιθέσεων (βλ. πίνακα αποτελεσμάτων 4.12), για το MNIST στην $\epsilon = 0.1$ περίπτωση, η καλύτερη μέθοδος αγγίζει certified robustness ακρίβεια πάνω από 98%, αποδεικνύοντας την καλή απόδοση των deterministic verification προσεγγίσεων στα μικρά σύνολα δεδομένων. Στα μεγαλύτερα σύνολα δεδομένων όμως, και πιο συγκεκριμένα στα CIFAR-10 με $\epsilon = 8/255$ και ImageNet με $\epsilon = 1/255$, η ακρίβεια πέφτει απότομα κοντά στο 40% και στις δύο περιπτώσεις.

Στην περίπτωση των ℓ_2 επιθέσεων, (βλ. πίνακα αποτελεσμάτων 4.13), τα αποτελέσματα έχουν καλύτερες επιδόσεις από ότι στην ℓ_∞ περίπτωση. Πιο συγκεκριμένα, για στο CIFAR-10 με $\epsilon = 0.25$ η ακρίβεια αγγίζει το 81%. Στο ImageNet με $\epsilon = 1.0$ και με $\epsilon = 2.0$, οι ακρίβειες αγγίζουν το 67% και 42% αντίστοιχα, μόνο με probabilistic verification προσεγγίσεις (όπως και στην ℓ_∞ περίπτωση).

Όπως αναφέραμε ήδη και παρατηρούμε από τα αποτελέσματα, αυτήν τη στιγμή, οι ντετερμινιστικές τεχνικές με αυστηρές προσεγγίσεις, δεν μπορούν να λειτουργήσουν για σύνολα δεδομένων μεγαλύτερα του CIFAR-10. Για το ImageNet σχεδόν όλες οι προσεγγίσεις είναι πιθανολογικές.

Έχει ιδιαίτερο ενδιαφέρον να αναφέρουμε αναλυτικά τις **καλύτερες επιδόσεις certified robustness** που έχουν παρουσιαστεί στο SoK Leaderboard από κάθε σύνολο δεδομένων και σε κάθε μοντέλο απειλής που εξετάζουμε, έως τη δεδομένη χρονική στιγμή, και άρα σύμφωνα με τις τελευταίες προσθήκες που έχουν γίνει στην ιστοσελίδα τους[7]:

- MNIST, ℓ_∞, $\epsilon = 0.1$: Certified Accuracy 98.22% (Determinstc)
- CIFAR-10, ℓ_∞, $\epsilon = 8/255$: Certified Accuracy 41.78% (Determinstc)
- CIFAR-10, ℓ_2, $\epsilon = 0.25$: Certified Accuracy 81.00% (Probabilistic)
- ImageNet, ℓ_∞, $\epsilon = 1/255$: Certified Accuracy 38.20% (Probabilistic)

Τα παραπάνω leaderboards με τις καλύτερες επιδόσεις που αναφέραμε, αφορούν τα δημοσιευμένα αποτελέσματα, με τελευταία ανανέωση τον Απρίλιο του 2023. Στην ιστοσελίδα του SoK, βρίσκονται οι ενημερωμένοι πίνακες κατάταξης με νέα μοντέλα που έχουν αυξήσει τις προηγούμενες μέγιστες τιμές robust accuracy.

Από όλα τα συγκεντρωμένα αποτελέσματα που έχουν παραχθεί από το SoK Leaderboard για τις περιπτώσεις που εξετάζουμε, μπορούμε να φτιάξουμε ένα διάγραμμα το οποίο δείχνει την εξέλιξη των δημοσιευμένων μοντέλων και τεχνικών για certified robustness ανά τα χρόνια (βλ. 4.14). Πιο συγκεκριμένα για τις 4 περιπτώσεις συνόλων δεδομένων και μοντέλα απειλών που εξετάζουμε, βλέπουμε σε κάθε κύκλο μια δημοσιευμένη μέθοδο και συνολικά σε όλη τη γραμμή, την πορεία της ακρίβειας ανά τα χρόνια έως τη δεδομένη χρονική στιγμή, και άρα σύμφωνα με τις τελευταίες προσθήκες που έχουν γίνει στην ιστοσελίδα.

Όπως έχουμε ήδη σχολιάσει, φαίνεται ότι στην περίπτωση του MNIST υπάρχει ανοδική πορεία, αλλά έχει ήδη ξεκινήσει με πολύ μεγάλες τιμές ακρίβειας. Για το CIFAR-10 με ℓ_∞, παρουσιάζει επίσης ανοδική πορεία με σχεδόν 20% συνολική αύξηση ανά τα χρόνια. Στην περίπτωση του CIFAR-10 με ℓ_2, ενώ δεν είναι τόσο ξεκάθαρα ανοδική η πορεία, παρουσιάζει διαχρονικά πολύ υψηλότερες τιμές ακρίβειας από την ℓ_∞ περίπτωση. Τέλος, για το ImageNet δεν υπάρχουν πολλές μέθοδοι, και παραμένει ακόμα σε χαμηλά επίπεδα κάτω του 40%.

[7] https://sokcertifiedrobustness.github.io/leaderboard/

	ℓ_∞ Adversary			
MNIST	$\epsilon = 0.1$		$\epsilon = 0.3$	
	98.22%	[140], Interval	**94.02%**	[63], Polyhedra
	98.14%	[141], Smooth Layers	93.40%	[141], Smooth Layers
	97.95%	[122], Interval	93.40%	[140], Interval
	97.95%	[127], Smooth Layers	93.20%	[127], Smooth Layers
	97.91%	[143], Duality	93.10%	[122], Interval
CIFAR-10	$\epsilon = 2/255$		$\epsilon = 8/255$	
	68.2%*	[98], Nym.-Prsn.	**40.39%**	[141], Smooth Layers
	63.8%*	[144], Nym.-Prsn.	40.06%	[127], Smooth Layers
	62.84%	[140], Interval	35.42%	[89], Smooth Layers
	60.5%	[148], Polyhedra	35.13%	[140], Interval
	56.94%	[141], Smooth Layers	34.97%	[122], Interval
ImageNet	$\epsilon = 1/255$		No work achieves $> 0\%$ certified accuracy under large ϵ yet.	
	38.2%*	[98], Nym.-Prsn.		
	28.6%*	[22], Nym.-Prsn.		

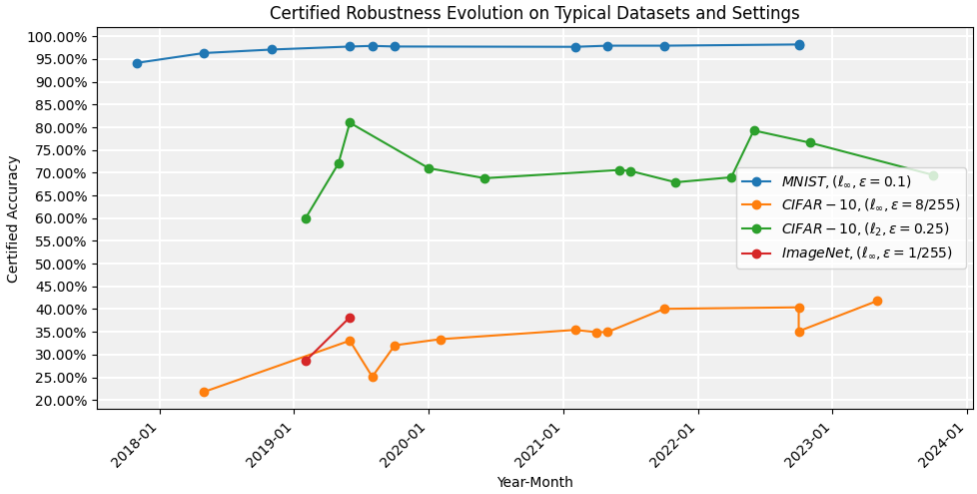
Σχήμα 4.12: *Συγκριτικός πίνακας (SoK Leaderboard) με τα 5 καλύτερα μοντέλα και μεθόδους certified robustness για τα MNIST, CIFAR-10, ImageNet datasets και διαταραχή ℓ_∞ για διάφορες τιμές ϵ [102]. Με '*' σημειώνονται οι probabilistic verification προσεγγίσεις, αλλιώς πρόκειται για deterministic verification προσεγγίσεις. Με **bold** σημειώνονται οι καλύτερες επιδόσεις σε κάθε κατηγορία.*

	ℓ_2 Adversary			
MNIST	$\epsilon = 0.5$		$\epsilon = 1.58$	
	98.2%*	[139], Nym.-Prsn.	**70.7%***	[139], Nym.-Prsn.($\epsilon = 1.75$)
	98.0%*	[142], Nym.-Prsn.	70.5%*	[142], Nym.-Prsn.($\epsilon = 1.75$)
	78.45%	[90], Curvature	69.79%	[90], Curvature
			69.0%*	[93], Divergence Based
			62.8%	[81], General Lipschitz
CIFAR-10	$\epsilon = 36/255$		$\epsilon = 0.25$	
	65.6%*	[93], Divergence Based	**81%***	[98], Nym.-Prsn.
	64.49%	[145], Smooth Layers	79.3%*	[146], Nym.-Prsn.
	62.96%	[87], Smooth Layers	76.6%*	[147], Nym.-Prsn.
	59.16%	[88], Smooth Layers	72%*	[144], Nym.-Prsn.
	58.4%	[81], General Lipschitz	71%*	[149], Nym.-Prsn.
ImageNet	$\epsilon = 1.0$		$\epsilon = 2.0$	
	67.0%*	[147], Nym.-Prsn.	**42.2%***	[147], Nym.-Prsn.
	54.3%*	[146], Nym.-Prsn.	30.4%*	[150], Nym.-Prsn.
	45%*	[98], Nym.-Prsn.	29.5%*	[146], Nym.-Prsn.
	44.6%*	[151], Nym.-Prsn.	28.6%*	[151], Nym.-Prsn.
	44.4%*	[150], Nym.-Prsn.	28%*	[98], Nym.-Prsn.

Σχήμα 4.13: *Συγκριτικός πίνακας (SoK Leaderboard) με τα 5 καλύτερα μοντέλα και μεθόδους certified robustness για τα MNIST, CIFAR-10, ImageNet datasets και διαταραχή ℓ_1 για διάφορες τιμές ϵ [102]. Με '*' σημειώνονται οι probabilistic verification προσεγγίσεις, αλλιώς πρόκειται για deterministic verification προσεγγίσεις. Με **bold** σημειώνονται οι καλύτερες επιδόσεις σε κάθε κατηγορία.*

4.7 Συμβιβασμοί Ευρωστίας

Όπως έχουμε ήδη αναφέρει στις προηγούμενες ενότητες, στην προσπάθεια ενίσχυσης της ευρωστίας ενός μοντέλου, μπορεί να επηρεαστούν άλλοι παράμετροι του μοντέλου αρνητικά ή να προκύψουν κάποιες μη θεμιτές παρενέργειες (side-effects). Οπότε υπάρχουν κάποιοι



Σχήμα 4.14: *Διάγραμμα εξέλιξης certified robustness από δημοσιευμένες μεθόδους δοκιμασμένες σε επιλεγμένα dataset και μοντέλα απειλών, σύμφωνα με τα δεδομένα του SoK Leaderboard. Ο x-άξονας αναπαριστά τον χρόνο χωρισμένο ανά έτος, ενώ ο y-άξονας αναπαριστά το certified robust accuracy της κάθε δημοσιευμένης μεθόδου.*

συμβιβασμοί (trade-offs) που πρέπει να ληφθούν υπόψη, όταν χρησιμοποιούνται τεχνικές για την ενίσχυση της ευρωστίας ενός μοντέλου.

4.7.1 Ακρίβεια

Σε ορισμένες περιπτώσεις, η εκπαίδευση και ο σχεδιασμός ενός ML συστήματος με γνώμονα την ασφάλεια, ενδέχεται να έρχονται σε αντίθεση με τον στόχο της υψηλής ακρίβειας, καθώς ορισμένες τεχνικές και μέθοδοι που χρησιμοποιούνται για να κάνουν τα ML μοντέλα πιο εύρωστα (και άρα ασφαλή) προκαλούν απώλεια στην κανονική ακρίβεια του μοντέλου. Παρόλο που οι θεωρητικές ενδείξεις δείχνουν ότι δεν υπάρχει εγγενές συμβιβασμός μεταξύ ευρωστίας και ακρίβειας, στην πράξη οι διαθέσιμες μέθοδοι αποτυγχάνουν να επιτύχουν αύξηση και στις δύο μετρήσεις [100].

4.7.1.1 Θεωρητική ερμηνεία

Ένας από τους πιο αποτελεσματικούς τρόπους για την επίτευξη ευρωστίας ενός ML μοντέλου είναι το adversarial training, το οποίο με την επαύξηση των δεδομένων εκπαίδευσης με adversarial examples, προσπαθεί να μειώσει το robust error (σφάλμα στις χειρότερες διαταραγμένες εισόδους) και άρα να αυξήσει το robust accuracy του μοντέλου. Στην πράξη όμως, έχει παρατηρηθεί ότι ταυτόχρονα **αυξάνεται το standard error** (σφάλμα στις καθαρές, μη διαταραγμένων, εισόδους) και άρα **μειώνεται το standard accuracy** του μοντέλου, χειροτερεύοντας τη δυνατότητα γενίκευσης του μοντέλου [118] (βλ. παράδειγμα 4.15).

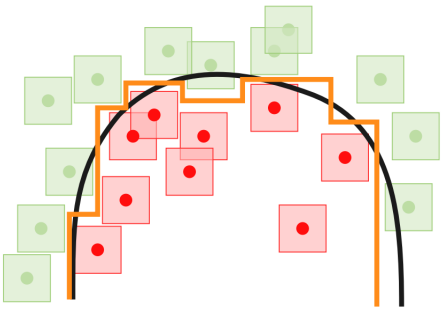
Μία από τις πιο εδραιωμένες εξηγήσεις για αυτό το φαινόμενο, στηρίζεται στο γεγονός ότι οι στόχοι της τυπικής απόδοσης ενός μοντέλου και του adversarial robustness είναι αντίθετοι, λόγω των εγγενών διαφορών μεταξύ των αναπαραστάσεων των χαρακτηριστικών που μαθαίνουν από την κανονική και την εύρωστη εκπαίδευση αντίστοιχα. Για παράδειγμα, κατά το adversarial training μειώνεται η εξάρτηση από μη εύρωστα χαρακτηριστικά, τα οποία όμως είναι πιθανώς χρήσιμα [120]. Άρα **υπάρχει ένας εγγενές συμβιβασμός μετα-**

	Standard training	Adversarial training
Robust test	3.5%	45.8%
Robust train	-	100%
Standard test	95.2%	87.3%
Standard train	100%	100%

Σχήμα 4.15: *Αποτελέσματα standard και robust accuracy σε δεδομένα εκπαίδευσης και δοκιμής για ένα μοντέλο εκπαιδευμένο τυπικά αλλά και ανταγωνιστικά στο CIFAR-10. Και στις δύο περιπτώσεις η κανονική ακρίβεια στα δεδομένα εκπαίδευσης είναι 100% αλλά μειώνεται αισθητά στα δεδομένα δοκιμής [119].*

ξύ standard accuracy και robust accuracy ενός μοντέλου και δεν προκύπτει ως κάποια παρενέργεια του adversarial training, το οποίο έχει αποδειχθεί σε σχετικά απλές και φυσικές συνθήκες (binary ταξινομητές) [121], αλλά παρουσιάζεται και εμπειρικά σε πιο μεγάλα μοντέλα και περίπλοκες συνθήκες (βλ. παρακάτω 4.7.1.2). Αυτός ο συμβιβασμός είναι υπαρκτό και σε certified robustness τεχνικές, οι οποίες παρέχουν αποδεδειγμένες εγγυήσεις για την ευρωστία του μοντέλου, αλλά αγνοούν την απόδοση των μοντέλων σε μη ανταγωνιστικά, αφήνοντας ανοιχτή τη θεωρητική αντιμετώπιση αυτού του συμβιβασμού [122].

Αντίθετα, με την παραπάνω εξήγηση, άλλες έρευνες από τη βιβλιογραφία, έχουν δείξει ότι **η ευρωστία και η ακρίβεια δεν είναι απαραίτητα αντίθετοι στόχοι** και μπορούν να επιτευχθούν παράλληλα σε πραγματικές συνθήκες, αλλά οι μέχρι τώρα τεχνικές δεν το επιτυγχάνουν. Η μία ερμηνεία για τη διαφορά αυτή μεταξύ θεωρίας και πράξης, βασίζεται σε δύο περιορισμούς των τρεχουσών τεχνικών: (i) είτε στην αποτυχία τους στο να επιβάλουν τοπικές Lipschitz ιδιότητες, (ii) είτε στην ανεπάρκεια γενίκευσης τους. Δηλαδή, σε πραγματικά σύνολα δεδομένων (π.χ. εικόνες), τα οποία ακολουθούν μια ιδιότητα φυσικού διαχωρισμού, μπορεί να υπάρξει ένας εύρωστος και απόλυτα ακριβής ταξινομητής, που μπορεί να ληφθεί μέσω της στρογγυλοποίησης μιας τοπικής Lipschitz συνάρτησης [100] (βλ. σχήμα 4.16).



Σχήμα 4.16: *Σχηματικό διάγραμμα δύο ταξινομητών σε ένα σύνολο 2 διαφορετικών ειδών δεδομένων. Ο ταξινομητής με το πορτοκαλί όριο έχει μικρές τοπικές Lipschitz ιδιότητες καθώς δεν αλλάζει σε ℓ_∞-σφαίρες γύρω από τα σημεία. Αντίθετα, ο ταξινομητής με το* **μαύρο** *όριο είναι ευάλωτος σε adversarial examples παρά την υψηλή ακρίβεια που έχει [100].*

Μία άλλη ερμηνεία, βασίζεται στο ότι σε ρεαλιστικές συνθήκες (με ανεπαίσθητες ℓ_∞ διαταραχές που διατηρούν τις ίδιες ετικέτες), η **χρήση περισσότερων (θεωρητικά άπειρων) δεδομένων εκπαίδευσης**, απορρίπτει την ύπαρξης αυτού του εγγενούς συμβιβα-

σμού μεταξύ robust accuracy και standard accuracy [118]. Πράγματι, το adversarial training για παράδειγμα, είναι μια διαδικασία που απαιτεί πολλά δεδομένα [123] και τα εμπειρικά αποτελέσματα (βλ. παρακάτω 4.7.2) δείχνουν να επιβεβαιώνουν στατιστικά την άποψη ότι αυτός ο συμβιβασμός ισχύει λόγο της ανεπάρκειας δεδομένων που υπάρχει και ότι η χρήση πρόσθετων δειγμάτων αρκεί για να το αντισταθμίσει [119].

Όμως, ακόμα και στην περίπτωση της ύπαρξης άπειρων δεδομένων, γενικότερα ο συμβιβασμός αυτός ισχύει ακόμη θεωρητικά, το οποίο δείχνει ότι είναι ένα πρόβλημα που δεν έχει να κάνει με τη μη ύπαρξη επαρκών δεδομένων αλλά με την κατανομή αυτών [121].

4.7.1.2 Εμπειρικές παρατηρήσεις

Αρχικά είναι σημαντικό να διαχωρίσουμε τον συμβιβασμό ευρωστίας και ακρίβειας σε 2 διαφορετικές περιπτώσεις: (**1**) τη διαφορά του standard accuracy μετά από κάποιου είδους εύρωστης εκπαίδευση (συνήθως adversarial training), δηλαδή στη διαφορά της πριν και μετά την εκπαίδευση πάνω σε καθαρά δεδομένα (βλ. παράδειγμα 4.15), και (**2**) τη διαφορά standard accuracy και robust accuracy μετά την εύρωστη εκπαίδευση, δηλαδή στη διαφορά της ακρίβειας μετά την εκπαίδευση πάνω σε καθαρά και διαταραγμένα δεδομένα (βλ. διάγραμμα 4.9). Στη θεωρητική προσέγγιση, επικεντρωθήκαμε κυρίως στη 1η περίπτωση, αλλά εδώ παρουσιάζουμε και τα εμπειρικά αποτελέσματα και για τις δύο περιπτώσεις.

Για τη **1η περίπτωση** υπάρχουν πολλές έρευνες που έχουν παρατηρήσει πτώση της καθαρής ακρίβειας του εύρωστα εκπαιδευμένου μοντέλου σε σχέση με το αντίστοιχο μη εύρωστο, με διαφορά πάνω από 10% ανάλογα τις συνθήκες, πιο συγκεκριμένα πτώση του standard accuracy έως και το 10% στο CIFAR-10 και 15% στο ImageNet [47].

Επίσης, από τα αποτελέσματα του RobustBench [98], μπορούμε να εξάγουμε κάποια ποσοτικά αποτελέσματα. Αρχικά από το διάγραμμα 4.9 που αναλύσαμε, φαίνεται η διαφορά standard accuracy σε σχέση με το robust accuracy υπολογισμένη από το AutoAttack με βάση τα αποτελέσματα του RobustBench 4.6.2 για τα CIFAR-10 και CIFAR-100 με ℓ_∞. Το πρώτο που παρατηρήσαμε είναι ότι το robust accuracy του μοντέλου μειώνεται σε σχέση με την κανονική του ακρίβεια, αλλά η χρήση παραπάνω δεδομένων στην εκπαίδευση φαίνεται να αντισταθμίζει κάπως αυτό τον συμβιβασμό. Επίσης, από την έρευνα του RobustBench [98] προκύπτει ότι για όλα τα μοντέλα με υψηλή ευρωστία ενάντια σε διαταραχές ℓ_∞ έως και $\epsilon = 8/255$, έχουν σημαντική πτώση στο standard accuracy τους, σε σχέση με τα αντίστοιχα τυπικά εκπαιδευμένα μοντέλα. Από τα ίδια αποτελέσματα όμως, για τα αντίστοιχα τα εύρωστα μοντέλα στο CIFAR-10 σε διαταραχές ℓ_2, παρουσιάζουν υψηλότερη standard accuracy από ότι τα τυπικά εκπαιδευμένα μοντέλα (95.74% και 94.78% αντίστοιχα), ενώ προφανώς έχουν υψηλότερο robust accuracy από ότι τα τυπικά (82.32% και 0.00% αντίστοιχα), και έχοντας παράλληλα ένα πολύ μικρό χάσμα μεταξύ robust accuracy και standard accuracy.

Ένα ακόμα παράδειγμα από τη βιβλιογραφία [121], δείχνει την επίπτωση που έχει η εύρωστη εκπαίδευση στην καθαρή ακρίβεια ενός ταξινομητή, ακόμα και με τη χρήση ενός μείγματος καθαρών και adversarial δειγμάτων (50% αναλογία συγκεκριμένα) σε κάθε κύκλο εκπαίδευσης. Στον πίνακα 4.17 φαίνονται τα αποτελέσματα standard και robust accuracy, σε 3 διαφορετικά σενάρια εκπαίδευσης (μόνο καθαρά δείγματα, μισά-μισά, μόνο ανταγωνιστικά δείγματα), για ℓ_2 και ℓ_∞ διαταραχές στα MNIST, CIFAR-10 dataset. Αν εστιάσουμε καθαρά στα αποτελέσματα καθαρής ακρίβειας, για κάθε μοντέλο απειλής και σύνολο δεδομένων, όσο αυξάνεται

το threshold ϵ τόσο πιο μεγάλη είναι και η πτώση της. Συγκεκριμένα, στο MNIST έχουμε μικρότερο ποσοστό πτώσης, 1.94% στη χειρότερη περίπτωση ($\ell_\infty, \epsilon = 0.3$), ενώ φαίνεται το μείγμα δειγμάτων να μην έχει μεγάλη επίδραση. Όμως, στο CIFAR-10, στη χειρότερη περίπτωση ($\ell_2, \epsilon = 320/255$) έχουμε πτώση 16.45% της καθαρής ακρίβειας, και στην περίπτωση του μείγματος δειγμάτων φαίνεται τώρα η επίδραση με πτώση ακρίβειας 10.46%.

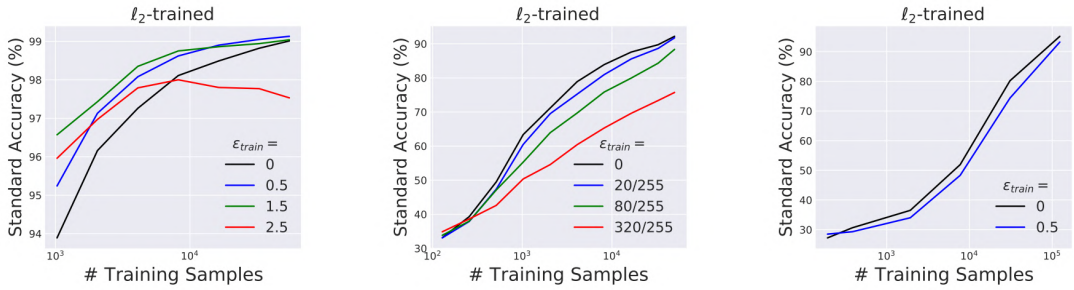
			Standard Accuracy			Robust Accuracy		
	Norm	ε	Standard	Half-half	Robust	Standard	Half-half	Robust
MNIST	ℓ_∞	0	99.31%	-	-	-	-	-
		0.1	99.31%	99.43%	99.36%	29.45%	95.29%	95.05%
		0.2	99.31%	99.22%	98.99%	0.05%	90.79%	92.86%
		0.3	99.31%	99.17%	97.37%	0.00%	89.51%	89.92%
	ℓ_2	0	99.31%	-	-	-	-	-
		0.5	99.31%	99.35%	99.41%	94.67%	97.60%	97.70%
		1.5	99.31%	99.29%	99.24%	56.42%	87.71%	88.59%
		2.5	99.31%	99.12%	97.79%	46.36%	60.27%	63.73%
CIFAR10	ℓ_∞	0	92.20%	-	-	-	-	-
		2/255	92.20%	90.13%	89.64%	0.99%	69.10%	69.92%
		4/255	92.20%	88.27%	86.54%	0.08%	55.60%	57.79%
		8/255	92.20%	84.72%	79.57%	0.00%	37.56%	41.93%
	ℓ_2	0	92.20%	-	-	-	-	-
		20/255	92.20%	92.04%	91.77%	45.60%	83.94%	84.70%
		80/255	92.20%	88.95%	88.38%	8.80%	67.29%	68.69%
		320/255	92.20%	81.74%	75.75%	3.30%	34.45%	39.76%

Σχήμα 4.17: *Αποτελέσματα standard και robust accuracy σε 3 σενάρια εκπαίδευσης:* ***Standard*** *(μόνο καθαρά δείγματα),* ***Half-Half*** *(μείγμα καθαρών και adversarial δειγμάτων),* ***Robust*** *(μόνο ανταγωνιστικά δείγματα). Εκπαίδευση ταξινομητή σε διαταραχή ℓ_∞, ℓ_2 για διάφορες τιμές ϵ στα MNIST, CIFAR-10 dataset [121].*

Επίσης, στην την ίδια έρευνα, από το σχήμα 4.18 μπορούμε να βγάλουμε εμπειρικά αποτελέσματα για την επίπτωση που έχουν το μέγεθος των δεδομένων εκπαίδευσης στην standard accuracy ενός ανταγωνιστικά εκπαιδευμένου μοντέλου. Όπως φαίνεται οι ταξινομητές που έχουν εκπαιδευτεί ενάντια σε διαταραχές ℓ_2 (το ίδιο ισχύει και για ℓ_∞) και για διάφορες τιμές ϵ, με λίγα δείγματα παρουσιάζουν χαμηλότερες τιμές ακρίβειας, όμως από ένα σημείο και μετά, κυρίως για τα μικρά σύνολα δεδομένων (δηλαδή στην περίπτωση του MNIST), ελαττώνεται ο ρυθμός αύξησης της ακρίβειας ή ακόμα η τιμή της ακρίβειας αρχίζει και πέφτει κάτω από την τιμή του μοντέλου με τυπική εκπαίδευση. Οπότε υπάρχει ένα όριο στο μέγεθος των δεδομένων, το οποίο αν ξεπεραστεί, τα επιπλέον αυτά δεδομένα έχουν αρνητική επίδραση στην standard accuracy του μοντέλου [121].

Για τη **2η περίπτωση** που επίσης υπάρχει ένα χάσμα standard accuracy και robust accuracy, μπορούμε να εξάγουμε κάποια εμπειρικές τιμές για αυτήν τη διαφορά από τα αποτελέσματα του RobustBench [98], το οποίο περιλαμβάνει τα καλύτερα μοντέλα από πλευράς empirical robustness. Για να το κάνουμε αυτό, υπολογίσαμε τη διαφορά μεταξύ καθαρής και εύρωστης ακρίβειας (**standard accuracy-robust accuracy**), για κάθε διαθέσιμο μοντέλο στα leaderboards σε κάθε σύνολο δεδομένων και μοντέλο απειλής. Έπειτα υπολογίσαμε τον μέσο όρο αυτής της διαφοράς και τη μέγιστη και ελάχιστη καταγεγραμμένη, τα οποία φαίνονται συγκεντρωτικά στον πίνακα 4.1.

Στην ℓ_2 περίπτωση εμφανίζεται η μικρότερη διαφορά, κοντά στο 15%, που όπως έχουμε δει έχει να κάνει και με το ότι στα πιο απλά σύνολα δεδομένων ή στις πιο μικρές διαταραχές,



(a) MNIST (b) CIFAR-10 (c) Restricted ImageNet

Σχήμα 4.18: *Διαγράμματα standard accuracy σε σχέση με το μέγεθος των δειγμάτων εκπαίδευσης για ταξινομητές εκπαιδευμένους ανταγωνιστικά ενάντια σε ℓ_2 διαταραχές (για διάφορες τιμές ϵ) στα MNIST, CIFAR-10, ImageNet datasets [121]. Με ϵ_{train} συμβολίζεται η καμπύλη τυπικής εκπαίδευσης ($\epsilon = 0$).*

είναι ευκολότερο να επιτευχθεί μεγάλο robust accuracy, οπότε και η διαφορά με την καθαρή ακρίβεια είναι πολύ μικρότερη. Όμως για ℓ_∞ διαταραχές, φαίνεται ότι υπάρχει ένα χάσμα 25%-35% σε όλα τα σύνολα δεδομένων, με μεγαλύτερη μέση διαφορά να εμφανίζεται στο CIFAR-10 και CIFAR-100 (κοντά στο 35%) και μικρότερη μέση διαφορά στο ImageNet (κοντά στο 25%). Αυτό όμως έχει να κάνει με το ότι υπάρχουν πολύ περισσότερες δεδομένα και μοντέλα στις πρώτες περιπτώσεις τα οποία ρίχνουν τη μέση διαφορά, καθώς αν παρατηρήσουμε την ελάχιστη διαφορά, η οποία έχει προκύψει κυρίως από τα μοντέλα με τις καλύτερες επιδόσεις ευρωστίας, βλέπουμε ότι εμφανίζεται στα CIFAR-10 και CIFAR-100 (12% και 10% αντίστοιχα), ενώ στο ImageNet η ελάχιστη διαφορά είναι σχεδόν στο 20%.

Dataset	Threat Model	Avg Diff.	Max Diff.	Min Diff.
CIFAR-10	$(\ell_\infty, \epsilon = 8/255)$	34.87%	93.53%	12.09%
CIFAR-10	$(\ell_2, \epsilon = 0.5)$	15.8%	22.61%	10.57%
CIFAR-100	$(\ell_\infty, \epsilon = 8/255)$	34.42%	46.49%	28.37%
ImageNet	$(\ell_\infty, \epsilon = 4/255)$	27.27%	76.52%	19.3%

Πίνακας 4.1: *Διαφορά **Standard-Robust accuracy** για κάθε dataset και threat model από τα διαθέσιμα αποτελέσματα των RobustBench Leaderboards. Στις 3 τελευταίες στήλες, φαίνεται μέσος όρος αυτής της διαφοράς από όλα τα μοντέλα, καθώς και η μέγιστη και ελάχιστη διαφορά αντίστοιχα.*

4.7.1.3 Τρόποι αντιμετώπισης

Όπως έχουμε αναφέρει ήδη, ένας τρόπος αντιμετώπισης αυτού του συμβιβασμού είναι με τη **χρήση περισσότερων δεδομένων εκπαίδευσης**, καθώς πρακτικά είναι μια από τις πιο απλές τεχνικές και δείχνει να έχει αποτέλεσμα χωρίς απαραίτητα να είναι η βέλτιστη λύση [119]. Αντίστοιχα, άλλες προτεινόμενες τεχνικές βασίζονται στην αρχή των περισσότερων δεδομένων, όπως η **Robust Self-Training (RST)** [124], που χρησιμοποιεί τυπική supervised εκπαίδευση για να αποκτήσει ψευδό-ετικέτες για τα δεδομένα που δεν έχουν, και τα τροφοδοτεί στον supervised αλγόριθμο εκπαίδευσης που χρησιμοποιείται για την ενίσχυση του adversarial robustness. Στην ουσία, είναι ένας τρόπος αξιοποίησης πολλών δεδομένα

χωρίς ετικέτες, και αποδεικνύεται ότι για γραμμική παλινδρόμηση ο RST εξαλείφει αυτόν τον συμβιβασμό, επιτυγχάνοντας παράλληλα το καλύτερο δυνατό robust error, σχεδόν στα ίδια επίπεδα του standard error [118]

Μια ακόμα τεχνική που κάνει χρήση επαύξησης δεδομένων, είναι η **Interpolated Adversarial Training (IAT)** [99], η οποία κατά την εκπαίδευση συνδυάζει παρεμβολές από adversarial examples μαζί με παρεμβολές από καθαρά και μη διαταραγμένα δείγματα, χρησιμοποιώντας τις τεχνικές των Mixup [125] και Manifold Mixup [126] ως τρόπους παρεμβολής αυτών των παραδειγμάτων. Με αυτήν την μέθοδο βελτιώνεται η γενίκευση σε καθαρά δείγματα (standard accuracy) διατηρώντας παράλληλα το adversarial robustness, κυρίως λόγω της αύξησης του συνόλου δεδομένων εκπαίδευσης και της συμπίεσης πληροφοριών που συμβαίνει στα χαρακτηριστικά που μαθαίνονται από τα βαθιά δίκτυα.

Άλλη μια αμυντική μέθοδος, που είναι σχεδιασμένη για την αντιμετώπιση του συγκεκριμένου συμβιβασμού είναι η **TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES)** [122], η οποία χρησιμοποιεί μια μέθοδο βελτιστοποίησης για να προσφέρει ένα άνω όριο στη διαφορά μεταξύ robust error και standard error. Με αυτήν τη μέθοδο, προσφέρονται θεωρητικές εγγυήσεις για τη διαφορά μεταξύ ακρίβειας και ευρωστίας των μοντέλων, ενώ ταυτόχρονα εμπειρικά (στο κάτω όριο) παρουσιάζει τις καλύτερες επιδόσεις, παίρνοντας και τη 1η θέση ανάμεσα σε πάνω από 2000 άλλες τεχνικές και μοντέλο στον διαγωνισμό NeurIPS 2018 Adversarial Vision Challenge [127].

Τα τελευταία χρόνια η έρευνα για εύρωστα ML συστήματα έχει επικεντρωθεί κυρίως στην ανάπτυξη νέων συναρτήσεων κόστους, ωστόσο η υπόλοιπη διαδικασία εκπαίδευσης (π.χ. τοπολογίες δικτύου, μέθοδοι βελτιστοποίησης, εργαλεία γενίκευσης) παραμένει ιδιαίτερα προσαρμοσμένη στην προώθηση της ακρίβειας. Για να επιτευχθούν ταυτόχρονα οι στόχοι της ευρωστία και ακρίβειας, πρέπει μελλοντικά να **επανασχεδιαστούν και άλλες πτυχές της εκπαίδευσης** των νευρωνικών δικτύων, συνδυαστικά με τη χρήση καλύτερων μεθόδων βελτιστοποίησης [100]. Υπάρχουν όμως και κάποιες άλλες αμυντικές τεχνικές, πέρα από τις εύρωστες εκπαιδεύσεις και τη χρήση δεδομένων, όπως η **PixelDefend** [91] η οποία χρησιμοποιεί μετασχηματισμούς που προσπαθούν να ελαχιστοποιήσουν την απώλεια ακρίβειας και παράλληλα να ανακτήσουν τη σωστή ταξινόμηση μιας ανταγωνιστικής εικόνας.

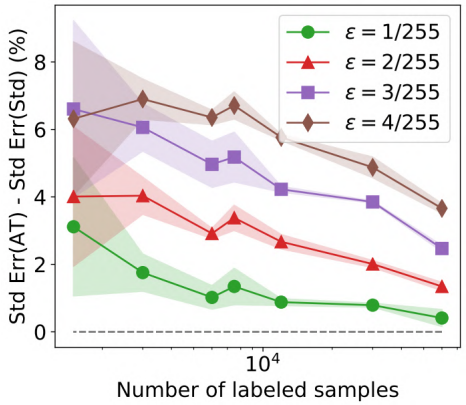
Τέλος, παρά την αρνητική επίπτωση που παρουσιάζουν οι εύρωστοι μέθοδοι εκπαίδευσης στην κανονική ακρίβεια, τα εύρωστα μοντέλα παρουσιάζουν και κάποια **απροσδόκητα οφέλη**, όπως το ότι μαθαίνουν σημαντικές αναπαραστάσεις χαρακτηριστικών που ευθυγραμμίζονται καλά με τα κύρια χαρακτηριστικά των δεδομένων και άρα είναι πιο αμετάβλητα σε άλλες τροποποιήσεις της εισόδου που μπορούν να ξεγελάσουν την ανθρώπινη αντίληψη [121]. Είναι λοιπόν σημαντικό να συνεχίζουν να αναπτύσσονται και να χρησιμοποιούνται εύρωστοι μέθοδοι εκπαίδευσης, καθώς η ευρωστία των μοντέλων δεν πρόκειται να προκύψει ποτέ από την απλή και κανονική εκπαίδευση [121].

4.7.2 Υπολογιστικό κόστος

Στην προσπάθεια αύξησης της ευρωστίας ενός μοντέλου, **αυξάνεται συνήθως και το υπολογιστικό κόστος** που χρειάζεται για να εκπαιδευτεί εύρωστα ή να γίνει επιβεβαίωση της ευρωστίας του.

Αρχικά όπως έχουμε ήδη αναφέρει, για να επιτευχθεί καλύτερη ακρίβεια κατά το adver-

sarial training, υπάρχει η ανάγκη **χρήσης περισσότερων δεδομένων εκπαίδευσης**, κάτι το οποίο την καθιστά πιο απαιτητική υπολογιστικά από την τυπική εκπαίδευση [40]. Όπως φαίνεται και εμπειρικά στο σχήμα 4.19, για να μειωθεί η διαφορά του standard error (άρα αύξηση καθαρής ακρίβειας) ενός ανταγωνιστικά εκπαιδευμένου μοντέλου σε σχέση με την αρχική στο CIFAR-10, χρειάζονται όλο και περισσότερα δεδομένα, και πιθανώς πλησιάζοντας τα άπειρα δεδομένα να φτάσει στην αρχική καθαρή ακρίβεια. Όμως, ακόμα και να ίσχυε θεωρητικά, θα ήταν υπολογιστικά αδύνατο ή πολύ δύσκολο να επιτευχθεί μια εκπαίδευση με άπειρα (ή κοντά σε αυτό) δεδομένα, τουλάχιστον με τις τωρινές μεθόδους εκπαίδευσης και υπολογιστική δύναμη [121].



Σχήμα 4.19: *Διάγραμμα της διαφοράς standard error με adversarial training με ℓ_∞ διαταραχές (για διάφορες τιμές ϵ) και standard error με κανονική εκπαίδευση σε σχέση με το μέγεθος των δειγμάτων εκπαίδευσης. Η διαφορά αυτή μειώνεται όσο αυξάνεται το μέγεθος των δεδομένων [118].*

Ακόμη, υπάρχει και το ζήτημα του **αριθμού και του είδους των διαταραχών** για τα οποία εκπαιδεύονται τα μοντέλα. Όλες οι μέθοδοι εύρωστης εκπαίδευσης εστιάζουν κάθε φορά σε ένα συγκεκριμένο είδος ℓ_p διαταραχών, για τον έλεγχο του μεγέθους της διαταραχής, το οποίο όμως προσφέρει ευρωστία μόνο για αυτά τα είδη διαταραχών και καμία εγγύηση για adversarial examples τα οποία δεν έχουν χρησιμοποιηθεί στην εκπαίδευση των μοντέλων (hidden adversarial samples) και περιέχουν άλλου είδους διαταραχές [40]. Ελάχιστη έρευνα έχει γίνει προς το παρόν, πάνω στην επέκταση της ευρωστίας σε πολλαπλά είδη ℓ_p-φραγμένων διαταραχών, αλλά τα αποτελέσματα δείχνουν την υπολογιστική δυσκολία και κλιμάκωση που εμφανίζεται ένα τέτοιο εγχείρημα και ενώ για μικρά μοντέλα και διαταραχές υπάρχουν υλοποιήσεις, στη γενική του μορφή παραμένει ακόμα ένα ανοιχτό πρόβλημα [128].

Επίσης, στον τομέα του robustness verification υπάρχει αντίστοιχος συμβιβασμός υπολογιστικού κόστους και **εγγυημένης ευρωστίας**, στο οποίο αναφερθήκαμε στη σύγκριση empirical και certified robustness 4.3.3, καθώς όσο αυξάνεται η πολυπλοκότητα του μοντέλου και του χώρου εισόδου, η εύρεση αυστηρών ορίων και εγγυήσεων γίνεται όλο και πιο δύσκολη [101]. Γενικότερα, είναι υπολογιστικά δύσκολη η εγγυημένη αξιολόγηση της ευρωστίας ακόμα και σε απλές ℓ_p-φραγμένες διαταραχές και οι ακριβείς προσεγγίσεις όπως είδαμε δεν μπορούν να κλιμακωθούν σε μεγάλα μοντέλα [104].

4.7.3 Χρόνος

Αντίστοιχα με την αύξηση του υπολογιστικού κόστους, στην προσπάθεια αύξησης της ευρωστίας ενός μοντέλου, **αυξάνεται συνήθως και ο χρόνος** της εύρωστης εκπαίδευσης.

Είναι ένα από τα μειονεκτήματα του adversarial training , η **αύξηση του χρόνου εκπαίδευσης**, καθώς πρέπει να υπολογίζονται νέες διαταραχές σε κάθε βήμα ενημέρωσης των παραμέτρων [121]. Για παράδειγμα, με χρήση του PGD αλγόριθμου, χρειάζονται τουλάχιστον 10 επιπλέον αριθμοί επαναλήψεων για να ληφθούν καλά adversarial examples, το οποίο μπορεί να μεταφραστεί σε περίπου 10-πλάσια επιβράδυνση σε σχέση με την τυπική εκπαίδευση [105]. Οπότε, το adversarial training στο CIFAR-10 μπορεί να διαρκέσει μερικές μέρες, ενώ στο ImageNet μερικές βδομάδες, ανάλογα και τους διαθέσιμους υπολογιστικούς πόρους.

Ακόμη, βλέποντας και τους χρόνους εύρωστης εκπαίδευσης για 4 διαφορετικές περιπτώσεις διαταραχών στα σύνολα δεδομένων από το RobustBench [98], παρατηρούμε ότι τα πιο εύρωστα μοντέλα χρειάζονται πολύ περισσότερο χρόνο (βλ. 4.20). Ενώ ο χρόνος επηρεάζεται και από το μέγεθος και την πολυπλοκότητα του μοντέλου και των δεδομένων, το πιο εύρωστο μοντέλο είναι για ℓ_2 στο CIFAR-10 με robust accuracy 78.80%, και χρόνο εκπαίδευσης λίγο πάνω από 15 ώρες. Αντίστοιχα, το λιγότερο χρόνο τον έχει το μοντέλο με το χαμηλότερο robust accuracy (25.31%) για ℓ_∞ στο ImageNet, με χρόνο εκπαίδευσης κοντά στη 1.5 ώρα.

Dataset	Leaderboard	Paper	Architecture	Clean acc.	Robust acc.	Time
CIFAR-10	ℓ_∞	Gowal et al. [50]	WRN-28-10	89.48%	62.82% ± 0.016	11.8 h
CIFAR-10	ℓ_2	Rebuffi et al. [111]	WRN-28-10	91.79%	78.80% ± 0.000	15.1 h
CIFAR-100	ℓ_∞	Wu et al. [155]	WRN-34-10	60.38%	28.84% ± 0.018	6.6 h
ImageNet	ℓ_∞	Salman et al. [117]	ResNet-18	52.92%	25.31% ± 0.010	1.6 h

Σχήμα 4.20: *Πίνακας αποτελεσμάτων για robust εκπαίδευση μοντέλων με το AutoAttack framework, ενάντια σε ℓ_∞, ℓ_2 διαταραχές στα CIFAR-10, CIFAR-100, ImageNet datasets με αναγραφή του clean και robust accuracy, και του χρόνου εκπαίδευσης σε κάθε περίπτωση [98]. Όλες οι εκτελέσεις γίνονται πάνω σε μία Tesla V100 GPU.*

Υπάρχουν όμως και υλοποιήσεις που υπολογίζουν το κομμάτι του χρόνου εκπαίδευσης και προσφέρουν **βελτιστοποίησης για την επίτευξη πιο γρήγορης εκπαίδευσης**. Κάποιες μέθοδοι, χρησιμοποιούν τροποποιημένες εκδόσεις του PGD ή του FGSM για να επιτύχουν αυτόν τον σκοπό, όμως με τον κίνδυνο της παραγωγής πιο αδύναμων adversarial δειγμάτων και άρα εκπαίδευσης ενός λιγότερου εύρωστου μοντέλου [40].

Επίσης, όπως αναφερθήκαμε και από πάνω, για την επίτευξη καλύτερης καθαρής ακρίβειας κατά το adversarial training , γίνεται **χρήση περισσότερων δεδομένων εκπαίδευσης**, κατά την οποία αυξάνεται πέρα από το υπολογιστικό κόστος και ο χρόνος εκπαίδευσης του μοντέλου.

4.8 Εργαλεία Ανοιχτού Κώδικα και Βιβλιοθήκες

Στην ενότητα των Robustness Benchmarks 4.6 αναφερθήκαμε στα 2 βασικότερα benchmarks που υπάρχουν αυτήν τη στιγμή για τη συγκριτική αξιολόγηση του empirical και certified robustness σε μοντέλα και τεχνικές. Όμως, υπάρχουν αρκετές άλλες βιβλιοθήκες και εργαλεία ανοιχτού κώδικα (open-source) που μπορούν να χρησιμοποιηθούν για την εκπαίδευση των

νευρωνικών δικτύων με σκοπό την αύξηση της ευρωστίας, αλλά και την αξιολόγηση αυτού μέσω ανταγωνιστικών επιθέσεων και αμυνών και μετρήσεων.

Σε αυτές τα εργαλεία και βιβλιοθήκες είναι συνήθως υλοποιημένες οι πιο ισχυρές adversarial άμυνες για την εκπαίδευση των μοντέλων ή για την ενσωμάτωση τους σε αυτά αν πρόκειται για τεχνικές που δεν αναγκάζουν σε επανεκπαίδευση του μοντέλου, οι πιο ισχυρές ανταγωνιστικές επιθέσεις για τον έλεγχο και τη δοκιμή του empirical robustness ενάντια σε μοντέλα, άμυνες αλλά και μετρήσεις για τη συγκριτική αξιολόγηση της ευρωστίας. Αυτές οι βιβλιοθήκες, επικεντρώνονται κυρίως σε μεθόδους επίτευξης και αξιολόγησης empirical robustness, καθώς είναι πολύ πιο εύκολα επιτεύξιμο χρονικά και υπολογιστικά για τα μεγάλα μοντέλα και σύνολα δεδομένων που χρησιμοποιούνται πλέον στις εφαρμογές ML. Αντίστοιχα, οι περισσότερες τεχνικές εφαρμόζονται σε μοντέλα βαθιάς μάθησης και σε τομείς εικόνας, φωνής και κειμένου καθώς πρόκειται για τις πιο συχνές εφαρμογές της τεχνητής νοημοσύνης.

Παρακάτω παρουσιάζουμε μερικά από τα πιο γνωστά και ευρέως χρησιμοποιούμενα εργαλεία για την πρακτική επίτευξη και αξιολόγηση ευρωστίας. Στον πίνακα 4.2 υπάρχει μια σύνοψη όλων των εργαλείων και βιβλιοθηκών που παρουσιάζουμε, με αναφορά των κύριων χαρακτηριστικών τους.

Βιβλιοθήκη	ML Frameworks	Πεδία	Επιθέσεις	Άμυνες	Metrics
ART	TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy	images, tables, audio, video	Evasion, Poisoing, Extraction, Inference	Preprocessing, Training, Transforming, Detection	Ναι
CleverHans	TensorFlow 2, PyTorch, JAX	images, tables	Evasion	Adversarial Training	Όχι
FoolBox	TensorFlow 2, PyTorch, JAX	images, tables	Gradient-based, Decision-based	Όχι	Όχι
AdverTorch	PyTorch	images, tables	Gradient-based, other	Adversarial Training	Όχι
AugLy	*Framework-Agnostic*	image, audio, video, text	Όχι	Data Augmentation	Όχι
TexAttack	*Framework-Agnostic*	text	NLP Attacks	Adversarial Training, Data Augmentation	Ναι

Πίνακας 4.2: *Εργαλεία και βιβλιοθήκες ανοιχτού κώδικα για εργασίες σχετικές με adversarial robustness.*

4.8.1 Adversarial Robustness Toolbox (ART)

Το **Adversarial Robustness Toolbox (ART)** [42] είναι μια Python βιβλιοθήκη που παρέχει εργαλεία για την αξιολόγηση, άμυνα και επαλήθευση μοντέλων και εφαρμογών ML ενάντια σε ανταγωνιστικές επιθέσεις (π.χ. evasion, poisoning επιθέσεις). Μπορεί να χρησιμοποιηθεί από blue teams και red teams για τη δημιουργία πιο ασφαλών μοντέλων και τον έλεγχο αυτής της ασφάλειας αντίστοιχα.

Πιο συγκεκριμένα, υποστηρίζει τα περισσότερα **ML frameworks** (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy), τους περισσότερους

τύπους δεδομένων (εικόνες, πίνακες, ήχος, βίντεο) και πολλές ML διεργασίες (ταξινόμηση, ανίχνευση αντικειμένων, αναγνώριση ομιλίας, generation, certification).

Επίσης, περιέχει διάφορα είδη και πλήθος **επιθέσεων**, όπως evasion white-box επιθέσεις (π.χ. FGSM, PGD, DeepFool, JSMA), evasion black-box επιθέσεις (π.χ. Square Attack, Pixel Attack), poisoning επιθέσεις (π.χ. Backdoor Attack). Ακόμη, περιέχει διάφορα είδη και πλήθος **αμυνών**, όπως preprocessing τεχνικές (π.χ. JPEG Compression, Label Smoothing, PixelDefend), τεχνικές εκπαίδευσης (π.χ. Adversarial Training, Certified Adversarial Training), και ανίχνευσης (π.χ. Basic detector based on inputs, Detection based on activations analysis). Τέλος, περιέχει πλήθος **metrics** (π.χ. Empirical Robustness, Loss Sensitivity, CLEVER) για την αξιολόγηση του robustness.

Πηγαίος κώδικας: https://github.com/Trusted-AI/adversarial-robustness-toolbox

4.8.2 CleverHans

Το **CleverHans** [129] είναι μια Python βιβλιοθήκη για τη διεξαγωγή benchmarks σε ML συστήματα για τον έλεγχο ευρωστίας ενάντια σε adversarial examples. Περιέχει τυποποιημένες υλοποιήσεις ανταγωνιστικών επιθέσεων και ανταγωνιστικής εκπαίδευσης και προσφέρεται κυρίως για την ανάπτυξη πιο εύρωστων ML μοντέλων μέσω της παροχής τυποποιημένων benchmarks για την απόδοσης των μοντέλων σε adversarial περιβάλλοντα. Χωρίς την τυποποιημένη και καλή υλοποίηση των ανταγωνιστικών επιθέσεων, ένα όχι ασφαλές μοντέλο μπορεί να δείξει καλές επιδόσεις.

Το CleverHans (από την έκδοση v4.0.0 και μετά) υποστηρίζει 3 από τα πιο δημοφιλή **ML frameworks** (TensorFlow 2, PyTorch, JAX) και περιέχει πλήθος state-of-the-art **επιθέσεων**, κυρίως evasion white-box επιθέσεων (π.χ. FGSM, PGD, C&W), αλλά περιέχει ελάχιστες **άμυνες**, όπως Adversarial Training. Επίσης, παρέχει πλήθος παραδειγμάτων και tutorials για την καλύτερη εξοικείωση με τη χρήση της και με τη δημιουργία τυποποιημένων benchmarks

Πηγαίος κώδικας: https://github.com/cleverhans-lab/cleverhans

4.8.3 Foolbox

Το **Foolbox** [130] είναι μια Python βιβλιοθήκη που επιτρέπει την εύκολη εκτέλεση ανταγωνιστικών επιθέσεων ενάντια σε ML μοντέλα για τη συγκριτική αξιολόγηση της ευρωστίας αυτών. Είναι framework-agnostic που σημαίνει ότι είναι κατάλληλο εργαλείο για τη σύγκριση ευρωστίας μεταξύ πολλών και διαφορετικών μοντέλων υλοποιημένων σε διαφορετικά ML framworks.

Το Foolbox (από την έκδοση 3, με την ονομασία Foolbox Native, και μετά) υποστηρίζει 3 από τα πιο δημοφιλή **ML frameworks** (TensorFlow 2, JAX), και υποστηρίζει κυρίως ML διεργασίες ταξινόμησης σε εφαρμογές εικόνας.

Επίσης, περιέχει μια πλήρη συλλογή state-of-the-art **επιθέσεων**, κυρίως gradient-based επιθέσεων (π.χ. FGSM, PGD, EAD, DeepFool) και decision-based επιθέσεων (π.χ. BoundaryAttack, PointwiseAttack, HopSkipJump). Δεν παρέχεται καμία άμυνα στη βιβλιοθήκη, αλλά παρέχεται μια πληθώρα κριτηρίων (criteria) και μέτρων απόστασης (distance measures) τα οποία μπορούν να χρησιμοποιηθούν για να ορίσουν κάτω υπό ποιες συνθήκες ένα δείγμα είναι adversarial, και για την ποσοτικοποίηση του μεγέθους του adversarial perturbation αντίστοιχα.

Πηγαίος κώδικας: https://github.com/bethgelab/foolbox

4.8.4 AdverTorch

Το **AdverTorch** [131] είναι μια Python βιβλιοθήκη για έρευνα adversarial robustness, παρέχοντας εργαλεία για τη δημιουργία adversarial examples, αλλά και για την άμυνα από τέτοιες ανταγωνιστικές επιθέσεις. Η βιβλιοθήκη είναι χτισμένη πάνω στο PyTorch ML framework, αξιοποιώντας τα πλεονεκτήματα του δυναμικού υπολογιστικού γράφου (dynamic computational graph) που υπάρχει στο framework, για να παρέχει συνοπτικές και αποδοτικές υλοποιήσεις.

Επίσης, περιέχει μια πλήρη συλλογή state-of-the-art **επιθέσεων**, κυρίως gradient-based επιθέσεων (π.χ. FGSM, PGD, C&W, Spatial Transformation) αλλά και άλλων κατηγοριών (π.χ. SinglePixelAttack, LocalSearchAttack). Ακόμη, περιέχει **άμυνες**, που χωρίζονται σε preprocessing τεχνικές (π.χ. JPEG Filtering, Bit Squeezing, Gaussian Smoothing), και τεχνικές εκπαίδευσης (π.χ. Adversarial Training).

Πηγαίος κώδικας: https://github.com/BorealisAI/advertorch

4.8.5 AugLy

Το **AugLy** [132] είναι μια Python βιβλιοθήκη που παρέχει στους ερευνητές της τεχνητής νοημοσύνης τεχνικές επαύξησης δεδομένων (**data augmentation**) για να αξιολογήσουν και να βελτιώσουν την ευρωστία των ML μοντέλων τους. Αυτήν τη στιγμή προσφέρονται 4 υπό-βιβλιοθήκες με παραπάνω από 100 διαφορετικές μεθόδους για επαυξήσεις δεδομένων εικόνας, ήχου, βίντεο και κειμένου, με χρήση κατάλληλων παραμέτρων για τον ορισμό του μεγέθους των μετασχηματισμών. Οι επαυξήσεις αυτές περιλαμβάνουν μια μεγάλη ποικιλία τροποποιήσεων, όπως περικοπή φωτογραφιών, αλλαγή τόνου της φωνής, προσθήκη κειμένου ή emoji πάνω σε φωτογραφίες καθώς και οτιδήποτε μπορεί να κάνουν οι άνθρωποι στο διαδίκτυο και στα μέσα κοινωνικής δικτύωσης, καθώς ένας από τους κύριους σκοπούς ανάπτυξης της συγκεκριμένης βιβλιοθήκης ήταν ο εντοπισμός αντιγράφων ή διπλοτύπων περιεχομένων.

Είναι model-agnostic και framework-agnostic που σημαίνει ότι μπορεί να χρησιμοποιηθεί σε οποιοδήποτε μοντέλο φτιαγμένο με οποιαδήποτε ML framework, αφού προσφέρεται μόνο για επαύξηση των δεδομένων ενός συνόλου δεδομένων και δεν επεμβαίνει στα ίδια τα μοντέλα.

Επίσης, περιέχει μια πλήρη συλλογή πάνω από 100 **μετασχηματισμών και επαυξήσεων**, από υπάρχουσες βιβλιοθήκες αλλά και καινούριες που δεν υπήρχαν προηγουμένως, για εικόνες (π.χ. περικοπή, θόλωμα, αλλαγή φωτεινότητας), ήχο (π.χ. αλλαγή συχνότητας ή έντασης, background noise), βίντεο (π.χ. αλλαγή ταχύτητας ή ανάλυσης), κείμενο (π.χ. αντικατάσταση λέξεων με συνώνυμες, αντικατάσταση χαρακτήρων με κοντινούς στο πληκτρολόγιο). Αυτές οι επαυξήσεις μπορούν να χρησιμοποιηθούν για την εκπαίδευση των μοντέλων, αλλά και στη δημιουργία ανταγωνιστικών επιθέσεων για την αξιολόγηση την ευρωστίας των μοντέλων σε τυχαίες διαταραχές οι οποίες δεν αλλάζουν τα κύρια χαρακτηριστικά των δεδομένων.

Πηγαίος κώδικας: https://github.com/facebookresearch/AugLy

4.8.6 TextAttack

Το **TextAttack** [133] είναι μια Python βιβλιοθήκη για τη διεξαγωγή adversarial attacks, την προσαύξηση δεδομένων (data augmentation) και την εκπαίδευση μοντέλων στον NLP

(Natural language processing) τομέα της μηχανικής μάθησης. Η αξιολόγηση και επίτευξη **adversarial robustness** σε **NLP deep learning models**, έχει αρκετές διαφορές σε σχέση με τις εφαρμογές εικόνας που έχουμε αναλύσει κυρίως. Το TextAttack συγκεντρώνει και υλοποιεί συστηματικά τις καλύτερες NLP ανταγωνιστικές επιθέσεις με χρήση ενός συστήματος 4 συνιστωσών: (i) μιας συνάρτησης στόχου που καθορίζει αν η επίθεση πέτυχε, (ii) με περιορισμούς που ορίζουν ποιες διαταραχές είναι έγκυρες, (iii) ενός μετασχηματισμού που δημιουργεί πιθανές τροποποιήσεις για κάθε είσοδο, (iv) μιας μεθόδου αναζήτησης που διασχίζει το χώρο αναζήτησης των πιθανών διαταραχών. Το TextAttack προσφέρεται κυρίως για την ανταγωνιστική εκπαίδευση εύρωστων NLP μοντέλων, την αξιολόγηση αυτών αλλά και την ανάπτυξη και εύκολη χρήση νέων επιθέσεων για την ανακάλυψη νέων ευπαθειών στα NLP ML μοντέλα, με συνδυασμό υπαρχόντων και νέων στοιχείων.

Είναι model-agnostic και framework-agnostic που σημαίνει ότι μπορεί να χρησιμοποιηθεί σε οποιοδήποτε μοντέλο φτιαγμένο με οποιαδήποτε ML framework, αρκεί να οποίο εξάγει IDs, tensors ή strings. Υπάρχουν ήδη προεκπαιδευμένα μοντέλα για διάφορες NLP εργασίες και αντίστοιχα κατάλληλοι wrappers για διάφορα ML frameworks και αποθετήρια μοντέλων (π.χ. Tensorflow, PyTorch, Scikit-Learn, Hugging Face), για την πιο εύκολη χρήση του εργαλείου και για δίκαιες σύγκριση μεταξύ επιθέσεων σε μοντέλα.

Το TextAttack περιέχει συλλογή 16 state-of-the-art adversarial NLP **επιθέσεων** (π.χ. BERT Attack), αλλά και τρόπους για συνδυασμούς αυτών. Παρέχει επίσης, μεθόδους για data augmentation μέσω κατάλληλων μετασχηματισμών (π.χ. τυχαία αλλαγή λέξεων με συνώνυμες, τυχαία διαγραφή λέξεων). Ακόμη, παρέχει τη δυνατότητα για **εκπαίδευση** του μοντέλου (π.χ. Adversarial Training), και αντίστοιχα **αξιολόγηση** της ευρωστίας του μοντέλου.

Πηγαίος κώδικας: https://github.com/QData/TextAttack

Κεφάλαιο 5

Ιδιωτικότητα στη Μηχανική Μάθηση - Επιθέσεις και Άμυνες

Στο κεφάλαιο αυτό περιγράφεται ορίζεται η έννοια της ιδιωτικότητας (privacy) στα μοντέλα μηχανικής μάθησης και αναλύονται οι επιθέσεις ιδιωτικότητας ενάντια σε μοντέλα μηχανικής μάθησης και τα δεδομένα τους καθώς και τεχνικές άμυνας ενάντια σε τέτοιες επιθέσεις. Επίσης, γίνεται μια αναφορά σε τεχνικές για τη γενικότερη προστασία και διατήρηση της ιδιωτικότητας των δεδομένων που μπορούν να εφαρμοστούν στα μοντέλα μηχανικής μάθησης.

5.1 Ιδιωτικότητα στη Μηχανική Μάθηση

Ως τώρα αναλύσαμε την ευρωστία (robustness) των ML μοντέλων ενάντια σε ανταγωνιστικές επιθέσεις που στοχεύουν στην παραβίαση της ακεραιότητας (integrity) ή της διαθεσιμότητας (availability) του συστήματος. Και οι δύο αυτοί τύποι επιθέσεων στοχεύουν στο να οδηγήσουν τα μοντέλα σε εσφαλμένη ταξινόμηση ή να μειώσουν αξιοπιστία του συστήματος, αυξάνοντας την αβεβαιότητα των προβλέψεών τους. Όμως, όπως αναφέραμε και στην ενότητα 3.4.1 για την ταξινόμηση των ανταγωνιστικών επιθέσεων, ένας άλλος στόχος των ανταγωνιστικών επιθέσεων είναι η και παραβίαση της ιδιωτικότητας (privacy) του ML συστήματος, όπου οι επιτιθέμενοι αποσκοπούν στο να αποκτηθούν προσωπικές πληροφορίες σχετικά με το σύστημα, τους χρήστες ή δεδομένα του [27]. Σε τομείς όπου το σύνολο των δεδομένων εκπαίδευσης περιέχει ευαίσθητες πληροφορίες που πρέπει να διατηρηθούν απόρρητες, όπως σε τομείς ιατρικών εφαρμογών, κρίνεται επιτακτική η ανάγκη προστασίας από τέτοιες επιθέσεις ιδιωτικότητας (**privacy attacks**) [4].

Πολλές φορές, τα ML μοντέλα μπορεί να θεωρηθούν εμπιστευτικά (confidential) λόγω ευαίσθητων δεδομένων εκπαίδευσης, της εμπορικής τους αξίας, της πνευματικής τους ιδιοκτησίας ή της εφαρμογής τους σε τομείς σχετιζόμενους με την ασφάλεια. Αυτά τα μοντέλα, μπορεί να είναι διαθέσιμα στο εύρη κοινό ή και διαθέσιμα προσβάσιμη για χρήση και άσκηση ερωτημάτων (π.χ. public API ή chatbots) [134]. Αυτό κάνει ακόμα πιο επιτακτική την ανάγκη τα ML μοντέλα να σχεδιάζονται να είναι και εύρωστο ενάντια σε επιθέσεις ιδιωτικότητας. Τα εύρωστα ML μοντέλα, δεν μπορούν να εγγυηθούν για την ασφάλεια της ιδιωτικότητας του μοντέλου ή των δεδομένων του, οπότε πρέπει να οριστούν νέες έννοιες για να αξιολογηθεί η ασφάλεια της ιδιωτικότητας (π.χ. privacy preserving guarantees), αλλά και νέοι τρόποι αντιμετώπισης τέτοιων επιθέσεων.

Ένα μεγάλο κομμάτι της βιβλιογραφίας ασχολείται με την προστασία της ιδιωτικότητας, είτε για την ανάπτυξη επιθέσεων που θα την παραβιάσουν (βλ. 5.2 για Επιθέσεις Ιδιωτικότητας), είτε για αμυντικές τεχνικές που θα την προστατέψουν, θα ανιχνεύσουν τις επιθέσεις ή

θα αποκρύψουν τις κατάλληλες πληροφορίες για την αποφυγή των επιθέσεων εξαρχής (βλ. 5.3 για Προστασία Ιδιωτικότητας). Πάντως, ανεξάρτητα από τον στόχο, οι περισσότερες επιθέσεις και άμυνες ιδιωτικότητας, σχετίζονται με την έκθεση ή την αποτροπή έκθεσης του μοντέλων και των δεδομένων εκπαίδευσης [4].

Όταν αναφερόμαστε στην ιδιωτικότητα στη μηχανική μάθηση γενικότερα, κάνουμε τον διαχωρισμό μεταξύ **ιδιωτικότητας μοντέλου** και **ιδιωτικότητας των δεδομένων**, όπου αναλύουμε το καθένα στις παρακάτω ενότητες.

Τέλος, σε αυτό το κεφάλαιο δεν αναφερόμαστε σε θέματα γενικότερης προστασίας των προσωπικών δεδομένων των χρηστών, είτε για την απόκτηση αυτών των δεδομένων, είτε για την αποθήκευση τους ή διακίνηση και κοινοποίηση τους, καθώς εμπίπτουν σε γενικότερα θέματα ασφάλειας των προσωπικών δεδομένων και προστασίας τις ιδιωτικότητας, που δε σχετίζονται με τα ML μοντέλα. Η αναφορά μας γίνεται αυστηρά στην ευαλωτότητα των ML μοντέλων να διατηρήσουν την προστασία των δεδομένων τους και επιθέσεις που εκμεταλλεύονται αυτές τις εγγενείς ευπάθειες.

5.1.1 Ιδιωτικότητα Μοντέλου

Εκτός από δεδομένα χρηστών, ακόμα και τα ίδια τα ML μοντέλα υπόκεινται σε επιθέσεις παραβίασης της ιδιωτικότητας τους, όπου γίνεται προσπάθειά εξαγωγής των χαρακτηριστικών τους, της αρχιτεκτονικής τους, τις παραμέτρους τους και οποιαδήποτε άλλη λεπτομέρεια του μοντέλου μπορεί να εξαχθεί. Όσο οι εφαρμογές μηχανικής μάθησης γίνονται όλο και πιο διαδεδομένες και διαθέσιμες σε μεγαλύτερο κοινό, η προστασία της ιδιωτικότητας των ML μοντέλων γίνεται πιο σημαντική κυρίως για δύο λόγους: (i) τα μοντέλα μπορεί να αποτελούν επιχειρηματικό πλεονέκτημα για τον ιδιοκτήτη του και (ii) ένας επιτιθέμενος μπορεί να χρησιμοποιήσει ένα κλεμμένο μοντέλο για να βρει adversarial examples τα οποία θα μπορούν να μεταφερθούν στο αρχικό μοντέλο για να προκαλέσουν σφάλματα στην ταξινόμηση ή πρόβλεψη [135]. Συνήθως, οι επιθέσεις αυτές που στοχεύουν στο να εξάγουν ευαίσθητες πληροφορίες για τα ML μοντέλα (βλ. **Model Extraction** 5.2.1), συμβαίνουν σε ένα black-box περιβάλλον, όπου οι επιτιθέμενοι δε γνωρίζουν καμία λεπτομέρεια για το σύστημα, και έχουν πρόσβαση στο μοντέλο μόνο μέσω κάποιες υπηρεσίας ή διεπαφής (π.χ. ML-as-a-service υπηρεσία με διαθέσιμο API) [134].

5.1.2 Ιδιωτικότητα Δεδομένων

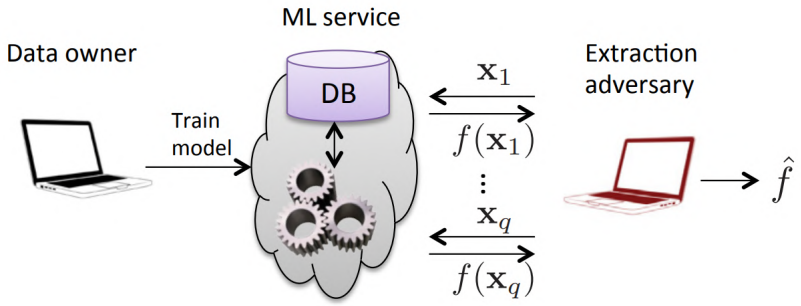
Τα δεδομένα είναι ζωτικής σημασίας για την καλή και ακριβής λειτουργία των ML μοντέλων, όμως πολλές φορές μπορεί να περιέχουν ευαίσθητες πληροφορίες για άτομα, όπως προσωπικά στοιχεία, συνήθειες και προτιμήσεις, κατάσταση υγείας κ.λπ. Επίσης, τα ML μοντέλα έχουν την ιδιότητα να συλλαμβάνουν και να απομνημονεύουν στοιχεία των δεδομένων εκπαίδευσης, οπότε είναι δύσκολο να δοθούν εγγυήσεις ότι η συμμετοχή ενός ατόμου σε ένα σύνολο δεδομένων δε βλάπτει την ιδιωτικότητα του, καθώς έχουν υπάρξει επιθέσεις οι οποίες μπορούν να μάθουν εάν ένα άτομο ή δείγμα βρίσκεται σε ένα σύνολο δεδομένων ή όχι (βλ. **Membership Inference Attack** 5.2.3), ή ακόμα με την κατάλληλη είσοδο να εξάγουν πληροφορίες για τα δεδομένα εκπαίδευσης (βλ. **Model Inversion** 5.2.2).

5.2 Επιθέσεις Ιδιωτικότητας

Όπως αναφέραμε ήδη, οι επιθέσεις ιδιωτικότητας (ή απορρήτου) ενάντια στα ML συστήματα, προσπαθούν να εξάγουν πληροφορίες είτε για το μοντέλο είτε για τα δεδομένα εκπαίδευσης. Σε αυτήν την ενότητα περιγράφουμε 3 δημοφιλείς κατηγορίες επιθέσεων για την παραβίαση της ιδιωτικότητας είτε του μοντέλου είτε των δεδομένων. Αυτές οι επιθέσεις μπορεί να έχουν εφαρμογή σε διάφορα ML μοντέλα, όπως νευρωνικά δίκτυα ή δέντρα αποφάσεων [134], και υποθέτουν black-box μοντέλο απειλής, δηλαδή καμία προηγούμενη γνώση για τις παραμέτρους των ML μοντέλων ή των δεδομένων εκπαίδευσης.

5.2.1 Model Extraction

Στις επιθέσεις εξαγωγής μοντέλου ή κλοπής μοντέλου (**Model Extraction** ή **Model Stealing**), οι επιτιθέμενοι προσπαθούν να αναδημιουργήσουν το υποκείμενο μοντέλο υποβάλλοντας τα κατάλληλα ερωτήματα (queries) στο μοντέλο, ώστε η λειτουργικότητα του νέου μοντέλου να είναι ίδια με αυτή του υποκείμενου μοντέλου. Αφού το μοντέλο κλαπεί και αντιγραφεί, μπορεί να αναστραφεί για να ανακτήσει πληροφορίες για τα χαρακτηριστικά του ή να κάνει συμπεράσματα σχετικά με τα δεδομένα εκπαίδευσης [43] (βλ. σχήμα 5.1).



Σχήμα 5.1: *Διάγραμμα της model extraction επίθεσης ενάντια σε ένα ML μοντέλο. Το μοντέλο f εκπαιδεύεται σε ένα σύνολο δεδομένων και επιτρέπει σε άλλους να κάνουν ερωτήματα προβλέψεων. Ένας επιτιθέμενος χρησιμοποιεί q ερωτήματα προβλέψεων για να εξαγάγει ένα μοντέλο $\hat{f} \approx f$ [134].*

Τεχνικά, δεν απαιτείται κάποιο ειδικό προνόμιο για να διεξαχθεί μια τέτοια επίθεση, καθώς ένας χρήστης στέλνει ή εισάγει ειδικά σχεδιασμένα ερωτήματα σε ένα ML σύστημα, όπως κάθε νόμιμος χρήστης [43]. Οι στόχοι των επιτιθεμένων για την κλοπή ενός μοντέλου, μπορεί να είναι πολλοί, όπως (i) σε ML μοντέλα ανίχνευσης επιθέσεων (π.χ. ανίχνευση malware, spam, anomaly detection), η εξαγωγή του μοντέλου μπορεί να διευκολύνει την εκτέλεση επιτυχημένων ανταγωνιστικών επιθέσεων [43], (ii) σε εμπορικά ML μοντέλα, η δημιουργία ενός αντίγραφου μπορεί να προσφέρει επιχειρηματικό πλεονέκτημα [135].

Οι μέθοδοι που χρησιμοποιούνται στις περισσότερες επιθέσεις σε αυτήν την κατηγορία είναι οι εξής:

- Επίλυση εξισώσεων (**Equation solving**): Όταν ένα μοντέλο επιστρέφει πιθανότητες κλάσεις στην έξοδο του, τότε μπορούν να δημιουργηθούν ερωτήματα για να προσδιοριστούν οι άγνωστες μεταβλητές του μοντέλου [43]
- Εύρεση διαδρομής (**Path finding**): Επίθεση η οποία εκμεταλλεύεται τις εξόδους του μοντέλου για να εξάγει τις αποφάσεις που λαμβάνονται από ένα δέντρο αποφάσεων κατά

την ταξινόμηση μιας εισόδου (εφαρμόσιμο για decision trees και regression trees) [134]

- Επίθεση μεταφοράς (**Transferability attack**): Εκπαίδευση τοπικού αντίγραφου μοντέλου (substitute model) μέσω εξόδων από ερωτήσεις στο αρχικό μοντέλο, με στόχο τη δημιουργία adversarial examples τα οποία μπορούν να μεταφερθούν στο αρχικό μοντέλο [136]

Για την επιτυχία του στόχου της εξαγωγής ενός μοντέλου, υπάρχουν κάποιες μετρήσεις (**metrics**) οι οποίες χρησιμοποιούνται για να αξιολογηθεί η *αποτελεσματικότητα μιας model extraction επίθεσης* [137]:

- **Αποτελεσματικότητα** (**Effectiveness**): Για τις επιθέσεις που επιχειρούν να εξάγουν τα χαρακτηριστικά του μοντέλου, μετριέται το κατά πόσον κοντά ή ίσες ήταν οι τιμές των χαρακτηριστικών που εξήχθησαν. Για τις επιθέσεις που επιχειρούν να αντιγράψουν τη συμπεριφορά του μοντέλου, μετριέται συνήθως η ακρίβεια (accuracy), η πιστότητα (fidelity) και η δυνατότητα μεταφοράς (transferability) του εξαγόμενου μοντέλου.
- **Αποδοτικότητα** (**Efficiency**): Μετριέται συνήθως ο αριθμός των ερωτημάτων (query budget) και ο χρόνος (timing) που χρειάζεται για να αντιγραφούν οι παράμετροι του μοντέλου.

Ένα παράδειγμα εφαρμογής **equation solving** επίθεσης για εξαγωγή μοντέλου είναι το [134], το οποίο εστιάζει στην εξαγωγή των παραμέτρων του μοντέλου με χρήση των πιθανοτήτων πρόβλεψης που επιστρέφουν τα API των μοντέλων σε κάθε πρόβλεψη. Η επίθεση είναι εννοιολογικά απλή και εφαρμόζει επίλυση εξισώσεων για να ανακτήσει τις παραμέτρους θ από τα σύνολα των παρατηρούμενων ζευγαριών εισόδου-εξόδου $(x, h_\theta(x))$. Όμως, αυτή η προσέγγιση δεν κλιμακώνεται σε σενάρια όπου ο αντίπαλος χάνει την πρόσβαση στις πιθανότητες που επιστρέφονται για κάθε κλάση, δηλαδή όταν έχει πρόσβαση μόνο στις ετικέτες, οπότε και δεν μπορεί να έχει μεγάλη πρακτική εφαρμογή.

Ένα άλλο παράδειγμα εφαρμογής **transferability** επίθεσης με χρήση substitute model για εξαγωγή μοντέλου είναι το [72], στο οποίο προτείνονται 2 υποκατάστατα μοντέλα, ένα πιο σύνθετο DNN και ένα απλούστερο Logistic Regression, για την κλοπή διαφόρων ML μοντέλων, όπως DNN, SVM, Decision Trees, kNN. Ο κύριος στόχος είναι η εκπαίδευση ενός μοντέλου με όριο απόφασης παρόμοιο με το αρχικό, όσον αφορά την τη δυνατότητα μεταφοράς, δηλαδή η προσέγγιση του αρχικού μοντέλου να επιτρέπει τη δημιουργία adversarial examples τα οποία με μεγάλη πιθανότητα θα ξεγελάνε το αρχικό μοντέλο. Έτσι το substitute model μπορεί να χρησιμοποιηθεί για να πραγματοποιηθούν επιθέσεις στο αρχικό μοντέλο.

5.2.2 Model Inversion

Στις επιθέσεις αντιστροφής μοντέλου (**Model Inversion**), μπορούν να ανακτηθούν οι ιδιωτικές λειτουργίες που χρησιμοποιούνται στα ML μοντέλα, μέσω προσεκτικά δημιουργημένων ερωτημάτων. Αυτό περιλαμβάνει την ανακατασκευή ιδιωτικών δεδομένων εκπαίδευσης στα οποία κανονικά δεν υπάρχει πρόσβαση από τους επιτιθέμενους. Αυτές οι επιθέσεις είναι γνωστές ως **hill climbing attacks** στον τομέα της βιομετρικών συστημάτων [138]. Αυτό μπορεί να επιτευχθεί με την εύρεση εισόδου η οποία μεγιστοποιεί το επιστρεφόμενο επίπεδο εμπιστοσύνης σε σχέση με την ταξινόμηση που αντιστοιχεί στον στόχο [43]. Οι στόχοι τώρα των επιτιθεμένων για την αντιστροφή του μοντέλου είναι κυρίως η εξαγωγή προσωπικών

πληροφοριών ή αποκάλυψη εμπιστευτικών δεδομένων.

Όπως και στην περίπτωση των model extraction επιθέσεων, τεχνικά δεν απαιτείται κάποιο ειδικό προνόμιο για να διεξαχθεί μια τέτοια επίθεση, καθώς ένας χρήστης στέλνει ή εισάγει ειδικά σχεδιασμένα ερωτήματα σε ένα ML σύστημα, όπως κάθε νόμιμος χρήστης, και παρατηρεί τη συμπεριφορά των προβλέψεων του μοντέλου για να εξάγει ιδιωτικά χαρακτηριστικά και πληροφορίες [43]. Ωστόσο, η είσοδος που εξάγεται δεν είναι στην πραγματικότητα ένα συγκεκριμένο σημείο του συνόλου δεδομένων εκπαίδευσης, αλλά μια μέση αναπαράσταση των εισόδων που ταξινομούνται σε μια κλάση, παρόμοια με αυτό που γίνεται από τους χάρτες προεξοχής (saliency maps) [4].

Το πρώτο παράδειγμα model inversion για εξαγωγή δεδομένων εκπαίδευση, προέρχεται από τον χώρο της ιατρικής, όπου για μια εργασία πρόβλεψης δόσης φαρμάκου, δείχνεται ότι απλά με την πρόσβαση στο μοντέλο και σε βοηθητικές πληροφορίες σχετικά με τη σταθερή δόση φαρμάκου από τον ασθενή, μπορούν να ανακτηθούν γονιδιωματικές (genomic) πληροφορίες για τον ασθενή [139]. Αν και η προσέγγιση απεικονίζει ανησυχίες για την παραβίαση της ιδιωτικότητας που μπορεί να προκύψει από την παροχή πρόσβασης σε ML μοντέλα που έχουν εκπαιδευτεί σε ευαίσθητα δεδομένα, δεν είναι σαφές εάν οι γονιδιωματικές πληροφορίες ανακτώνται λόγω ευπάθειας του ML μοντέλου ή της ισχυρής συσχέτισης μεταξύ των βοηθητικών πληροφοριών στις οποίες έχει επίσης πρόσβαση ο επιτιθέμενος (τη δόση του ασθενούς) [4].

Ένα παράδειγμα διαφορετικής model inversion επίθεσης ενάντια σε ένα ML σύστημα αναγνώρισης προσώπων φαίνεται στην εικόνα 5.2, όπου ένας επιτιθέμενος μπορεί να δημιουργήσει μια αναγνωρίσιμη εικόνα ενός ανθρώπου από το σύνολο δεδομένων, έχοντας στην κατοχή του μόνο πρόσβαση και δυνατότητα άσκησης ερωτημάτων στο σύστημα και το όνομα του ανθρώπου για το οποίο αναγνωρίζεται το πρόσωπό του [140].

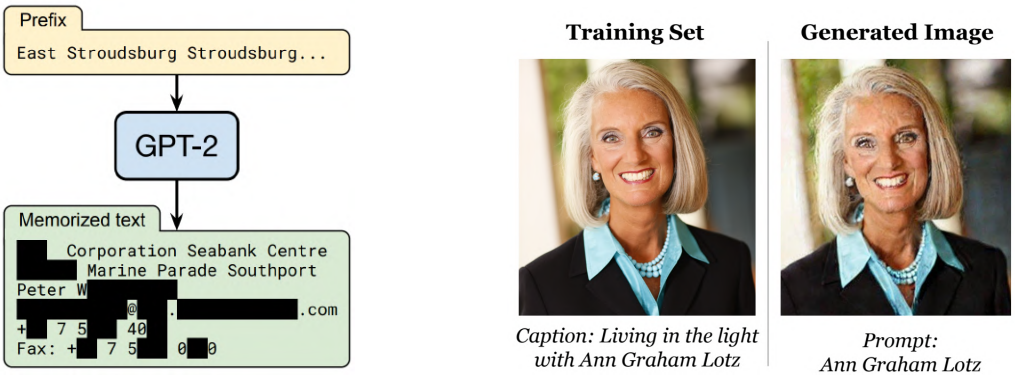
Σχήμα 5.2: *Παράδειγμα εικόνας που ανακτήθηκε χρησιμοποιώντας μια επίθεση model inversion (αριστερά) και μια εικόνα του συνόλου δεδομένων εκπαίδευσης του μοντέλου στόχος (δεξιά). Στον επιτιθέμενο έχει δοθεί μόνο το όνομα του ατόμου και πρόσβαση σε ένα σύστημα αναγνώρισης προσώπου που επιστρέφει βαθμό εμπιστοσύνης για κάθε κλάση [140].*

Οι περισσότερες model inversion επιθέσεις δουλεύουν σε **white-box σενάριο επιθέσεων**, καθώς είναι υπολογιστικά πολύ πιο δύσκολο να γίνουν οι υπολογισμοί των κλίσεων που απαιτούνται για την ανακατασκευή των δεδομένων σε black-box σενάριο [140]. Μια από τις ελάχιστες προτεινόμενες **black-box επιθέσεις ανακατασκευής**, χρησιμοποιεί έναν επιπλέον ταξινομητή για να εκτελεί την αντιστροφή από την έξοδο του αρχικού μοντέλου σε

μια υποψήφια έξοδο, κάπως παρόμοιο με τη λειτουργία των autoencoders. Όταν είναι διαθέσιμο το πλήρες διάνυσμα πρόβλεψης, η επίθεση εκτελεί καλή ανακατασκευή, αλλά με λιγότερες διαθέσιμες πληροφορίες, το παραγόμενο σημείο δεδομένων μοιάζει περισσότερο απλά με ένα αντιπροσωπευτικό δείγμα της κλάσης [2].

Επίσης, έχουν προταθεί αρκετές διαφορετικές τεχνικές για την επίτευξη της ανακατασκευής εισόδου, που περιλαμβάνουν: (i) δεδομένου ότι υπάρχει πρόσβαση στο μοντέλο και είναι γνωστά τα ευαίσθητα και μη χαρακτηριστικά του, η επίθεση περιλαμβάνει την εκτίμηση των τιμών των ευαίσθητων χαρακτηριστικών, δεδομένων των τιμών των μη ευαίσθητων χαρακτηριστικών και της ετικέτας εξόδου, (ii) τη χρήση των ετικετών στόχων και βοηθητικών λεπτομερειών για την ανακατασκευή της εισόδου λύνοντας ένα πρόβλημα βελτιστοποίησης (π.χ. με χρήση gradient descent), (iii) τη χρήση GANs για την εκμάθηση βοηθητικών πληροφοριών για τα δεδομένα εκπαίδευσης και την παραγωγή καλύτερων αποτελεσμάτων [2].

Τέλος, έχουν προταθεί και πολύ πρόσφατες **model inversion** επιθέσεις για την εξαγωγή των δεδομένων εκπαίδευσης από **Large Language Models (LLMs)** [141] όπως το GPT-2, και **Diffusion Models** [142] όπως το DALL-E 2, οι οποίες κάνοντας τα κατάλληλα ερωτήματα σε ορισμένο περιβάλλον, μπορούν να εξάγουν προσωπικά δεδομένα ατόμων (ονόματα, διευθύνσεις, τηλεφωνικούς αριθμούς, διευθύνσεις email) (βλ. 5.3α′) και αντίστοιχα εικόνες ή λογότυπα που χρησιμοποιήθηκαν κατά την εκπαίδευση (βλ. 5.3β′). Το εγγενές πρόβλημα που υπάρχει και στα δύο αυτά διαφορετικά είδη μοντέλων τα οποία έχουν εφαρμογή σε δύο διαφορετικούς τομείς, είναι η τάση τους να *απομνημονεύουν μεμονωμένα δεδομένα (memorization)* είτε πρόκειται για ολόκληρες φωτογραφίες είτε ολόκληρα κομμάτια κειμένου, τα οποία με τα κατάλληλα ερωτήματα και συμφραζόμενα (prompts with context) μπορούν να τα διαρρεύσουν. Ειδικότερα, στην περίπτωση των LLMs εκφράζεται η διαπίστωση ότι όσο μεγαλύτερα είναι τα μοντέλα, τόσο περισσότερο ευάλωτα είναι σε τέτοιες επιθέσεις.



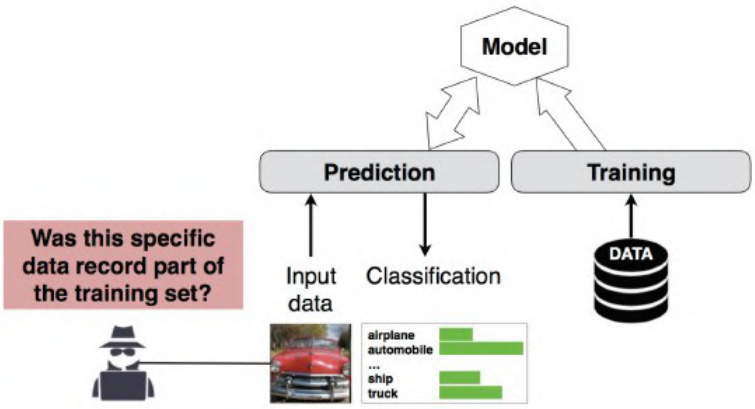
(α′) *LLM model inversion* (β′) *Diffusion model inversion*

Σχήμα 5.3: *Παράδειγμα ανάκτησης δεδομένων εκπαίδευσης από LLMs και Image Diffusion μοντέλα, όπου απομνημονεύουν ολόκληρα δεδομένα τα οποία μπορούν να εξαχθούν με το κατάλληλο prompt [45].*

5.2.3 Membership Inference Attack

Στις επιθέσεις συμπερασμάτων ιδιοτήτων μέλους (**Membership Inference Attack**), οι επιτιθέμενοι μπορούν να αναγνωρίσουν εάν ένα δεδομένο ήταν μέρος του συνόλου δεδομένων εκπαίδευσης του μοντέλου ή όχι [43]. Οι επιτιθέμενοι εκμεταλλεύονται τις πιθανότητες εξόδου

του μοντέλου για να εξάγουν συμπεράσματα για την κατάστασης μέλους ενός δεδομένου δείγματος, συγκρίνοντας τις διαφορές των προβλέψεων για δείγματα που χρησιμοποιήθηκαν έναντι εκείνων που δε συμπεριλήφθηκαν (βλ. σχήμα 5.4). Οι στόχοι των επιτιθεμένων για τις membership inference επιθέσεις είναι η εξαγωγή πληροφοριών για τα δεδομένα εκπαίδευσης, όπως π.χ. για τη γνώση αν το ιατρικό αρχείο ενός συγκεκριμένου ατόμου χρησιμοποιήθηκε για την εκπαίδευση ενός ML μοντέλου που σχετίζεται με μια συγκεκριμένη ασθένεια [143].



Σχήμα 5.4: *Σχηματικό παράδειγμα ενός membership inference attack για τη διαπίστωση ενός δείγματος αν ανήκει στο αρχικό σύνολο δεδομένων εκπαίδευσης [43].*

Τα ML μοντέλα, συχνά υπερπαραμετροποιούνται κάνοντας τα να απομνημονεύουν πληροφορίες σχετικά με τα δεδομένα εκπαίδευσης, και εκπαιδεύονται σε πεπερασμένα σύνολα δεδομένων σε πολλαπλές εποχές, με αποτέλεσμα να εμφανίζουν διαφορετική συμπεριφορά σε δείγματα που ανήκουν στα δεδομένα εκπαίδευσης (members) σε σχέση με δείγματα που δεν ανήκουν (non-member), όπως το να **ταξινομούν τα members με υψηλή βεβαιότητα** και τα **non-members με χαμηλή βεβαιότητα**, επιτρέποντας έτσι σε επιτιθέμενους να δημιουργήσουν membership inference επιθέσεις [144]. Όπως και στις δύο προηγούμενες περιπτώσεις επιθέσεων, τεχνικά δεν απαιτείται κάποιο ειδικό προνόμιο για να διεξαχθεί μια τέτοια επίθεση, καθώς ένας χρήστης στέλνει ή εισάγει ειδικά σχεδιασμένα ερωτήματα σε ένα ML σύστημα, όπως κάθε νόμιμος χρήστης, για να συμπεράνει αν το εισαγόμενο δείγμα ανήκει στα δεδομένα εκπαίδευσης [43].

Το πρώτο παράδειγμα membership inference επίθεσης, έχει εφαρμογή στον χώρο της ιατρικής, καθώς ερευνητές μπόρεσαν να προβλέψουν τη διαδικασία που ακολούθησε ένας ασθενής (π.χ. χειρουργική επέμβαση στην οποία υποβλήθηκε) με βάση τα χαρακτηριστικά του (π.χ. ηλικία, φύλο, νοσοκομείο), από ένα μοντέλο εκπαιδευμένο σε σύνολο δεδομένων νοσοκομείου που περιλαμβάνει ανώνυμες πληροφορίες για τους ασθενείς καταγράφοντας μόνο, τις εξωτερικές αιτίες τραυματισμού (π.χ. αυτοκτονία, κατάχρηση φαρμάκων), τη διάγνωση (π.χ. σχιζοφρένεια, παράνομη άμβλωση), τη διαδικασία που ακολούθησε και τα γενικά χαρακτηριστικά του [141].

Οι membership inference επιθέσεις δουλεύουν σε **white-box σενάριο επιθέσεων**, όπου η κατανομή των δεδομένων εκπαίδευσης, η αρχιτεκτονική και οι παράμετροι του μοντέλου είναι γνωστά, αλλά και σε **black-box σενάρια επιθέσεων**, όπου οι επιτιθέμενοι δε γνωρίζουν τίποτα και έχουν μόνο δυνατότητα υποβολής ερωτημάτων (εισόδων) και λήψης

προβλέψεων (εξόδων). Όμως και στις δύο περιπτώσεις η κατανομή των δεδομένων εκπαίδευσης μπορεί να θεωρείται διαθέσιμη στους επιτιθέμενους, καθώς οι επιτιθέμενοι μπορούν να αποκτήσουν ένα **shadow dataset**, το οποίο περιέχει εγγραφές δεδομένων ίδιας κατανομής με τα δεδομένα εκπαίδευσης. Αυτό είναι εφικτό, διότι το shadow dataset μπορεί να ληφθεί με σύνθεση βάσει στατιστικών όταν η κατανομή δεδομένων είναι γνωστή, και με σύνθεση βάσει μοντέλου όταν η κατανομή δεδομένων είναι άγνωστη [144].

Ανάλογα με την **προσέγγιση** (**approach**) που επιτιθέμενου, οι membership inference επιθέσεις μπορούν να κατηγοριοποιηθούν σε:

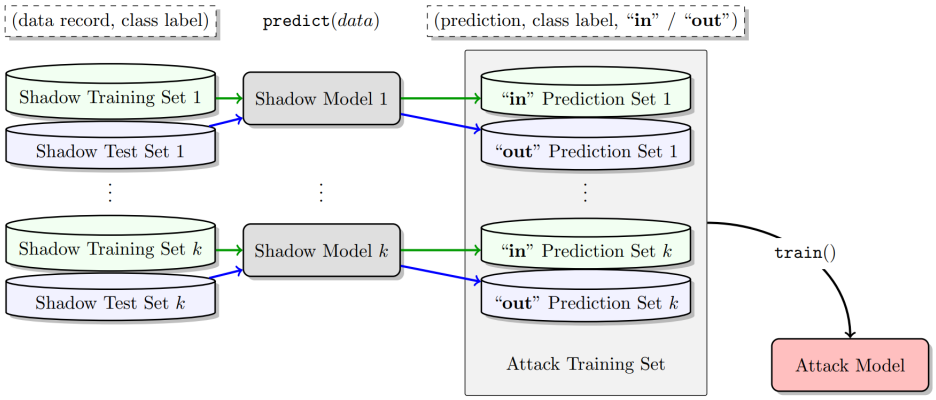
- Βασισμένες σε ταξινομητές (**Classifier Based**): Αυτές οι προσεγγίσεις βασίζονται συνήθως στην εκπαίδευση ενός **binary ταξινομητή**, για να διακρίνει τη συμπεριφορά μεταξύ των members και non-members [144].
- Βασισμένες σε μετρήσεις (**Metric Based**): Πρόκειται για πιο απλές και λιγότερο υπολογιστικές προσεγγίσεις, οι οποίες βασίζονται σε μετρήσεις των διανυσμάτων πρόβλεψης που συγκρίνονται με ένα προκαθορισμένο όριο για να λάβουν απόφαση για την κατάσταση μέλους ενός δείγματος. Υπάρχουν 4 βασικά είδη μετρήσεων: correctness-based, loss-based, confidence-based, και entropy-based προβλέψεις [144].

Η πρώτη αποτελεσματική τεχνική membership inference επίθεσης, η οποία χρησιμοποιείται ευρέως, ονομάστηκε **shadow training** [141] και βασίζεται σε binary ταξινομητή. Η κύρια ιδέα είναι ότι ένας επιτιθέμενος μπορεί να δημιουργήσει πολλαπλά shadow models για να μιμηθεί τη συμπεριφορά του μοντέλου στόχου, επειδή ο επιτιθέμενος υποτίθεται ότι γνωρίζει τη δομή και τον αλγόριθμο εκμάθησης του μοντέλου στόχου άλλα δεν έχει γνώση για τα δεδομένα εκπαίδευσης. Για αυτά τα shadow models υπάρχει η γνώση των δεδομένων εκπαίδευσης και δοκιμής, οπότε μπορεί να δημιουργηθεί ένα σύνολο δεδομένων που περιέχει χαρακτηριστικά και ετικέτες για το αν είναι μέλη των εγγραφών δεδομένων εκπαίδευσης και δοκιμής, τα οποία χρησιμοποιούνται για να εκπαιδευτεί ο binary ταξινομητής να αναγνωρίζει αν ένα δείγμα ανήκει στα δεδομένα εκπαίδευσης (βλ. 5.5). Η επίθεση αυτή δουλεύει και σε black-box σενάριο επίθεσης, όπου δεν είναι γνωστός ο αλγόριθμος εκμάθησης του μοντέλου, απλά στην white-box περίπτωση υπάρχουν περισσότερες λεπτομέρειες που φέρνουν πιο γρήγορα και ακριβή αποτελέσματα [144]. Η συγκεκριμένη επίθεση παρήγαγε ακρίβεια αναγνώρισης 70% με 95% συνολικά για το μοντέλο στόχος, με σχεδόν μηδενικά false negatives.

Μεταγενέστερες επιθέσεις χρησιμοποιούν πιο χαλαρές υποθέσεις για το μοντέλο στόχο και τα δεδομένα του, που τις καθιστούν πιο επικίνδυνες. Συγκεκριμένα, στο [145] αποδεικνύεται ότι αρκεί *μόνο ένα shadow model και ένα μοντέλο επίθεσης για να είναι η επίθεση αποτελεσματική*. Επίσης, επιδεικνύονται επιθέσεις μεταφοράς δεδομένων **data transferring attack** οι οποίες παρακάμπτουν την ακριβή διαδικασία παραγωγής συνθετικών δεδομένων, επιτυγχάνοντας όμως αντίστοιχη απόδοση.

Οι λόγοι για τους οποίους τα ML εμφανίζουν αυτήν τη συμπεριφορά, που τα καθιστά ευάλωτα σε membership inference επιθέσεις μπορούν να συνοψιστούν κυρίως σε τρεις:

- Υπερπροσαρμογή (**Overfitting**): Τα μεγάλα και πολύπλοκα ML μοντέλα, όπως τα DNN, παρουσιάζουν το φαινόμενο της υπερπροσαρμογής, το οποίο οφείλεται στην υψηλή πολυπλοκότητα του μοντέλου και στο περιορισμένο μέγεθος του συνόλου δεδομένων εκπαίδευσης. Τα μοντέλα βαθιάς μάθησης είναι συνήθως υπερπαραμετροποιημένα, επι-



Σχήμα 5.5: *Σχηματικό παράδειγμα εκπαίδευσης ενός ταξινομητή membership inference επίθεσης, πάνω στις εισόδους και εξόδους των shadow models [141].*

τρέποντάς τους να μαθαίνουν από μεγάλα δεδομένα αλλά και να απομνημονεύουν θόρυβο ή λεπτομέρειες των δεδομένων εκπαίδευσης. Επίσης, τα πεπερασμένα σύνολα δεδομένων συχνά αποτυγχάνουν να αναπαραστήσει ολόκληρη την κατανομή δεδομένων, καθιστώντας δύσκολη τη γενίκευση [141].

- Τύποι μοντέλων (**Target Model Types**): Η επιτυχία αυτών των επιθέσεων καθορίζεται και σε μεγάλο βαθμό από τον τύπο ML μοντέλου που χρησιμοποιείται, κυρίως το κατά πόσο είναι πιθανό μια συγκεκριμένη εγγραφή δεδομένων να επηρεάσει το όριο απόφασης του μοντέλου. Για παράδειγμα, η έρευνα στο [146] έδειξε ότι τα δέντρα αποφάσεων είναι τα πιο ευάλωτα από άλλα είδη μοντέλων (συγκεκριμένα DNN, Logistic Regression, Naive Bayes, kNN, Trees σε 7 διαφορετικά dataset), καθώς μια εγγραφή που εμφανίζει ένα μοναδικό χαρακτηριστικό μπορεί να το προκαλέσει να αναπτύξει έναν εντελώς νέο κλάδο, αλλάζοντας δραστικά το όριο απόφασης.
- Ποικιλομορφία δεδομένων εκπαίδευσης (**Training Data Diversity**): Γενικά όσο πιο αντιπροσωπευτικά είναι τα δεδομένα εκπαίδευσης, δηλαδή να μπορούν να αντιπροσωπεύουν καλύτερα ολόκληρη τη διανομή δεδομένων, τότε το μοντέλο θα είναι λιγότερο ευάλωτο σε τέτοιες επιθέσεις, καθώς βοηθάνε το μοντέλο να γενικεύσει καλύτερα τα δεδομένα δοκιμής [144].

Τέλος, ενώ πολλές membership inference επιθέσεις εστιάζουν σε μοντέλα ταξινόμησης, υπάρχει μεγάλη έρευνα για τέτοιες επιθέσεις και σε άλλα ML μοντέλα, όπου οι στόχοι των επιθέσεων είναι ελάχιστα διαφορετικοί, όπως και οι λόγοι που είναι ευάλωτα αυτά το μοντέλα σε τέτοιες επιθέσεις. Ενδεικτικά μερικά είδη ML μοντέλων στα οποία εφαρμόζονται membership inference επιθέσεις: **Classification**, **Generative**, **Embedding**, **Regression** [144].

5.3 Προστασία Ιδιωτικότητας

Στην προηγούμενη ενότητα, αναλύσαμε όλα τα είδη των επιθέσεων που έχουν σκοπό την παραβίαση της ιδιωτικότητας (ή απορρήτου) ενάντια στα ML συστήματα. Σε αυτήν την ενότητα, αναφέρουμε τις τεχνικές που μπορούν να χρησιμοποιηθούν για την προστασία της ιδιωτικότητας των μοντέλων αλλά και των δεδομένα τα οποία χρησιμοποιούνται για την εκπαίδευση αυτών. Επίσης, κάνουμε μια πιο αναλυτική παρουσίαση στις πιο δημοφιλείς τεχνικές που χρησιμοποιούνται για τον σκοπό αυτό.

5.3.1 Προστασία Ιδιωτικότητας Μοντέλου

Οι τεχνικές για προστασία της ιδιωτικότητας του μοντέλου, δηλαδή από model extraction επιθέσεις, μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες: (i) αντιδραστικές (**reactive**) π.χ. για την ανίχνευση τωρινών ή παλαιότερων επιθέσεων, και (ii) προληπτικές (**proactive**) π.χ. για την πρόληψη μιας επίθεσης και μείωσης του αντίκτυπου της [137].

Τις **reactive** άμυνες μπορούμε να τις διακρίνουμε σε περαιτέρω κατηγορίες ανάλογα με τον στόχος τους: (i) επαλήθευσης ιδιοκτησίας (**ownership verification**) για να αποδειχθεί η ιδιοκτησία ενός κλεμμένου μοντέλου μέσω μοναδικών αναγνωριστικών, και (ii) ανίχνευση επιθέσεων (**attack detection**) για να διαπιστωθεί εάν ένα μοντέλο δέχεται model extraction επιθέσεις μέσω παρακολούθησης των ερωτημάτων ή εισόδων. Οι **proactive** άμυνες προσπαθούν να μετριάσουν μια αναμενόμενη επίθεση μέσω τροποποίησης ορισμένων χαρακτηριστικών του μοντέλου, όπως αρχιτεκτονική, μαθημένες παραμέτρους, όρια απόφασης ή γενικά τη συνολική του αποτελεσματικότητα. Γενικά, οι reactive μέθοδοι πολλές φορές κρίνονται ως καλύτερες άμυνες, διότι μπορούν να ενημερώσουν τους κατόχους των μοντέλων για περιστατικά model extraction, αλλά και δεν επηρεάζουν αρνητικά τους κανονικούς χρήστες, όπως οι proactive άμυνες που μπορεί να μειώσουν την ακρίβεια του μοντέλου [137].

5.3.1.1 Αντιδραστικές τεχνικές για προστασία από model extraction

Μια κατηγορία αντιδραστικών (reactive) τεχνικών για την προστασία της ιδιωτικότητας των μοντέλων είναι αυτή του **Unique Model Identifier (UMI)**, όπου αναγνωρίζεται μια μοναδική ιδιότητα του μοντέλου η οποία θα μεταφερθεί σε substitute μοντέλα κατά την εξαγωγή του αρχικού, και θα μπορεί να χρησιμοποιηθεί για να αποδειχθεί η πραγματική ιδιοκτησία του μοντέλου. Αυτή η ιδιότητα συνήθως προκύπτει από το μοντέλο και τα δεδομένα που έχει εκπαιδευτεί [147], ή ακόμα και με χρήση adversarial examples τα οποία μεταφέρονται μόνο σε substitue μοντέλα [148]. Μια άλλη μέθοδος, για την απόδειξη της ιδιοκτησίας ενός κλεμμένου μοντέλου είναι και η τεχνική του **Watermarking** [149], όπου ενεργά γίνεται ενσωμάτωση κρυφών πληροφοριών στο μοντέλο, που μόνο οι νόμιμοι ιδιοκτήτες γνωρίζουν πως να εξάγουν. Αυτή η πληροφορία συνήθως εισάγεται κατά την εκπαίδευση, όπου το μοντέλο μαθαίνει να προβλέπει μια προκαθορισμένη τιμή για ένα δείγμα πολύ μακριά από το όριο απόφασης (outlier), όποτε μόνο κάποιος που το γνωρίζει θα μπορεί να επερωτήσει το μοντέλο με το outlier δείγμα και να ελέγξει αν παράγει την προκαθορισμένη τιμή [137].

Μια άλλη κατηγορία μεθόδων, βασίζεται στην ανίχνευση κακόβουλων χρηστών και ερωτημάτων που στοχεύουν στο να εξάγουν πληροφορίες για το μοντέλο (**Monitor-based**). Μια τέτοια μέθοδος είναι και η **PRADA** [135], η οποία βασίζεται στην υπόθεση ότι τα 'adversarial' ερωτήματα που μοντέλων που προσπαθούν να εξερευνήσουν τα όρια απόφασης θα έχουν διαφορετική κατανομή από τα κανονικά, όποτε αναλύει την κατανομή των ερωτηθέντων δειγμάτων, και ανιχνεύει εάν είναι adversarial από την απόκλιση τους από την κανονική κατανομή. Η άμυνα αυτή έχει δείξει να ανιχνεύει όλες τις πιθανές επιθέσεις εξαγωγής χωρίς false positives, καθώς δεν κάνει καμία υπόθεση για τα δεδομένα εκπαίδευσης (π.χ. όπως άλλες adversarial detection άμυνες), αλλά εξετάζει μόνο την κατανομή τον δειγμάτων εισόδου. Επίσης, τα **DefenseNet** [150] και **SEAT** [151], πρόκειται για ML μοντέλα τα οποία έχουν εκπαιδευτεί για να αναγνωρίζουν εάν ένα δείγμα είναι adversarial ή όχι, και μπορούν επίσης να χρησιμοποιηθούν για να αναγνωρίσουν model extraction επιθέσεις.

5.3.1.2 Προληπτικές τεχνικές για προστασία από model extraction

Οι προληπτικές (proactive) τεχνικές δεν μπορούν να αποτρέψουν από το να συμβεί μια επίθεση εξαγωγής του μοντέλου, αλλά στοχεύουν στο να καταστήσουν την ποιότητα του κλεμμένου μοντέλου πολύ χαμηλή ώστε να μην είναι χρήσιμο το μοντέλο ή τα χαρακτηριστικά που θα εξαχθούν.

Ο πιο απλός τρόπος προστασίας, πρόκειται για την **επιστροφή μόνο των ετικετών** ως απάντηση από τα ML μοντέλα και όχι την αποκάλυψη περισσότερων πληροφοριών όπως πιθανότητες πρόβλεψης, ή τις εξόδους των softmax και logit επιπέδων [134]. Πρόκειται για έναν αρκετά περιοριστικό τρόπο προστασίας, αλλά αρκετά αποτελεσματικό. Αντίστοιχα, υπάρχουν και άλλοι τρόποι για να μειωθεί η πληροφορία που επιστρέφεται από τα μοντέλα, παραμένοντας όμως χρήσιμα σε εφαρμογές, όπως το να μην **απαντάνε σε ελλιπή ερωτήματα** που δεν ταιριάζουν σε κάποια καθορισμένη μορφή εισόδου (κυρίως για NLP εφαρμογές) [43]. Όμως, ακόμη και εάν παρέχονται μόνο οι ετικέτες κλάσεων ως έξοδος από τα μοντέλα, είναι ακόμα δυνατόν να υπάρξουν πετυχημένες επιθέσεις [134], οπότε για αυτό υπάρχουν και πιο ισχυρές τεχνικές προστασίας.

Μια άλλη κατηγορία μεθόδων χρησιμοποιούν, διαταραχές στα δείγματα (**Data perturbation**) για να προσφέρουν προστασία από model extraction επιθέσεις, κάνοντας τα μοντέλα να επιστρέφουν μια ανακριβή έξοδο, διατηρώντας παράλληλα την ακεραιότητα της πρόβλεψης. Τα δεδομένα μπορούν να διαταραχθούν σε 3 διαφορετικά στάδια: στην είσοδο του μοντέλου, κατά την πρόβλεψη, και στην έξοδο του μοντέλου [137]. Μια τεχνική που προσθέτει input perturbations, προστατεύει τα μοντέλα εικόνων προσθέτοντας θόρυβο στα ασήμαντα pixels που επιλέγονται με την μέθοδο **Gradient-weighted Class Activation Mapping (Grad-CAM)** [152]. Τεχνικές που προσθέτουν output perturbations, όπως αυτή στο [134], προτείνουν στρογγυλοποίηση των προβλεπόμενων τιμών αξιοπιστίας (**Confidence Rounding**) που επιστρέφουν ως έξοδο τα μοντέλα, καθώς οι κανονικοί χρήστες δε χρειάζονται πολλαπλά δεκαδικά ψηφία ακρίβειας, και μειώνει αντίστοιχα την πληροφορία που μπορεί να αξιοποιηθεί για επιθέσεις.

Η τεχνική του **Differential Privacy** που χρησιμοποιείται κυρίως για να προστατέψει την ιδιωτικότητα των δεδομένων με εγγυήσεις (βλ. λεπτομερή ανάλυση στο 5.3.3), σε εφαρμογές ML αλλά και σε οποιαδήποτε εφαρμογή χρησιμοποιεί σύνολα δεδομένων, μπορεί να χρησιμοποιηθεί επίσης για να μειώσει την επίπτωση των επιθέσεων εξαγωγής στα ML μοντέλα. Μια τέτοια εφαρμογή αυτής της μεθόδου, προσθέτει διαταραχές στις εξόδους του μοντέλου με μια τεχνική που ονομάζεται **Boundary Differential Privacy Layer (BDPL)** [153], ώστε να κάνει το αποτέλεσμα όλων των δειγμάτων που είναι κοντά στο όριο απόφασης να μη διακρίνονται μεταξύ τους. Αυτή η μέθοδος προσφέρει εγγυήσεις ιδιωτικότητας (privacy guarantees), δηλαδή ότι ένας επιτιθέμενος δεν μπορεί να μάθει το όριο απόφασης με μια προκαθορισμένη ακρίβεια, ανεξάρτητα από το πόσα ερωτήματα και να γίνουν στο API πρόβλεψης του μοντέλου.

Τέλος, για τη σωστή προστασία των μοντέλων πρέπει να γίνει κατάλληλη επιλογή της αμυντικής στρατηγικής που θα εφαρμοστεί με βάσει τους στόχους και της συνθήκες κάθε μοντέλου. Εάν ο πρωταρχικός στόχος είναι η παρακολούθηση των κακόβουλων χρηστών ενός ML API, τότε μπορούν να εφαρμοστούν τεχνικές μοναδικών χαρακτηριστικών ή watermarking. Εάν όμως οι επιτιθέμενοι δε δημοσιεύσουν πότε τα κλεμμένα μοντέλα, τότε οι τεχνικές

αυτές είναι αναποτελεσματικές. Επίσης, οι άμυνες ανίχνευσης μπορεί να αργήσουν να ανιχνεύσουν μια επίθεση, και το μοντέλο να κλαπεί, οπότε μπορεί να χρειάζονται proactive άμυνες για να μετριάσουν τις επιπτώσεις. Ο συνδυασμός όμως διαφορετικών αμυντικών τεχνικών μπορεί να είναι μια λύση που ενισχύσει την προστασία σε πολλαπλά επίπεδα [137].

Defense	Objective	Approach	Details
Dataset Inference [147]	Reactive	Unique Model Identifier	Check specific training dataset
Conferrable AE [148]	Reactive	Unique Model Identifier	Adversarial example fingerprint
Watermarking [149]	Reactive	Watermarking	Embed hidden information
PRADA [135]	Reactive	Detection-based	Distribution of queries analysis
DefenseNet [150]	Reactive	Detection-based	Adversarial example detection
SEAT [151]	Reactive	Detection-based	Adversarial example detection
Label-only return [134]	Proactive	Basic Defense	Predection API minimization
Well-formed queries [43]	Proactive	Basic Defense	Reject non complete queries
Grad-CAM [152]	Proactive	Input perturbations	Input noise to unimportant pixels
Confidence Rounding [134]	Proactive	Output perturbations	Confidence score rounding
BDPL [153]	Proactive	Differential Privacy	DP for decision boundary

Πίνακας 5.1: *Συγκριτικός πίνακας των τεχνικών προστασίας της ιδιωτικότητας μοντέλων (**model privacy protection**) με τα κύρια χαρακτηριστικά τους*

5.3.2 Προστασία Ιδιωτικότητας Δεδομένων

Για την προστασία της ιδιωτικότητας των δεδομένων από privacy attacks σε ML μοντέλα, έχουν αναπτυχθεί τεχνικές, πολλές από τις οποίες προέρχονται από άλλους κλάδους της πληροφορικής (π.χ. κρυπτογραφία) ή βασίζονται σε παραδοσιακές τεχνικές προστασίας δεδομένων (π.χ. access controls, anonymization). Γενικά τα συστήματα τεχνητής νοημοσύνης θα πρέπει να είναι σε θέση να διατηρούν το απόρρητο των δεδομένων, από τη συλλογή τους, κατά την αποθήκευση τους, κατά την εκπαίδευση τους όσο και κατά τη λειτουργία του μοντέλου. Για την κάλυψη αυτής της ανάγκης έχει δημιουργηθεί και η έννοια των ML μοντέλων που διατηρούν την ιδιωτικότητα (**Privacy Preserving ML**) [143].

5.3.2.1 Privacy Preserving Machine Learning (PPML)

Η έννοια του PPML αναφέρεται σε ML μοντέλα τα οποία έχουν σχεδιαστεί με στόχο την εκπαίδευση τους χωρίς να έχουν άμεση πρόσβαση στα δεδομένα ή ακόμα και αυτή η πρόσβαση να προσφέρει εγγυήσεις ότι τα δεδομένα είναι προστατευμένα από διαρροές. Αυτό δημιουργεί αρχικά τις συνθήκες ώστε να επιτρέπεται σε πολλαπλούς συμμετέχοντες να εκπαιδεύσουν συνεργατικά τα μοντέλα τους, περιορίζοντας την πρόσβαση ή την κοινή χρήση τους και χρησιμοποιώντας δεδομένα από πολλαπλές πηγές χωρίς να τα αποδεσμεύσουν, να τα δημοσιεύσουν, ή να τα μοιραστούν στην αρχική τους ευαίσθητη μορφή [143].

Πολλές από τις τεχνικές που εφαρμόζονται για τη δημιουργία PPML models αναλύονται παρακάτω, όμως ενδεικτικά κάποιες από αυτές είναι: (i) η **Differential Privacy** (βλ. 5.3.3) όπου εφαρμόζεται θόρυβος (input perturbation) κατανεμημένα στα δεδομένα προσφέροντας μαθηματικές εγγυήσεις για το απόρρητο των ατόμων σε ένα σύνολο δεδομένων χωρίς να παραμορφώνει τις στατιστικές ιδιότητες του αρχικού συνόλου δεδομένων [141], (ii) η **Homomorphic Encryption** (βλ. 5.3.5) όπου εφαρμόζεται κρυπτογράφηση στα δεδομένα με

τέτοιο τρόπο ώστε να επιτρέπεται η πράξη της πρόσθεσης και του πολλαπλασιασμού (και άρα πιο σύνθετων συναρτήσεων) πάνω στα κρυπτογραφημένα δεδομένα, διασφαλίζοντας έτσι ότι δεν εμφανίζονται καθόλου τα ακατέργαστα δεδομένα [143], (iii) το **Secure Multiparty Computation** (βλ. 5.3.4) όπου τα κατανεμημένα δεδομένα από πολλαπλές πηγές χρησιμοποιούν υπολογισμούς και συναθροίζονται με ασφαλή τρόπο, μειώνοντας τη διαρροή πληροφοριών κατά την εκπαίδευση [145], (iv) το **Federated Learning** (βλ. 5.3.6) όπου γίνεται αποκεντρωμένη εκπαίδευση τοπικών μοντέλων με τα αρχικά ακατέργαστα δεδομένα, αποφεύγοντας την ανταλλαγή και διαμοιρασμό δεδομένων, κοινοποιώντας μόνο τις ενημερώσεις των παραμέτρων των μοντέλων, εξασφαλίζοντας την προστασία των δεδομένων εκπαίδευσης [2].

Ο στόχος όλων αυτών των τεχνικών είναι η εκτέλεση της εκπαίδευσης όπως στην κανονική περίπτωση, με ιδανικά ελάχιστες επιπτώσεις στην απόδοση του μοντέλου, ενισχύοντας την προστασία των δεδομένων εκπαίδευσης, τόσο κατά την εκπαίδευση όσο και από ανταγωνιστικές επιθέσεις, όπως για παράδειγμα στην περίπτωση της Differential Privacy, όπου τα μοντέλα προστατεύονται και από επιθέσεις όπως membership inference, καθώς εξ ορισμού προστατεύουν τα σύνολα δεδομένων από την ταυτοποίηση κάποιου συγκεκριμένου δείγματος μέσα στο σύνολο [141]. Αντίστοιχα, στην περίπτωση της Homomorphic Encryption, εφόσον τα αρχικά δεδομένα δεν αποκαλύπτονται ποτέ δεν μπορούν και εξορισμού να αντιστραφούν από το τελικό μοντέλο μέσω model inversion επιθέσεων [154].

Στις περισσότερες από αυτές τις τεχνικές, *υπάρχει ένα κόστος για την προστασία της ιδιωτικότητας των δεδομένων*, το οποίο μπορεί να είναι είτε υπολογιστικό, είτε στην απόδοση και ακρίβεια του μοντέλου ή στην ευαισθησία και δυνατότητα γενίκευσης του. Όπως και στην περίπτωση του adversarial robustness, έχει αποδειχτεί ότι υπάρχει ένας συμβιβασμός μεταξύ ιδιωτικότητας και ακρίβειας (privacy/accuracy trade-off), το οποίο όμως είναι διαφορετικό για κάθε τεχνική [155]. Στην περίπτωση της Differential Privacy, η εισαγωγή θορύβου επηρεάζει την ακρίβεια της πρόβλεψης, ανάλογα και με το επιλεγμένο προϋπολογισμό ιδιωτικότητας (**privacy budget**) [156]. Αντίστοιχα, στην περίπτωση της Homomorphic Encryption, εφόσον κάθε δεδομένο είναι κρυπτογραφημένο, με τη χρήση της αυξάνεται το υπολογιστικό κόστος, η ακρίβεια και εισάγονται περιορισμοί στον σχεδιασμό της εκπαίδευσης καθώς περιορίζεται το σύνολο των αριθμητικών πράξεων που είναι διαθέσιμες [4]. Τέλος, η χρήση του Federated Learning, απαιτεί άφθονους υπολογιστικούς πόρους και εύρος ζώνης από τις τοπικές συσκευές και αντίστοιχο κόστος επικοινωνίας, λόγω της κατανεμημένης αρχιτεκτονικής αυτής της εκπαίδευσης [157].

Technique	Strength	Weakness
Differential privacy	Provable guarantee of privacy	Accuracy drop
Secure Multiparty Computation	Computing on encrypted data	High computation cost
Homomorphic Encryption	Encrypted training data	Numeric Data only
Federated Learning	Decentralised training	High communication cost

Πίνακας 5.2: *Συγκριτικός πίνακας των τεχνικών της απόκρυψης των δεδομένων εκπαίδευσης ML μοντέλων (**PPML**) με τα κύρια χαρακτηριστικά τους*

5.3.2.2 Τεχνικές για προστασία από model inversion

Οι τεχνικές για την προστασία των δεδομένων εκπαίδευσης ενός μοντέλου από model inversion που προσπαθούν να εξάγουν δεδομένα από το σύστημα, διακρίνονται κυρίως σε (i) βασικές τεχνικές που τροποποιούν τα μοντέλα, ελέγχουν τις εισόδους και μορφοποιούν τις εξόδους τους, και σε (ii) τεχνικές που τροποποιούν τα δεδομένα.

Πιο συγκεκριμένα για τις βασικές τεχνικές προστασίας, υπάρχουν πολλά μέτρα ασφαλείας που μπορούν να εφαρμοστούν στα ML συστήματα πριν καν ένα ερώτημα ή μια είσοδος φτάσει στο μοντέλο. Κάποια από αυτά είναι ο περιορισμός των ερωτημάτων που επιτρέπονται από ένα χρήστη στο σύστημα (**rate limiting**), καθώς για να υλοποιηθούν αυτές οι επιθέσεις συνήθως χρειάζεται ένας μεγάλος αριθμός ερωτήσεων προς το μοντέλων. Επίσης, η επικύρωση εισόδων ή ερωτημάτων (**input validation**) είναι μια τεχνική που μπορεί να εφαρμοστεί, στην οποία ορίζεται εξαρχής η μορφή των ερωτημάτων που θα γίνονται δεκτά και σε κάθε αίτημα προς το μοντέλο, πραγματοποιείται έλεγχος για το ερώτημα, και απορρίπτεται οποιαδήποτε είσοδος δεν ικανοποιεί της προϋποθέσεις [43]. Αντίστοιχα, στην έξοδο των μοντέλων μια τεχνική είναι η επιστροφή μόνο των απολύτως απαραίτητων πληροφοριών που απαιτούνται για να είναι χρήσιμη η απάντηση, το οποίο μπορεί να εφαρμοστεί είτε με επιστροφή μόνο των προβλέψεων-απαντήσεων είτε ακόμα και με στρογγυλοποίηση της βαθμολογίας που παράγεται από το softmax επίπεδο του μοντέλου (**softmax score rounding**), καθώς ακόμα και με στρογγυλοποίηση μπορούν να παραχθούν βαθμολογίες εμπιστοσύνης που είναι χρήσιμες για πολλούς σκοπούς, καθιστώντας όμως το μοντέλο ανθεκτικό σε επιθέσεις ανακατασκευής.

Για τις τεχνικές που τροποποιούν τα δεδομένα, αρχικά μπορούν να αξιοποιηθούν μέθοδοι επιμέλειας και φιλτραρίσματος των δεδομένων πριν την εκπαίδευση (**data curation**), για να μπορέσουν να αποφευχθούν εξαρχής να βρεθούν στο σύνολο δεδομένων, ευαίσθητα προσωπικά στοιχεία [141]. Ακόμη, είναι σημαντικό να αφαιρούνται επίσης τα πολλαπλά αντίγραφα ή διπλότυπα δεδομένων που μπορούν να υπάρχουν (**data deduplication**), ειδικότερα στις περιπτώσεις των Generative μοντέλων (π.χ. GPT-2, GPT-3, Stable Diffusion) έχουν την τάση να απομνημονεύουν ολόκληρα δεδομένα εκπαίδευσης (memorization), είτε πρόκειται για προτάσεις κειμένου είτε και για ολόκληρες εικόνες [142]. Για παράδειγμα, σε ένα diffusion model εκπαιδευμένο στο CIFAR-10, όταν εκπαιδεύεται στο ίδιο σύνολο δεδομένων από το οποίο έχουν αφαιρεθεί όλες οι διπλότυπες εικόνες (ομοιότητα πάνω από 85%), το νέο σύνολο δεδομένων έχει 10.55% λιγότερα δείγματα (44,725 έναντι 50,000), και το μοντέλο αναπαράγει 23% λιγότερα παραδείγματα σε σχέση με το αρχικό (986 έναντι 1280).

Οι προηγούμενες τεχνικές αποτελούν μια πρώτη γραμμή υπεράσπισης και είναι καλές πρακτικές για την καλύτερη ασφάλεια των δεδομένων, όμως δεν προσφέρονται για την απόλυτη πρόληψη κατά των διαρροών της ιδιωτικότητας. Για να συμβεί αυτό χρειάζονται πιο αυστηρές μέθοδοι, οι οποίες όμως μπορεί να έχουν επιπτώσεις σε άλλες λειτουργίες του μοντέλου. Μια από αυτές είναι και η **Differential Privacy** [158], για να προστεθεί θόρυβος και να αποκρυφτούν οι ευαίσθητες πληροφορίες, παρέχοντας ισχυρές εγγυήσεις για την ιδιωτικότητα των μεμονωμένων δεδομένων στο σύνολο δεδομένων. Η τεχνική αυτή εφαρμόζεται στα ML μοντέλα, και ειδικότερα στα νευρωνικά, κατά την εκπαίδευση μέσω του αλγορίθμου **DP-SGD** (βλ. 5.3.3), όπου οι κλίσεις του μοντέλου περιορίζονται και εισάγεται θόρυβος για να αποτραπεί η διαρροή ουσιαστικών πληροφοριών σχετικά με την παρουσία οποιασδήποτε μεμονωμένης

εισόδου στο σύνολο δεδομένων [142]. Λόγω όμως του συμβιβασμού μεταξύ ακρίβειας και ιδιωτικότητας (όσο μεγαλύτερος θόρυβος, τόσο μεγαλύτερη πτώση ακρίβειας [159]) και επειδή αυξάνει τον χρόνο εκπαίδευσης, αποφεύγεται να χρησιμοποιείται σε πολύ μεγάλα μοντέλα όπως LLMs ή Diffusion Models, τα οποία η ακρίβεια τους περιορίζονται κυρίως από το κόστος εκπαίδευσης.

Μια άλλη τεχνική που προστατεύει τα αρχικά δεδομένα είναι και η **Homomorphic Encryption** (βλ. 5.3.5), η οποία με την κρυπτογράφηση των δεδομένων και εκτέλεση υπολογισμών απευθείας πάνω στην κρυπτογραφημένη είσοδο, αποτρέπει τις ευαίσθητες πληροφορίες από το να διαρρεύσουν. Όμως, ένα μειονέκτημα αυτής της κρυπτογράφησης είναι ότι πάσχει από τεράστια αναποτελεσματικότητα και δεν ισχύει για όλες τις λειτουργίες ενός νευρωνικού δικτύου [159].

Defense	Approach	Details
Rate Limiting [43]	Basic Input/Output defense	API queries rate limiting
Input Validation [43]	Basic Input/Output defense	Reject non-valid input
Softmax Score Rounding [43]	Basic Input/Output defense	Rounding of confidence scores
Data Curation [141]	Data Preprocessing	Filtering of sensitive data
Data Deduplication [142]	Data Preprocessing	Delete duplicate data
DP-SGD [159]	Differential Privacy	Provable guarantee of privacy
Homomorphic Encryption [159]	Homomorphic Encryption	Computing on encrypted data

Πίνακας 5.3: *Συγκριτικός πίνακας των τεχνικών προστασίας της ιδιωτικότητας των δεδομένων εκπαίδευσης ML μοντέλων από* ***model inversion*** *επιθέσεις με τα κύρια χαρακτηριστικά τους*

5.3.2.3 Τεχνικές για προστασία από membership inference

Οι τεχνικές για την προστασία των δεδομένων εκπαίδευσης ενός μοντέλου από membership inference επιθέσεις που προσπαθούν να αναγνωρίσουν εάν ένα δείγμα ανήκει στο σύνολο δεδομένων εκπαίδευσης, μπορούν να κατηγοριοποιηθούν σε τέσσερις κατηγορίες (i) απόκρυψη βαθμολογίας εμπιστοσύνης (**confidence score masking**), (ii) κανονικοποίησης (**regularization**), (iii) απόσταξης γνώσης (**knowledge distillation**), και (iv) **differential privacy** [144].

Οι **confidence score masking** τεχνικές, στοχεύουν στο να αποκρύψουν τις πραγματικές βαθμολογίες εμπιστοσύνης που επιστρέφονται από ταξινομητές ώστε να μετριάσουν την αποτελεσματικότητα των membership inference επιθέσεων. Μια τέτοια μέθοδος είναι και ο περιορισμός του διανύσματος πιθανοτήτων στις k πιο πιθανές κλάσεις (**top-k classes restriction**) [141], όπου όταν ο αριθμός κλάσεων είναι μεγάλος, πολλές κλάσεις μπορεί να έχουν μικρές πιθανότητες στο διάνυσμα πρόβλεψης του μοντέλου, οπότε και προστίθεται ένα φίλτρο για να περιορίσει της πληροφορίες που διαρρέει το μοντέλο. Στην πιο ακραία περίπτωση αυτής της τεχνικής, το μοντέλο επιστρέφει μόνο την ετικέτα την πιο πιθανής κατηγορίας χωρίς να αναφέρει την πιθανότητα της (**prediction label only**) [141], προσφέροντας την πιο περιορισμένη γνώση σε επιτιθέμενους, παραμένοντας όμως ευάλωτη σε ισχυρές επιθέσεις [144]. Μια άλλη κατηγορία τεχνικών, προσπαθεί να στρογγυλοποιήσει τις τιμές των διανυσμάτων πιθανοτήτων σε d floating points (**prediction precision rounding**) [141], όπου όσο μικρότερο το d τόσο λιγότερη πληροφορία διαρρέεται από το μοντέλο. Τέλος, μια άλλη κατηγορία τεχνι-

κών, προσθέτει προσεκτικά κατασκευασμένο θόρυβο στο διάνυσμα πρόβλεψης για να καλύψει τις πραγματικές βαθμολογίες. Μια τέτοια είναι και η **MemGuard** [160], η οποία αξιοποιεί μεθόδους ανταγωνιστικής μηχανικής μάθησης, μετατρέποντας αυτό το θορυβώδες διάνυσμα σε adversarial example, ώστε να μην μπορούν οι επιτιθέμενοι να αναγνωρίσουν εάν ένα δείγμα είναι member ή όχι με βάση αυτήν την έξοδο. Η μέθοδος αυτή δεν απαιτεί επανεκπαίδευση του αρχικού μοντέλου, και ούτε επηρεάζει την ακρίβεια του, αλλά ενώ μπορεί να μετριάσει το αποτέλεσμα των membership inference επιθέσεων, παραμένει ευάλωτη σε αυτές [155]. Γενικά, όλες οι παραπάνω τεχνικές δεν απαιτούν επανεκπαίδευση του μοντέλου, δεν επηρεάζουν την ακρίβεια του μοντέλου, παρά μόνο τα διανύσματα πρόβλεψης, οπότε είναι εύκολο να υλοποιηθούν, αλλά δεν μπορούν να αποτελέσουν τις μοναδικές μεθόδους προστασίας του μοντέλου, καθώς δεν προσφέρουν καμία ελάχιστη επιβεβαίωση ιδιωτικότητας [144].

Οι **regularization** τεχνικές, στοχεύουν στο να μειώσουν τον βαθμό υπερπροσαρμογής (overfitting) των μοντέλων, το οποίο έχει σαν συνέπεια και τον μετριασμό των membership inference επιθέσεων. Υπάρχουν πολλές κλασσικές μέθοδοι regularization που μπορούν να αξιοποιηθούν, όπως: **L2-norm Regularization** όπου κατά την εκπαίδευση τιμωρούνται οι μεγάλοι παράμετροι [141], **Dropout** όπου απορρίπτονται τυχαία ένα προκαθορισμένο ποσό νευρώνων κατά την εκπαίδευση, και **Model Stacking** όπου συνδυάζονται πολλαπλά μοντέλα που εκπαιδεύονται ξεχωριστά [145], άλλα και άλλες όπως **Label Smoothing**, **Early Stopping**, **Data Augmentation** [144]. Έχουμε ήδη αναφερθεί στις περισσότερες από αυτές, ως τεχνικές που μπορούν να προστατεύσουν τα μοντέλα από επιθέσεις με adversarial examples, καθώς βελτιώνουν τη γενίκευση των μοντέλων, αναγκάζοντας τα να παράγουν παρόμοιες κατανομές εξόδου για members ή non-members των δεδομένων εκπαίδευσης. Οι regularization τεχνικές χρειάζονται να εφαρμοστούν κατά την εκπαίδευση των μοντέλων, αλλάζοντας τις εσωτερικές παραμέτρους τους, έχοντας αντίστοιχα επίπτωση στην ακρίβεια των μοντέλων, και επίσης μπορεί να μην παρέχουν ικανοποιητική προστασία της ιδιωτικότητας σε σχέση με άλλες μεθόδους [155].

Οι **knowledge distillation** τεχνικές, χρησιμοποιούν τα αποτελέσματα ενός μεγάλου μοντέλου για να εκπαιδεύσουν ένα μικρότερο μοντέλο, ώστε να μεταφερθεί η γνώση από το μεγάλο μοντέλο στο μικρό, επιτρέποντας του να έχει παρόμοια ακρίβεια με το μεγάλο [144]. Μια τέτοια τεχνική είναι και η **Distillation For Membership Privacy (DMP)** [161], όπου χρειάζεται ένα ιδιωτικό σύνολο δεδομένων εκπαίδευσης και ένα σύνολο δεδομένων χωρίς ετικέτες. Πρώτα εκπαιδεύεται ένα 'μεγάλο' μοντέλο με τα ιδιωτικά δεδομένα ώστε να επισημανθούν με ετικέτες στο σύνολο που δεν έχει, και ύστερα εκπαιδεύεται ένα 'μικρό' μοντέλο με βάση το νέο σύνολο. Η ιδέα πίσω από αυτήν την τεχνική είναι ότι ο νέος ταξινομητής δεν έχει άμεση πρόσβαση στα δεδομένα εκπαίδευσης του αρχικού μοντέλου, μειώνοντας έτσι σημαντικά τη διαρροή πληροφοριών μέλους.

Η **differential privacy**, είναι ένας πιθανολογικός μηχανισμός ιδιωτικότητας που παρέχει μια θεωρητική εγγύηση απορρήτου των δεδομένων, ο οποίος μπορεί να εφαρμοστεί πάνω σε ML μοντέλα για την προστασία από membership inference επιθέσεις. Όταν ένα ML μοντέλο εκπαιδεύεται με differential privacy, το μοντέλο δε μαθαίνει ούτε θυμάται τα στοιχεία συγκεκριμένων χρηστών εάν ο προϋπολογισμός ιδιωτικότητας (privacy budget) είναι αρκετά μικρός (βλ. 5.3.3). Εξ ορισμού λοιπόν, τα differentially private μοντέλα περιορίζουν φυσικά την πιθανότητα επιτυχίας των membership inference επιθέσεων, και ακόμα και με χρήση

βοηθητικών πληροφοριών, η ανωνυμία τον δεδομένων διατηρείτε και ένας επιτιθέμενος δεν μπορεί να αυξήσει την απώλεια της ιδιωτικότητας [143]. Ο πιο συχνός τρόπος εκπαίδευσης differentially private ML μοντέλων είναι μέσω του DP-SGD αλγορίθμου [162], ο οποίος εισάγει θόρυβο στις κλίσεις του μοντέλου κατά την εκπαίδευση. Όπως όμως αναφερθήκαμε και στην προηγούμενη ενότητα, ενώ είναι μια τεχνική που προσφέρει privacy guarantess, αλλά υπάρχει ένας εγγενές συμβιβασμός με την ακρίβεια των μοντέλων αυτών, όπου θυσιάζεται σημαντικά για την προστασία της ιδιωτικότητας.

Defense	Approach	Details
Top-k Classes Restriction [141]	Confidence Score Masking	Return k most probable classes
Prediction Label Only [141]	Confidence Score Masking	Return labels only
Prediction Rounding [141]	Confidence Score Masking	Rounding of prediction scores
MemGuard [160]	Confidence Score Masking	Adversarial noise on prediction
L2-norm Regularization [141]	Regularization	Normal regularization
Dropout [145]	Regularization	Drop random neurons
Model Stacking [145]	Regularization	Combination of models
other classic techniques [144]	Regularization	Label Smooth., Early Stop., Data Augm.
DMP [161]	Knowledge Distillation	No access to private data
DP-SGD [162]	Differential Privacy	Provable guarantee of privacy

Πίνακας 5.4: *Συγκριτικός πίνακας των τεχνικών προστασίας της ιδιωτικότητας των δεδομένων εκπαίδευσης ML μοντέλων από* ***membership inference*** *επιθέσεις με τα κύρια χαρακτηριστικά τους*

5.3.3 Differential Privacy

Η διαφορική ιδιωτικότητα ή **Differential Privacy (DP)** [158], πρόκειται για μια μέθοδο που παρέχει εγγυήσεις ιδιωτικότητας (**privacy guarantees**) για κάθε εγγραφή από αλγορίθμους που επεξεργάζονται ή αναλύουν δεδομένα πάνω σε σύνολα ή βάσεις δεδομένων. Η μέθοδος βασίζεται στην ιδέα του 'μην μαθαίνεις τίποτα για ένα άτομα, μαθαίνοντας παράλληλα χρήσιμες πληροφορίες για όλο τον πληθυσμό', δηλαδή σε περίπτωση δύο συνόλων δεδομένων τα οποία έχουν διαφορά μόνο σε ένα δείγμα (adjacent datasets), και χρησιμοποιούνται από τον ίδιο αλγόριθμο (ή μηχανισμό), η έξοδος αυτού του αλγορίθμου θα πρέπει να είναι παρόμοια [2] (βλ. σχήμα 5.6). Τυπικά, ο ορισμός της DP είναι ο εξής: ένας τυχαίος μηχανισμός $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ με πεδίο $\mathcal{D}$ και εύρος $\mathcal{R}$ ικανοποιεί το ϵ-differential privacy αν για κάθε δύο υποσύνολα δεδομένων d, d' που διαφέρουν κατά ένα δείγμα, και για κάθε υποσύνολο $S \subseteq \mathcal{R}$, ισχύει ότι:

$$Pr\left[\mathcal{M}(d) \in S\right] \leq e^{\epsilon} Pr\left[\mathcal{M}(d') \in S\right]$$

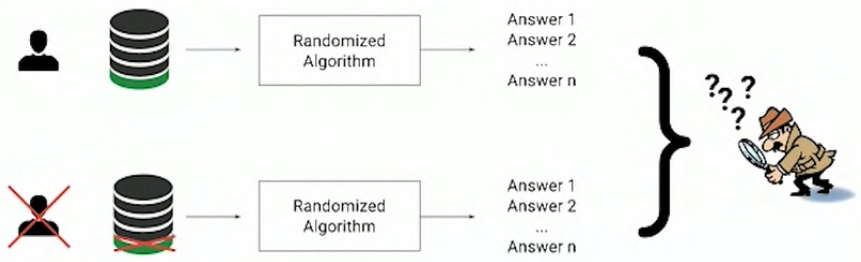
Τη μεταβλητή ϵ την ονομάζουμε προϋπολογισμός ιδιωτικότητας (**privacy budget**) και καθορίζει τον *συμβιβασμό μεταξύ ακρίβειας και διαρροής ιδιωτικότητας* του μηχανισμού $\mathcal{M}$. Όσο πιο μικρή είναι η τιμή του ϵ, τόσο πιο μικρή είναι η διαρροή ιδιωτικότητας και άρα τόσο πιο ισχυρή είναι η εγγύηση και το επίπεδο ιδιωτικότητας [156].

Σε μεταγενέστερη έκδοση αυτού του ορισμού, εισήχθηκε και η μεταβλητή δ στη δεξιά μεριά της εξίσωσης, ως χαλάρωση που επιτρέπει σε ορισμένες εξόδους να μην οριοθετούνται

από e^{ϵ} και δίνει τον ορισμό της (ϵ, δ)-differential privacy:

$$Pr\left[\mathcal{M}(d) \in S\right] \leq e^{\epsilon} Pr\left[\mathcal{M}(d') \in S\right] + \delta$$

η εισαγωγή του δ, το οποίο ονομάζεται πιθανότητα αποτυχίας (failure probability), επιτρέπει τη πιθανότητα η απλή ϵ-differential privacy να σπάσει με πιθανότητα δ (η οποία είναι κατά προτίμηση μικρότερη από $1/|d|$) [162]. Στην περίπτωση $\delta = 0$, ο τυχαίος μηχανισμός $\mathcal{M}$ παρέχει ϵ-differential privacy με τον αυστηρότερο ορισμό του.



Σχήμα 5.6: *Σχηματικό παράδειγμα εφαρμογής DP, όπου σε δύο βάσεις δεδομένων που διαφέρουν κατά μια εγγραφή χρήστη, από το αποτέλεσμα ενός τυχαίο αλγόριθμου πάνω στη βάση που δεν περιέχει την εγγραφή, δεν μπορεί να διακριθεί αν αυτή περιέχεται ή όχι*[1].

Η DP δουλεύει συνήθως προσθέτοντας στατιστικό θόρυβο είτε στις εισόδους (**local DP**), το οποίο προσφέρει ιδιωτικότητα εξαρχής, είτε στις εξόδους (**global DP**) ενός μοντέλου ή ερωτήματος, για το οποίο χρειάζεται εμπιστοσύνη σε αυτόν που διαχειρίζεται τη βάση δεδομένων ή το μοντέλο. Μια ακόμα περίπτωση είναι αυτή για εφαρμογή θορύβου στα NN όπου μπορεί να εισαχθεί στα κρυμμένα επίπεδα (hidden layer) του μοντέλου. Το μέγεθος του θορύβου που θα προστεθεί καθαρίζεται από το privacy budget ϵ. Όσο μικρότερο το ϵ, τόσο πιο μεγάλος θόρυβος εισάγεται και άρα πιο υψηλό επίπεδο ιδιωτικότητας, μειώνοντας όμως τη χρησιμότητα του αποτελέσματος, καθώς χάνονται όλο και περισσότερες πληροφορίες (*συμβιβασμός μεταξύ ιδιωτικότητας και χρησιμότητας*). Οι δύο πιο συνηθισμένοι μηχανισμοί που χρησιμοποιούνται για εισαγωγή θορύβου στην DP είναι: ο Laplace και ο Gaussian μηχανισμός [156].

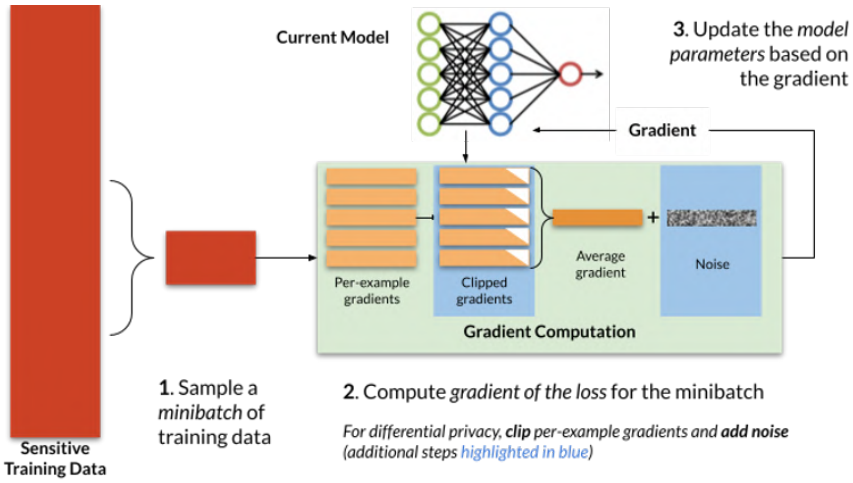
5.3.3.1 Εφαρμογές στη Μηχανική Μάθηση

Η εφαρμογή της DP σε ML εφαρμογές και μοντέλα, συνήθως γίνεται μέσω του **Differentially Private Stochastic Gradient Descent (DP-SGD)** αλγορίθμου [162], ο οποίος πρόκειται για μια παραλλαγή του SGD με 2 επιπλέον βήματα σε κάθε ανανέωση των κλίσεων:

- Αποκοπή των κλίσεων ώστε η L_2 νόρμα τους να είναι κάτω από ένα όριο C.
- Εισαγωγή θορύβου, μεγέθους ανάλογο με τον κανόνα αποκοπής C, στη μέση τιμή των ανανεωμένων και αποκομμένων κλίσεων. Ο θόρυβος είναι κανονικής (Gauss) κατανομής με τυπική απόκλιση $C\sigma$.

Η εφαρμογή του DP-SGD όμως επηρεάζει και την ακρίβεια των μοντέλων, ανάλογα το privacy budget ϵ, και τη 'δυσκολία' του συνόλου δεδομένων. Πιο συγκεκριμένα, στο MNIST dataset, ένα τυπικό ML μοντέλο έχει ακρίβεια 100%, ενώ το DP μοντέλο για $\epsilon = 1 - 3$

[1] https://blog.openmined.org/maintaining-privacy-in-medical-data-with-differential-privacy/
[2] https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy



Σχήμα 5.7: *Σχηματικό διάγραμμα λειτουργίας του SGD και DP-SGD αλγόριθμου. Τα βήματα που είναι αφορούν τον DP-SGD έχουν μπλε φόντο*[2].

έχει 98%-99% ακρίβεια [163]. Στο CIFAR-10, το τυπικό μοντέλο έχει ακρίβεια 98%, ενώ το DP μοντέλο για $\epsilon = 3$ έχει μόλις 69% ακρίβεια [163], αλλά με μια καλύτερη υλοποίηση για $\epsilon = 4$ φτάνει 73.5% ακρίβεια [164]. Στο ImageNet, το τυπικό μοντέλο έχει ακρίβεια 87%, ενώ το DP μοντέλο για $\epsilon = 3$ έχει μόλις 32.4% ακρίβεια, η οποία δεν είναι λειτουργική [164]. Όμως αν το μοντέλο έχει προεκπαιδευτεί πρώτα κανονικά σε ένα μεγάλο σύνολο δεδομένων και μετά γίνει fine-tune με ευαίσθητα δεδομένα και εφαρμογή DP, μπορούν να επιτευχθούν πολύ καλύτερες ακρίβειες, όπως 84.4% για $\epsilon = 1$ και 86% για $\epsilon = 4$ [164].

Με εφαρμογή του DP-SGD, η DP εφαρμόζεται από την εκπαίδευση ενός ML μοντέλου (**hidden layer DP**), μπορεί όμως να προσφερθεί και κατά το inference, όπου η εισαγωγή θορύβου θα γίνει είτε στην είσοδο (**input layer DP**) είτε στα αποτελέσματα του μοντέλου (**output layer DP**). Όμως αυτές οι περιπτώσεις έχουν την τάση να υποβαθμίζουν ακόμα περισσότερο την ακρίβεια των προβλέψεων, καθώς η ποσότητα του θορύβου που εισάγεται αυξάνεται με τον αριθμό ερωτημάτων που απαντώνται από το ML μοντέλο [4].

5.3.3.2 Προστασία Ιδιωτικότητας

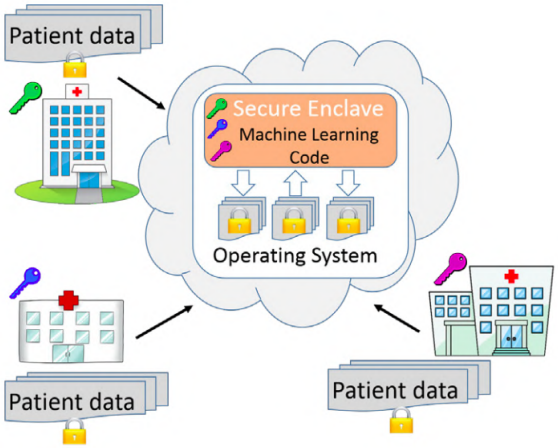
Εξ' ορισμού η DP μπορεί να προστατεύσει τα ML μοντέλα από membership inference επιθέσεις, καθώς η βασική της ιδιότητα είναι η δυνατότητα απόκρυψης ενός δεδομένου από το αν είναι μέρος ενός συνόλου δεδομένων ή όχι. Μπορεί επίσης να χρησιμοποιηθεί και για προστασία από model inversion επιθέσεις, περιορίζοντας την έξοδο πρόβλεψης του μοντέλου [143]. Όμως από αξιολόγηση διαφορετικών DP ML μοντέλων, φαίνεται ότι για να υπάρχει ικανοποιητική προστασία από τέτοιες επιθέσεις, θα πρέπει να θυσιαστεί αρκετά η χρησιμότητα τους [2].

Το **πλεονέκτημα** αυτή της μεθόδου είναι η τυποποιημένη εγγύηση ιδιωτικότητας που προσφέρει άρα και η προστασία από privacy attacks, αλλά και η κοινωνικοποίηση (regularization) που εφαρμόζει στα μοντέλα. Το **μειονέκτημα** όμως αυτή της μεθόδου είναι ο συμβιβασμός μεταξύ χρησιμότητας και ιδιωτικότητας, καθώς όσο πιο ισχυρός θόρυβος εισάγεται τόσο τα μοντέλα έχουν την τάση να χάνουν την ιδιαιτερότητά της κατανομής τον δεδομένων, και άρα να μειώνεται η χρησιμότητα τους. Επίσης, η εφαρμογή DP σε μεγάλα μοντέλα (π.χ.

LLMs) μπορεί να εισάγει και μεγάλο υπολογιστικό και χρονικό κόστος εκπαίδευσης, μειώνοντας αντίστοιχα και την ακρίβεια, η οποία είναι περιορισμένη πολλές φορές από το κόστος εκπαίδευσης [141].

5.3.4 Secure Multi-party Computation (MPC)

Ο ασφαλής υπολογισμός πολλών μερών ή **Secure Multi-party Computation (MPC)**, πρόκειται για ένα κλάδο της κρυπτογραφίας, όπου ο στόχος είναι η δημιουργία μεθόδων για να υπολογίζεται από πολλαπλά μέρη από κοινού μια συνάρτηση πάνω σε εισόδους, οι οποίες παραμένουν ιδιωτικές (μέσω κρυπτογράφησης) και κανένα μέρος δεν έχει πρόσβαση στις εισόδους των άλλων μελών [165]. Ο σκοπός ανάπτυξης αυτού του κλάδου έχει να κάνει με εργασίες όπου υπάρχει ανάγκη ενός κοινού υπολογισμού από πολλές πηγές δεδομένων, όμως για τη διασφάλιση της ιδιωτικότητας των δεδομένων, απαγορεύεται συγκέντρωση των δεδομένων σε ένα σημείο ή η αποκάλυψη δεδομένων σε άλλους συντελεστές, οι οποίοι μπορεί να είναι κακόβουλοι ή αξιόπιστοι. Για παράδειγμα, για εφαρμογές ανάλυσης δεδομένων πάνω σε ιστορικά ασθενών, είναι πολύ δύσκολο να δοθεί πρόσβαση σε βάσεις ή σύνολα δεδομένων νοσοκομειακών εγκαταστάσεων που κατέχουν τα ιστορικά των ασθενών, όμως με MPC μεθόδους, μπορούν να συνεισφέρει το κάθε νοσοκομείο τα δεδομένα του για τον υπολογισμό κάποιον συναρτήσεων, χωρίς να χρειαστεί να αποκαλυφθεί ποτέ κάποιο ευαίσθητο προσωπικό δεδομένο ασθενή στα υπόλοιπα μέλη (βλ. σχήμα 5.8).



Σχήμα 5.8: *Σχηματικό παράδειγμα εφαρμογής MPC σε πολλαπλά νοσοκομεία, όπου τα σύνολα δεδομένων κρυπτογραφούνται με διαφορετικό κλειδί. Τα νοσοκομεία τρέχουν ένα προσυμφωνημένο ML αλγόριθμο σε έναν μεσάζοντα (π.χ. cloud data center), όπου μοιράζονται τα κλειδιά τους, ώστε ο μεσάζοντας να επεξεργαστεί συγκεντρωτικά τα σύνολα δεδομένων για να εκπαιδεύσει ένα κρυπτογραφημένο ML μοντέλο [166].*

5.3.4.1 Εφαρμογές στη Μηχανική Μάθηση

Αυτή η μέθοδος έχει βρει εφαρμογή και σε εφαρμογές ML, όπου η ιδιωτικότητα των δεδομένων είναι υψίστης σημασίας, και η διακίνηση αυτών είναι απαγορευμένη. Έχουν προταθεί μέθοδοι και εργαλεία για MPC πάνω σε ML μοντέλα όπως SVM, NN, Decision Trees, K-means Clustering και άλλα [167]. Τα εργαλεία τα οποία επιτρέπουν την εφαρμογή MPC σε αλγόριθμους μηχανικής μάθησης, όπως π.χ. το CrypTen, προσφέρουν τη δυνατότητα εκτέλεσης των βασικότερων υπολογισμών που είναι απαραίτητη για τα περισσότερα ML frameworks

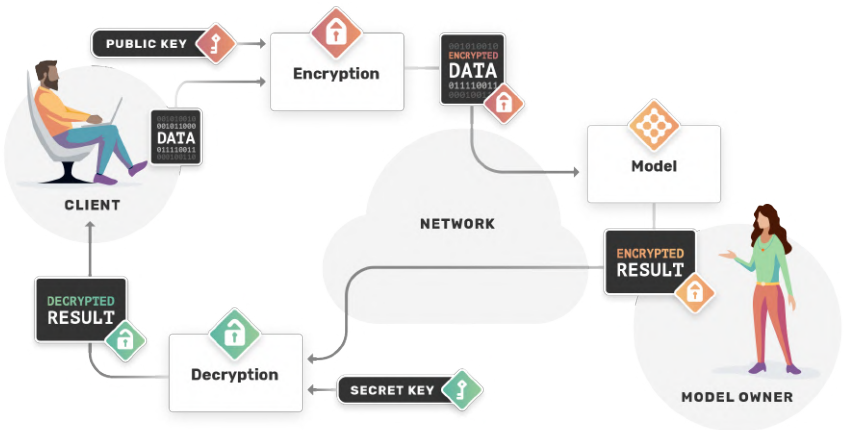
όπως tensor computations και automatic differentiation [166].

5.3.4.2 Προστασία Ιδιωτικότητας

Το **πλεονέκτημα** αυτή της μεθόδου υπολογισμών είναι ότι πολλαπλά μέρη μπορούν να έχουν κοινή κτήση ενός σύνολου δεδομένων, χωρίς όμως να ξέρουν μεμονωμένα τις όλες τις τιμές του, καθώς το κάθε μέρος είναι κρυπτογραφημένο, και το κάθε δεδομένο μπορεί να αξιοποιηθεί μονάχα αν επιτραπεί από κάθε μέλος του συνόλου που συνεισφέρει στον υπολογισμό. Επίσης, πέρα από τα δεδομένα, μπορεί και το ίδιο το μοντέλο να είναι κρυπτογραφημένο [167]. Το **μειονέκτημα** αυτή της μεθόδου υπολογισμών είναι ότι πρόκειται για μια υπολογιστικά ακριβή διαδικασία, λόγω της κρυπτογράφησης που εμπεριέχει και του κόστους επικοινωνίας μεταξύ των μελών, αλλά με πιο σύγχρονες μεθόδους και επεξεργαστές αυτό βελτιώνεται χρόνο με τον χρόνο [141]. Επίσης, ένα μοντέλο που έχει εκπαιδευτεί έτσι μπορεί να είναι ακόμα *ευάλωτο σε membership inference επιθέσεις*, καθώς συνήθως ο αλγόριθμος εκπαίδευσης είναι ο ίδιος όπως στις κλασσικές περιπτώσεις [141].

5.3.5 Homomorphic encryption

Η ομομορφική κρυπτογράφηση ή **Homomorphic encryption** είναι μια μορφής κρυπτογράφηση που επιτρέπει την εκτέλεση μαθηματικών υπολογισμών σε κρυπτογραφημένα δεδομένα, χωρίς να χρειάζεται πρώτα να αποκρυπτογραφηθούν, καταλήγοντας σε ένα κρυπτογραφημένο αποτέλεσμα, το οποίο όταν αποκρυπτογραφηθεί θα είναι πανομοιότυπο με αυτό που θα είχε παραχθεί αν οι υπολογισμοί γινόντουσαν στα αρχικά αποκρυπτογραφημένα δεδομένα εξ αρχής [168]. Στο επίπεδο εκπαίδευσης ML μοντέλων, αυτό επιτρέπει στο να εκπαιδευτούν μοντέλα με κρυπτογραφημένα δεδομένα, διασφαλίζοντας έτσι την ασφάλεια της ιδιωτικότητας των δεδομένων εκπαίδευσης, καθώς δεν μπορούν να παρθούν από το τελικό μοντέλο [154]. Ενώ αυτή η μέθοδος δεν προσφέρει εγγυήσεις ιδιωτικότητας όπως η differential privacy μέθοδος, προστατεύει το απόρρητο του κάθε δεδομένου κατά το inference ενός ML μοντέλου [4].



Σχήμα 5.9: *Σχηματικό παράδειγμα εφαρμογής Homomorphic Encryption σε ένα MLaaS περιβάλλον όπου τα δεδομένα του χρήστη-πελάτη κρυπτογραφούνται και γίνεται εκπαίδευση ενός κρυπτογραφημένου μοντέλου από τον ιδιοκτήτη της υπηρεσίας ή του μοντέλου, το οποίο μόνο ο αρχικός χρήστης μπορεί να αποκρυπτογραφήσει για να χρησιμοποιήσει*[3].

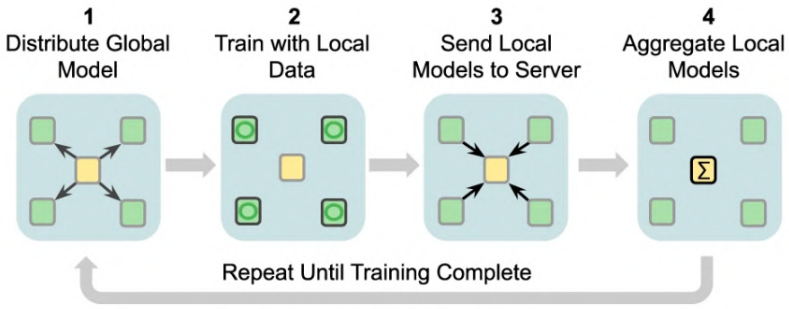
Το **πλεονέκτημα** αυτών των υπολογισμών είναι ξεκάθαρο, καθώς τα αρχικά δεδομένα

[3] https://blog.openmined.org/what-is-homomorphic-encryption/

δεν αποκαλύπτονται ούτε μοιράζονται παρά μόνο στον κάτοχο, καθώς είναι κρυπτογραφημένα, οπότε μπορούν να αξιοποιηθούν από άλλους χρήστες ή οργανισμούς και σε μη αξιόπιστα περιβάλλοντα. Πρόκειται επίσης για μια αποτελεσματική λύση σε model inversion επιθέσεις καθώς τα δεδομένα δε δίνονται στο μοντέλο στην αρχική τους μορφή, οπότε δεν υπάρχει και η δυνατότητα να εξαχθούν [143]. Το **μειονέκτημα** όμως, όπως και με τις MPC μεθόδους, είναι ότι πρόκειται για μια υπολογιστικά ακριβή και πολύπλοκη διαδικασία, λόγω της κρυπτογράφησης, η οποία δύσκολα κλιμακώνεται [157], όμως έχουν υπάρξει βελτιστοποιήσεις οι οποίες αποφεύγουν περιττούς υπολογισμούς και ειδικότερα πολλαπλασιασμούς που θα εκτελεστούν στα κρυπτογραφημένα δεδομένα [154] ή κάνουν χρήση προσεγγιστικών υπολογισμών για να μειωθεί ο χρόνος εκτέλεσης [169]. Επίσης, είναι αντίστοιχα *ευάλωτο σε membership inference επιθέσεις*, καθώς συνήθως ο αλγόριθμος εκπαίδευσης είναι ο ίδιος όπως στις κλασσικές περιπτώσεις, ακόμα και αν δεν υπάρχει πρόσβαση στα αρχικά δεδομένα [141].

5.3.6 Federated Learning

Το **Federated Learning (FL)** [170] πρόκειται για μια τεχνική κατανεμημένης εκπαίδευσης αλγορίθμων μηχανικής μάθησης, όπου η εκπαίδευση με τα ακατέργαστα δεδομένα πραγματοποιείται σε τοπικά μοντέλα σε συσκευές (local ή edge nodes) και δεν απαιτεί ανταλλαγή δεδομένων από συσκευές πελάτη σε κεντρικούς κόμβους, αλλά κοινοποιούνται μόνο οι ενημερώσεις μοντέλων, αυξάνοντας έτσι το απόρρητο των δεδομένων. Παρουσιάστηκε πρώτη φορά το 2017, από έρευνα προερχόμενη από την Google και είχε απευθείας εφαρμογή, χρησιμοποιώντας αυτήν τη μέθοδο για να βελτιώσουν τις προβλέψεις που παρείχε το Google Keyboard (Gboard), όπως πρόβλεψης επόμενης λέξης ή πρόταση emoji, αξιοποιώντας τα android κινητά των χρηστών για να εκπαιδευτούν τοπικά μοντέλα με τα δεδομένα τους (όταν το κινητό είναι ανενεργό, συνδεδεμένο στο wifi και φορτίζει), χωρίς την ανάγκη μεταφοράς τους στο cloud, και με κρυπτογραφημένη κοινοποίησή των αλλαγών τον τοπικών μοντέλων κεντρικά, ώστε να εκπαιδευτεί το κεντρικό ML μοντέλο (global model). Τέλος, οι αλλαγές του τελικώς εκπαιδευμένου κεντρικού μοντέλου, μοιράζονται πίσω στις συσκευές υπό τη μορφή update για να ολοκληρωθεί η μετάδοση του εκπαιδευμένου μοντέλου (βλ. σχήμα 5.10).



Σχήμα 5.10: *Σχηματικό παράδειγμα εφαρμογής Federated Learning, όπου στέλνονται αρχικά στα local nodes τα βάρη του μοντέλου, ο κάθε node εκπαιδεύει ένα τοπικό μοντέλο, μοιράζει τις αλλαγές των βαρών του στο κεντρικό, το κεντρικό μοντέλο συναθροίζει (aggregation) όλα τα βάρη που λαμβάνει και τέλος ανανεώνει τα βάρη όλων των local nodes σύμφωνα με τα τελικά βάρη [171].*

Οι χρήσεις του FL είναι πάρα πολλές και έχουν μέγιστη εφαρμογή σε περιπτώσεις που

υπάρχουν πολλά δεδομένα κατανεμημένα σε πολλούς κόμβους (χρήστες, βάσεις δεδομένων, εταιρίες ή οργανισμούς), τα οποία όμως απαιτείται να μην μπορούν να συγκεντρωθούν μαζικά, διότι θα παραβιαστεί η ιδιωτικότητα των δεδομένων ή γιατί θα είναι πολύ ακριβό υπολογιστικά να αξιοποιηθούν όλα σε ένα κεντρικό μοντέλο [172]. Τέτοιες εφαρμογές, πέρα αυτή των κινητών συσκευών που είδαμε, είναι π.χ. σε νοσοκομεία ή φαρμακευτικές εταιρίες για την αποφυγή κοινοποίησης ευαίσθητων προσωπικών δεδομένων [173] ή σε μοντέλα αυτόνομη οδήγησης όπου υπάρχει τεράστιος όγκος δεδομένων που παράγεται από ένα όχημα καθημερινά [174].

5.3.6.1 Προστασία Ιδιωτικότητας

Η FL τεχνική προσφέρει *προστασία στα δεδομένα εκπαίδευσης*, καθώς δε χρειάζεται να είναι διαθέσιμα αποκλειστικά σε ένα μοντέλο, και ο κεντρικός κόμβος λαμβάνει μόνο τα ανανεωμένα βάρη του μοντέλου, όμως αυτό από μόνο του δεν αρκεί για την προστασία απέναντι σε membership inference ή model inversion επιθέσεις, καθώς το μοντέλο μπορεί πάλι να κάνει memorize ολόκληρα δεδομένα από κάποιο τοπικό μοντέλο [157]. Για την προστασία από τέτοιες επιθέσεις, μπορεί να γίνει χρήση *Differential Privacy μαζί με Federated Learning* [175], για την προσφορά εγγυήσεων για την ιδιωτικότητα των δεδομένων των χρηστών. Υπάρχουν πολλές τεχνικές που συνδυάζουν αυτές τις δύο μεθόδους, αλλά η γενική ιδέα είναι ότι γίνεται εισαγωγή θορύβου είτε από τους τοπικούς κόμβους στις εκπαιδευμένες παραμέτρους του τοπικού μοντέλου, πριν σταλθούν και γίνει η συνάθροιση (Locally Differentially Private Federated Averaging), είτε από το μοντέλο κεντρικά μετά τη συνάθροιση (Centrally Differentially Private Federated Averaging) [157].

Τα **πλεονεκτήματα** γενικότερα της FL τεχνική είναι η δυνατότητα κατανεμημένης εκπαίδευσης των δεδομένων χωρίς ανταλλαγές δεδομένων, το οποίο ενισχύει την ιδιωτικότητα των ακατέργαστων δεδομένων εκπαίδευσης, αλλά και τον εκπαιδευμένων παραμέτρων του μοντέλου και μειώνει τον όγκο των δεδομένων που μεταφέρονται στο δίκτυο. Επίσης, λόγω της κατανεμημένης εκπαίδευσης, κατανέμετε και η υπολογιστική ισχύς που απαιτείται, χωρίς να χρειάζεται ένας πανίσχυρος κόμβος για να επεξεργαστεί τον μεγάλο όγκο των δεδομένων [172]. Αντίστοιχα όμως, τα **μειονεκτήματα** της FL τεχνικής είναι ότι υπάρχει μεγάλο κόστος επικοινωνίας, συντονισμού και διαθεσιμότητας κόμβων, λόγω της φύσης της κατανεμημένης εκπαίδευσης και της μεγάλης στατιστικής ετερογένειας που παρουσιάζεται στα δεδομένα εκπαίδευσης ανά κόμβο [157].

5.4 Εργαλεία Ανοιχτού Κώδικα και Βιβλιοθήκες

Για τη διασφάλιση της ασφάλεια και της ιδιωτικότητας των μοντέλων και των δεδομένων σε ML συστήματα (PPML), έχουν κυκλοφορήσει διάφορα εργαλεία και open-source βιβλιοθήκες μέχρι στιγμής, όπως για εφαρμογή Differential Privacy, Federated Learning, Secure Multi-party Computation, και Homomorphic encryption.

Παρακάτω παρουσιάζουμε μερικά από τα πιο γνωστά και ευρέως χρησιμοποιούμενα εργαλεία για την ενίσχυση της ιδιωτικότητας των δεδομένων στα ML μοντέλα. Στον πίνακα 5.5 υπάρχει μια σύνοψη όλων των εργαλείων και βιβλιοθηκών που παρουσιάζουμε, με αναφορά των κύριων χαρακτηριστικών τους.

Βιβλιοθήκη	ML Frameworks	Πεδία	Τεχνικές
TensorFlow Privacy	TensorFlow	*generic*	Differential Privacy
PyTorch Opacus	PyTorch	*generic*	Differential Privacy
TensorFlow Federated	TensorFlow	*generic*	Federated Learning
CrypTen	PyTorch	*generic*	Secure Multiparty Computation
PySyft	*agnostic*	*generic*	Homomorphic encryption, Differential Privacy, Federated Learning
SyferText	*agnostic*	NLP	Secure Multiparty Computation, Federated Learning

Πίνακας 5.5: *Εργαλεία και βιβλιοθήκες ανοιχτού κώδικα για εργασίες σχετικές με data privacy protection για* ***PPML****.*

5.4.1 TensorFlow Privacy

Το **TensorFlow Privacy**[4] είναι μια Python βιβλιοθήκη που περιλαμβάνει υλοποιήσεις TensorFlow optimizers (π.χ. SGD, Adam) για εκπαίδευση ML μοντέλων με χρήση differential privacy και εργαλεία για την ανάλυση και υπολογισμό των παρεχόμενων privacy guarantees, παρέχοντας τη δυνατότητα για εκπαίδευση privacy-preserving μοντέλων με λίγες επιπλέον γραμμές κώδικα. Η βιβλιοθήκη χρησιμοποιεί τον DP-SGD αλγόριθμο για την εκπαίδευση τον μοντέλων με DP για τον μετριασμό του κινδύνου έκθεσης ευαίσθητων δεδομένων εκπαίδευσης.

Πηγαίος κώδικας: https://github.com/tensorflow/privacy

5.4.2 PyTorch Opacus

Το **PyTorch Opacus** [176] είναι μια Python βιβλιοθήκη που επιτρέπει την εκπαίδευση PyTorch μοντέλων, με differential privacy, με ελάχιστες αλλαγές στον κώδικα και μικρό αντίκτυπο στην απόδοση της εκπαίδευσης. Επίσης, δίνεται η δυνατότητα στον χρήστη να παρακολουθήσει το privacy budget που δαπανάται ανά πάσα στιγμή. Η βιβλιοθήκη προσφέρεται για επαγγελματίες που θέλουν να εκπαιδεύσουν privacy-preserving μοντέλα, εύκολα και χωρίς πολλές αλλαγές, αλλά και για ερευνητές που θέλουν να πειραματιστούν. Η βιβλιοθήκη χρησιμοποιεί τον DP-SGD αλγόριθμο για την εκπαίδευση τον μοντέλων με DP, εκτελώντας διανυσματοποιημένους (vectorized) υπολογισμούς κλίσεων ανά δείγμα, το οποίο δείχνει να είναι 10 φορές ταχύτερο από την τεχνική του microbatching.

Πηγαίος κώδικας: https://github.com/pytorch/opacus

5.4.3 TensorFlow Federated

Το **TensorFlow Federated (TFF)**[5] πρόκειται για ένα Python framework για μηχανική μάθηση και εκτέλεση υπολογισμών σε κατανεμημένα δεδομένα, κυρίως αναπτυγμένο για να διευκολύνει την έρευνα και τον πειραματισμό με το Federated Learning, όπου ένα κοινό global μοντέλο εκπαιδεύεται σε πολλούς συμμετέχοντες clients που διατηρούν τα δεδομένα εκπαίδευσης τους τοπικά. Η βιβλιοθήκη παρέχει δομικά στοιχεία και διεπαφές υψηλού επιπέδου (APIs) που επιτρέπουν την υλοποίηση Federated Learning ή Federated Analytics και την αξιολόγηση τους σε υπάρχοντα TensorFlow μοντέλα. Η TFF βιβλιοθήκη χωρίζεται σε δύο επίπεδα: το *Federated Learning API* όπου προσφέρει τις διεπαφές υψηλού επιπέδου για υλοποίηση FL σε

[4] https://www.tensorflow.org/responsible_ai/privacy/guide
[5] https://www.tensorflow.org/federated

TensorFlow μοντέλα, και το *Federated Core API* όπου προσφέρει τις διεπαφές χαμηλότερου επιπέδου για υλοποίηση αλγορίθμων σε κατανεμημένα περιβάλλοντα.

Πηγαίος κώδικας: https://github.com/tensorflow/federated

5.4.4 CrypTen

Το **CrypTen** [166] πρόκειται για ένα Python framework που έχει δημιουργηθεί πάνω στο PyTorch για να διευκολύνει την έρευνα για PPML και διασφάλιση της ιδιωτικότητας, μέσω Secure Multi-party Computation ως μηχανισμό κρυπτογράφησης των δεδομένων μεταξύ πολλαπλών συμμετεχόντων. Το framework επιτρέπει σε ερευνητές που δεν είναι ειδικοί στην κρυπτογραφία να πειραματιστούν εύκολα με ML μοντέλα χρησιμοποιώντας τεχνικές ασφαλούς υπολογισμού, με αντικείμενα που μοιάζουν ακριβώς με PyTorch Tensors. ToCrypTen, παρόλο που είναι χτισμένο με γνώμονα πραγματικές προκλήσεις, πρόκειται κυρίως για ένα ερευνητικό εργαλείο το οποίο δεν έχει δοκιμαστεί στην παραγωγή.

Πηγαίος κώδικας: https://github.com/facebookresearch/CrypTen

5.4.5 PySyft

Το **PySyft** [177] είναι ένα Python framework που προσφέρει μεθόδους για την υλοποίηση ασφαλέστερων privacy-perserving ML μοντέλων και data science εργασιών. Το PySyft αποσυνδέει τα προσωπικά δεδομένα από την εκπαίδευση μοντέλων, χρησιμοποιώντας τεχνικές όπως Federated Learning, Differential Privacy και Encrypted Computation, στηριζόμενο σε διεπαφές που μοιάζουν με numpy πίνακες, ώστε να μπορούν να ενσωματωθούν εύκολα με οποιοδήποτε ML framework. Το PySyft επιτρέπει σε data scientists, να κάνουν ερωτήσεις σχετικά με ένα σύνολο δεδομένων, εντός κάποιον ορίων ιδιωτικότητας που έχει ορίσει ο κάτοχος των δεδομένων, και να λαμβάνει απαντήσεις, χωρίς να λαμβάνει αντίγραφο των ίδιων των δεδομένων, εξασφαλίζοντας ασφάλεια και προστασία των προσωπικών δεδομένων.

Πηγαίος κώδικας: https://github.com/OpenMined/PySyft

5.4.6 SyferText

Το **SyferText** [178] είναι μια Python βιβλιοθήκη για την προστασία της ιδιωτικότητας στον τομέα του Natural Language Processing (NLP), χρησιμοποιώντας τεχνικές όπως: Federated Learning και Secure Multi-Party Computations, για ML εφαρμογές σε ευαίσθητα δεδομένα που δεν μπορούν να κοινοποιηθούν ή συγκεντρωθούν σε κάποιο κεντρικό μηχάνημα. Το SyferText μπορεί να χρησιμοποιηθεί για εργασία σε σύνολα δεδομένων που βρίσκονται είτε τοπικά είτε σε απομακρυσμένα μηχανήματα. Τα δύο κύρια σενάρια χρήσης αυτής της βιβλιοθήκης είναι τα εξής: (i) *ασφαλής προεπεξεργασία plain-text κειμένων*, για την ασφαλή επεξεργασία δεδομένων κειμένου που βρίσκονται σε απομακρυσμένο μηχάνημα χωρίς την παραβίαση της ιδιωτικότητας, (ii) *ασφαλής ανάπτυξη NLP pipelines*, όπου τα στοιχεία που θα εκτελούν την προεπεξεργασία δεδομένων μέσω του SyferText, σε συνδυασμό με εκπαιδευμένα PPML μέσω του PySyft, θα γίνονται deploy με ασφάλεια.

Πηγαίος κώδικας: https://github.com/OpenMined/SyferText

Κεφάλαιο 6

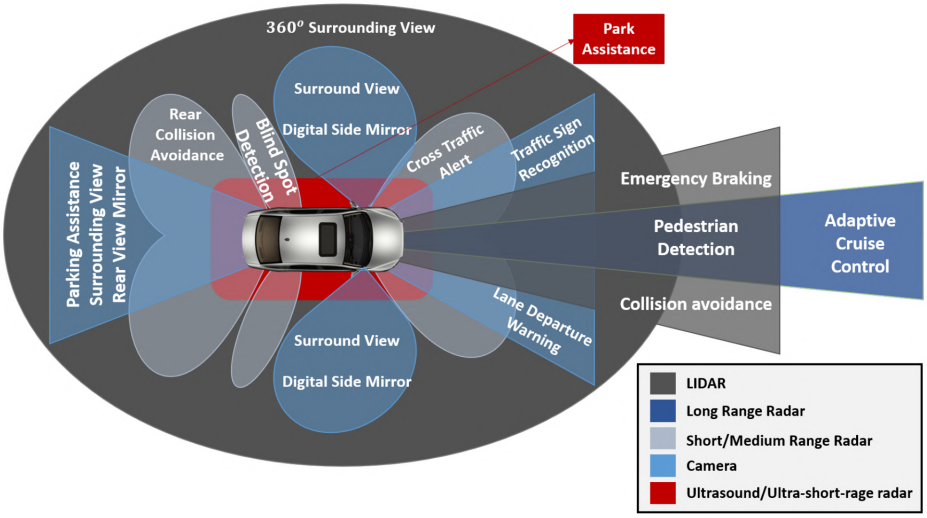
Μελέτη Ευρωστίας Συστημάτων Τεχνητής Νοημοσύνης σε Κρίσιμα Πεδία

Στο κεφάλαιο αυτό γίνεται εστίαση στη χρήση της Τεχνητής Νοημοσύνης (AI) σε κρίσιμα ερευνητικά και βιομηχανικά πεδία όπως τα Οχήματα και η Αυτόνομη Οδήγηση, η Φροντίδα Υγείας και Ιατρική Διάγνωση, και η Ενέργεια και τα Έξυπνα Δίκτυα. Τα συγκεκριμένοι πεδία έχουν επιλεχθεί λόγω της έντονης χρήσης του AI ειδικά τα τελευταία χρόνια και λόγο του ποικίλου φάσματος που αντιπροσωπεύουν συλλογικά αυτά τα πεδία. Κάθε πεδίο φέρνει στο προσκήνιο μοναδικές προκλήσεις και εφαρμογές, συνεισφέροντας διακριτές διαστάσεις στον γενικότερο διάλογο για την ασφάλεια και την ευρωστία της τεχνητής νοημοσύνης. Σε αυτές τις ενότητες, εμβαθύνουμε στην περίπλοκη αλληλεπίδραση μεταξύ τεχνητής νοημοσύνης και ασφάλειας σε αυτά τα πεδία, διερευνώντας σχολαστικά τις εφαρμογές της τεχνητής νοημοσύνης και τους εγγενείς κινδύνους που ενέχουν τα ανταγωνιστικά παραδείγματα και ελέγχοντας τις ανταγωνιστικές επιθέσεις που είναι πιο επικίνδυνες για αυτά τα AI συστήματα ανά πεδίο. Επιπλέον, παρέχεται μια ανάλυση ισχυρών αμυντικών μηχανισμών προσαρμοσμένων για την ενίσχυση αυτών των πεδίων έναντι των απειλών, αλλά και τρόπους για την αξιολόγηση αυτών των αμυντικών στρατηγικών.

6.1 Οχήματα και Αυτόνομη Οδήγηση

Το πεδίο της οδήγησης και των αυτοκινήτων έχει αναπτυχθεί ραγδαία τα τελευταία χρόνια, με την προσθήκη όλο και περισσότερης τεχνολογίας μέσα στα αυτοκίνητα, όσο και λόγο της ανάπτυξης και χρήσης ηλεκτροκίνητων οχημάτων. Η αυτοκινητοβιομηχανία έχει μετασχηματιστεί, ώστε να παράγει αυτοκίνητα με τεχνολογία αιχμής για να βελτιώσει την οδηγική εμπειρία. Η έλευση των προηγμένων συστημάτων υποστήριξης οδηγού (**Advanced Driver Assistance System - ADAS**) αλλά και των συστημάτων αυτόνομης οδήγησης (**Autonomous Driving Systems - ADS**) οδήγησε σε αυτή την επανάσταση, με την πρωτοπόρο της βιομηχανίας Tesla να πρωτοστατεί, αλλά και με άλλες εταιρίες όπως η Uber και η Google (πλέον Waymo) [179]. Η τεχνητή νοημοσύνη (AI) έχει παίξει προφανώς σημαντικό ρόλο στην ανάπτυξη αυτού του πεδίου, προσφέροντας τη δυνατότητα αυτόνομης οδήγησης διαφόρων επιπέδων, από απλή υποβοηθητική οδήγηση (**Driving Assistance**) σε πλήρη αυτόνομη οδήγηση (**Autonomous or Self Driving**). Για την εξέλιξη αυτή, κύριο ρόλο παίζουν επίσης και οι αισθητήρες οι οποίοι προσφέρουν τις κατάλληλες εισόδους στα AI συστήματα, όπως κάμερες, RADAR, LIDAR (Light Detection and Ranging), και αισθητήρες υπερήχων, οι οποίο χρησιμοποιούνται στρατηγικά για την παροχή δεδομένων από πολλαπλές διαφορετικών εμβελειών, όπως φαίνεται και στο σχήμα 6.1.

Μια ακόμη εξέλιξη σχετική με την αυτόνομη οδήγηση, είναι και τα Αυτόνομα Διασυνδε-

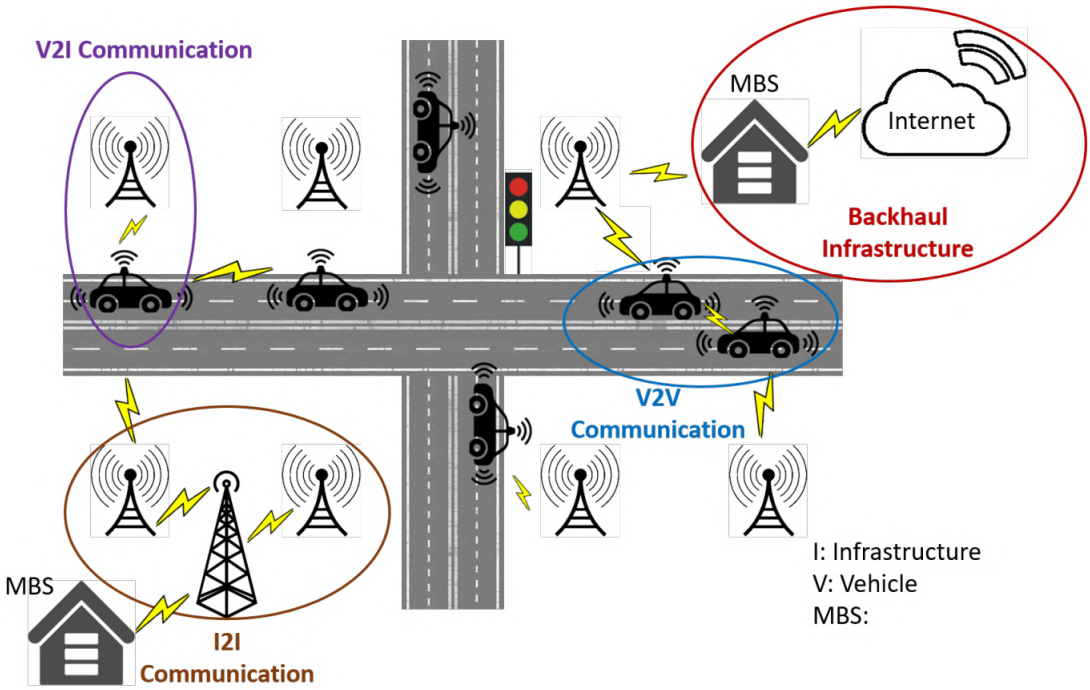


Σχήμα 6.1: *Σχηματικό παράδειγμα των κύριων αισθητήρων που μπορεί να υπάρχουν σε ένα όχημα αυτόνομης οδήγησης, όπου φαίνεται ενδεικτικά η εμβέλεια και η τοποθέτηση του κάθε είδους αισθητήρα [180].*

δεμένα Οχήματα (**Connected and Autonomous Vehicle - CAVs**), τα οποία πρόκειται για οχήματα τα οποία δεν είναι μόνο αυτόνομα αλλά είναι επίσης εξοπλισμένα με προηγμένα συστήματα επικοινωνίας, που τους επιτρέπουν να αλληλεπιδρούν έξυπνα μεταξύ τους και με τη γύρω υποδομή. Τα οχήματα αυτά μπορούν να επικοινωνούν μεταξύ τους (Vehicle-to-Vehicle - V2V), με τις κατάλληλες υποδομές (Infrastructure-to-Vehicle - I2V), αλλά και οι υποδομές μεταξύ τους (Infrastructure-to-Infrastructure - I2I) μέσω ανταλλαγής μηνυμάτων για την παροχή κατάλληλων κρίσιμων πληροφοριών (π.χ. κυκλοφοριακές και οδικές συνθήκες), όπως φαίνεται και στο σχήμα 6.2. Πρόκειται για ένα ενεργό και ταχύτατα αναπτυσσόμενο πεδίο έρευνας, μαζί με το πεδίο της αυτόνομης οδήγησης, το οποίο έχει αναπτυχθεί κυρίως λόγο των μεθόδων μηχανικής μάθησης (ML), και ιδιαίτερα της βαθιάς μάθησης (DL), που χρησιμοποιούνται για τη λήψη αποφάσεων σε διαφορετικά επίπεδα [180].

Μια ειδική αναφορά αξίζει να γίνει και στα **επίπεδα αυτόνομης οδήγησης** που έχει ορίσει η Εταιρεία Μηχανικών Αυτοκινήτου (Society of Automotive Engineers - SAE), τα οποία είναι 6 στο σύνολο. Ο SAE έχει ορίσει τις δυνατότητες του αυτοματισμού οδήγησης σε κάθε ένα επίπεδο, τα οποία φαίνονται στον πίνακα 6.1. Αυτήν τη χρονική στιγμή τα περισσότερα μοντέλα αυτοκίνητων με αυτονομία οδήγησης ανήκουν στο επιπέδου 2 όπως το Tesla Autopilot (αν και υπάρχει διαφωνία σχετικά με το αν ανήκει στο επίπεδο 2 ή 3)[1], πολύ λιγότερα στο επίπεδο 3 όπως το Drive Pilot σύστημα της Mercedes-Benz's[2], ελάχιστα στο επίπεδο 4 όπως της Waymo[3] τα οποία όμως πρόκειται για δοκιμαστικά και πιλοτικά μοντέλα, και κανένα στο επίπεδο 5 ακόμα[4].

[1] https://www.jurist.org/commentary/2021/09/william-widen-philip-koopman-autonomous-vehicles
[2] https://www.mbusa.com/en/owners/manuals/drive-pilot
[3] https://waymo.com
[4] https://www.cnet.com/roadshow/news/self-driving-car-guide-autonomous-explanation



Σχήμα 6.2: *Σχηματικό παράδειγμα της αρχιτεκτονικής ενός βασικού συστήματος CAV με τις 3 ειδών επικοινωνίες: V2V, I2I, I2V [180].*

6.1.1 Εφαρμογές Τεχνητής Νοημοσύνης στο πεδίο

Όπως ήδη αναφερθήκαμε, όλες οι νέες δυνατότητες στο πεδίο της οδήγησης έχουν γίνει εφικτές χάρη στο AI/ML και ειδικότερα στην ανάπτυξη και βελτίωση των Deep Learning τεχνικών και των DNN, τα οποία χρησιμοποιούνται για τη λήψη αποφάσεων σε διαφορετικά επίπεδα [180]. Οι εφαρμογές που έχει η τεχνητή νοημοσύνη στο πεδίο της οδήγησης και των αυτοκινήτων ποικίλει και περιλαμβάνει αυτά που αναφέραμε παραπάνω, όπως: Αυτόνομη οδήγηση (**ADS**), Υποβοηθητική οδήγηση (**Driving Assistance**), Αυτόνομα διασυνδεδεμένα δίκτυα οχημάτων (**CAVS**), Αναγνώρισή του περιβάλλοντός (**Environment Recognition**) αλλά και Παρακολούθηση του οδηγού (**Driver Monitoring**).

Για τις εφαρμογές αυτόνομης οδήγησης, ο τρόπος που δουλεύουν συνήθως τα μοντέλα αυτά είναι ο εξής (όπως φαίνεται και ενδεικτικά στο σχήμα 6.3): (i) Δέχονται δεδομένα εισόδου από τους αισθητήρες (π.χ. LiDAR και κάμερες), (ii) ένα βαθύ νευρωνικό δίκτυο προβλέπει τον έλεγχο του οχήματος (συνήθης πρόκειται για CNNs καθώς έχουν εξαιρετική απόδοση, απαιτούν λιγότερους νευρώνες και καταναλώνουν λιγότερους πόρους), και (iii) οι αποφάσεις του νευρωνικού καταλήγουν σε κινήσεις του οχήματος μέσω αλλαγής της γωνίας του τιμονιού ή της ταχύτητας.

Σχετικά με την αυτόνομη οδήγηση, οι λειτουργίες και οι διαδικασίες που έχουν να διαχειριστούν τα συστήματα αυτά είναι πολλές και συνήθως κάποιο νευρωνικό ή Deep Learning διαδικασία είναι υπεύθυνη για αυτές. Μία από τις πιο σημαντικές είναι και ο υπολογισμός ή πρόβλεψη πορείας (**Trajectory Prediction**), η οποία είναι υπεύθυνη για την πρόβλεψη μελλοντικών χωρικών συντεταγμένων διαφόρων οδικών παραγόντων, όπως οχημάτων και πεζών [181]. Μία άλλη σημαντική λειτουργία είναι και η παρακολούθηση του οδηγού (**Driver**

Επίπεδο	Ονομασία	Περιγραφή
0	No automation	Καμία αυτονομία: Όλες οι εργασίες οδήγησης και τα κύρια συστήματα ελέγχονται από έναν άνθρωπο οδηγό.
1	Function Specific Automation	Αυτονομία ειδικής λειτουργίας: παρέχει περιορισμένη βοήθεια οδηγού, π.χ. έλεγχος πλευρικής ή διαμήκους κίνησης.
2	Partial Driving	Μερική αυτονομία οδήγησης: τουλάχιστον δύο κύριες λειτουργίες ελέγχου συνδυάζονται για την εκτέλεση μιας ενέργειας, π.χ. υποβοήθηση διατήρησης λωρίδας και adaptive cruise control.
3	Conditional Driving Automation	Αυτονομία οδήγησης υπό συνθήκες: επιτρέπει την περιορισμένη αυτόνομη οδήγηση, δηλαδή επιτρέπει στον οδηγό να αποσπάσει προσωρινά την προσοχή του από την οδήγηση για να εκτελέσει άλλη δραστηριότητα, αλλά η παρουσία του οδηγού είναι πάντα απαραίτητη για να ανακτήσει τον έλεγχο μέσα σε λίγα δευτερόλεπτα.
4	High Driving Automation	Υψηλή αυτονομία οδήγησης: ένα αυτοματοποιημένο σύστημα οδήγησης που εκτελεί όλες τις δυναμικές εργασίες οδήγησης, π.χ. παρακολούθηση του περιβάλλοντος και έλεγχος κίνησης. Ωστόσο, ο οδηγός είναι σε θέση να έχει τον πλήρη έλεγχο των κρίσιμων για την ασφάλεια λειτουργιών του οχήματος υπό ορισμένα σενάρια.
5	Self-Driving Automation	Αυτόματη οδήγηση: ένα αυτοματοποιημένο σύστημα οδήγησης που εκτελεί όλες τις δυναμικές λειτουργίες οδήγησης και παρακολουθεί το κοντινό περιβάλλον για ολόκληρο το ταξίδι, χωρίς καμία ανθρώπινη παρέμβαση ανά πάσα στιγμή.

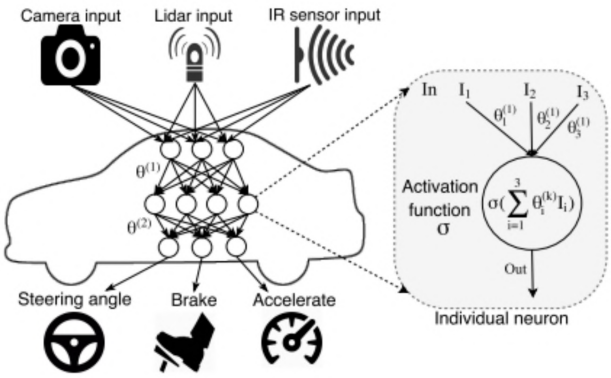
Πίνακας 6.1: *Πίνακας με τα επίπεδα αυτόνομης οδήγησης σύμφωνα με τον SAE [180].*

Monitoring), καθώς η υπνηλία και η απόσπαση της προσοχής του οδηγού είναι σημαντικοί παράγοντες σε μεγάλο αριθμό ατυχημάτων, και τέτοια μοντέλα είναι υπεύθυνα για την παρακολούθηση μέσω κάμερας των κινήσεων του οδηγού για ανίχνευση κλειστών ματιών ή χεριών που δε βρίσκονται στο τιμόνι για πολύ ώρα [182].

Κάποιες από τις πολλές λειτουργίες που εκτελεί ένα τέτοιο σύγχρονο σύστημα αυτόνομης ή βοηθητικής οδήγησης είναι και οι εξής: (i) Ανίχνευση και αποφυγή εμποδίων (Obstacle detection and avoidance), (ii) Πρόβλεψη λωρίδας (Lane prediction), (iii) Υποβοήθηση διατήρησης λωρίδας (Lane-keeping assistance), (iv) Αυτοματοποιημένο κεντράρισμα λωρίδας (Automated Lane Centering), (v) Προειδοποίηση εμπρόσθιας σύγκρουσης (Forward Collision Warning), (vi) Υπολογισμός ή πρόβλεψη πορείας (Trajectory Generation or Prediction), (vii) Αναγνώριση σημάτων οδήγησης (Traffic sign recognition), (viii) Προσαρμοστικό cruise control (Adaptive cruise control) [181], [183], [184].

6.1.1.1 Μοντέλα Αυτόνομης Οδήγησης

Πολλές εταιρίες και οργανισμοί έχουν επενδύσει στη δημιουργία και στην ανάπτυξη state-of-the-art AI μοντέλων αυτόνομης οδήγησης. Η Tesla αναπτύσσει και χρησιμοποιεί το **Autopilot**, του οποίου η αρχιτεκτονική δεν είναι διαθέσιμη στο ευρύ κοινό, αλλά είναι γνωστό ότι έχει περάσει από διάφορες εκδόσεις. Σαν είσοδο χρησιμοποιείται πλέον μόνο βίντεο που καταγράφουν οι 8 εξωτερικές κάμερες του, το οποίο περνάει σε νευρωνικά τύπου ResNets για



Σχήμα 6.3: *Η επισκόπηση μιας λειτουργίας αυτόνομης οδήγησης [7].*

την εξαγωγή των χαρακτηριστικών, τα οποία μετά συνδυάζονται και χρησιμοποιούνται από διάφορα νευρωνικά, όπως CNNs, RNNs, και Transformers για να γίνουν οι διάφορες διαδικασίες όπως πρόβλεψη λωρίδων, ανίχνευση εμποδίων και αναγνώριση σημάτων οδήγησης[5].

Η Waymo (θυγατρική της Alphabet) αναπτύσσει και χρησιμοποιεί το **ChauffeurNet**, το οποίο και αυτό δεν είναι διαθέσιμο. Σαν είσοδο χρησιμοποιεί πολλαπλά είδη εισόδων από κάμερες, LiDARs και RADARs, τα οποία τροφοδοτούνται σε ένα σύνολο νευρωνικών τύπου RNNs, τα οποία εκπαιδεύονται μέσω deep reinforcement learning για να πάρουν αποφάσεις και υπολογισμός πορείας[6].

Ένα λίγο διαφορετικό παράδειγμα είναι αυτό του **openpilot** της comma.ai, το οποίο πρόκειται για ένα open-source ADAS σύστημα του οποίου το hardware μπορεί να προσαρμοστεί σε πολλά οχήματα και αξιοποιώντας το CAN network του οχήματος να εκτελέσει λειτουργίες όπως προσαρμοστικό cruise control και υποβοήθηση διατήρησης λωρίδας. Το κύριο δίκτυο του συστήματος βασίζεται στο EfficientNet [185] της Google το οποίο είναι ένα CNN, στο οποίο τροφοδοτούνται frames εικόνων από τις κάμερες του οχήματος και είναι υπεύθυνο για την πρόβλεψη λωρίδων, την πρόβλεψη θέσεως και ταχύτητας των γύρω αντικειμένων και για την λήψη αποφάσεων [186].

6.1.2 Επιθέσεις - Κενά Ασφαλείας

6.1.2.1 Εφαρμογές Επιθέσεων

Η ταχεία ανάπτυξη και χρήση αλγόριθμων ML και Deep learning τεχνικών, ενώ έχουν βοηθήσει στη βελτίωση της αυτόνομης οδήγησης, μπορεί να είναι ευάλωτα σε επιθέσεις, όπως φυσικές επιθέσεις, κυβερνοεπιθέσεις αλλά και ανταγωνιστικές επιθέσεις που βασίζονται στη μάθηση [55], όπου και σε αυτές εστιάζουμε. Η ευαλωτότητα κυρίως των νευρωνικών δικτύων σε τέτοιες επιθέσεις μπορεί να θέσει σε κίνδυνο την ασφάλεια των αυτόνομων οχημάτων και των οδηγών. Πρόσφατες μελέτες έχουν δείξει ότι οι ανταγωνιστικές επιθέσεις μπορούν να προκαλέσουν σημαντική μείωση στην ακρίβεια ανίχνευσης, των DNN μοντέλων ανίχνευσης αντικειμένων σε 3D χώρο. Ενώ όμως η ασφάλεια οδήγησης είναι το απόλυτο μέλημα για την αυτόνομη οδήγηση, δεν υπάρχει ολοκληρωμένη μελέτη για τη σχέση μεταξύ της απόδοσης των μοντέλων βαθιάς μάθησης και της ασφάλειας οδήγησης των **αυτόνομων οχημάτων**

[5] https://www.thinkautonomous.ai/blog/how-tesla-autopilot-works
[6] https://www.thinkautonomous.ai/blog/how-googles-self-driving-cars-work

υπό ανταγωνιστικές επιθέσεις [187].

Ενώ οι επιθέσεις που μπορούν να εφαρμοστούν με σε συστήματα αυτόνομης οδήγησης, από εξωτερικούς εισβολείς μπορούν εύκολα να προκαλέσουν τροχαία ατυχήματα και να θέσουν σε κίνδυνο την προσωπική ασφάλεια, αντίστοιχοι κίνδυνοι υπάρχουν και από 'επιθέσεις' των ίδιων τον οδηγών οι οποίοι μπορεί να αναζητήσουν τρόπους για να ξεγελάσουν τα συστήματα **παρακολούθησης του οδηγού (driver monitoring)**. Αυτά τα συστήματα παρακολούθησης βασίζονται σε εικόνες ή βίντεο που καταγράφουν οι κάμερες στην καμπίνα του οχήματος και χρησιμοποιούνται ως είσοδοι σε μοντέλα βαθιάς εκμάθησης [182], τα οποία μπορεί να είναι και αυτά ευάλωτα στις ίδιες ανταγωνιστικές επιθέσεις με τα συστήματα αυτόνομης οδήγησης.

Οι αλγόριθμοι λοιπόν που βασίζονται σε Deep learning τεχνικές και αξιοποιούνται σχεδόν σε όλα τα συστήματα σχετιζόμενα με την αυτόνομη οδήγηση, ακόμη και μια μικρή αλλαγή στις εικόνες εισόδου μπορεί να οδηγηθούν σε εντελώς λανθασμένα αποτελέσματα με αρκετά μεγάλη πιθανότητα. Αυτό εγείρει ανησυχίες σχετικά με τη χρήση του Deep learning σε κρίσιμες για την ασφάλεια εφαρμογές αυτοκινήτων. Πιο συγκεκριμένα τα **μοντέλα ταξινόμησης (classification)** που βασίζονται σε CNN αρχιτεκτονικές, μπορούν εύκολα να ξεγελαστούν από adversarial examples, τα οποία κατασκευάζονται με την εφαρμογή μικρών διαταραχών σε επίπεδο pixel στις εικόνες εισόδου. Ωστόσο, δεν είναι σίγουρο σε ποιο βαθμό τα **μοντέλα παλινδρόμησης (regression)**, όπως τα μοντέλα οδήγησης, είναι ευάλωτα σε ανταγωνιστικές επιθέσεις, με τον ίδιο τρόπο που είναι τα μοντέλα ταξινόμησης, όπως και πόσο αποτελεσματικές είναι και οι τεχνικές άμυνες που έχουν αναπτυχθεί. Για ένα μοντέλο ταξινόμησης εικόνας, μια ανταγωνιστική επίθεση μπορεί να θεωρηθεί επιτυχής όταν μια ανταγωνιστική είσοδος ταξινομηθεί σε διαφορετική κατηγορία σε σύγκριση με την αρχική. Ωστόσο, στα μοντέλα αυτόνομης οδήγησης, που πρόκειται συνήθως για μοντέλα παλινδρόμησης που προβλέπουν συνεχείς τιμές, οι αντίθετες επιθέσεις ορίζονται σε σχέση με ένα αποδεκτό εύρος σφάλματος, γνωστό ως **adversarial threshold** [7].

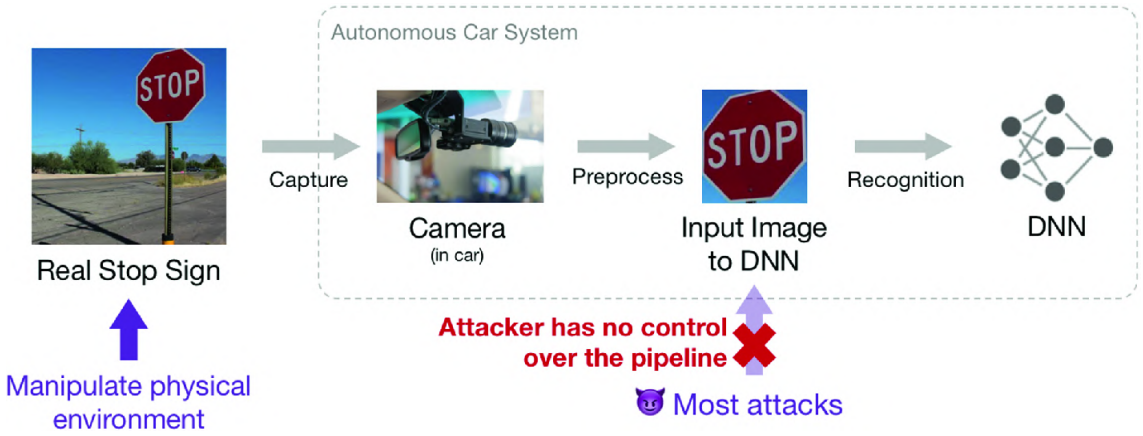
6.1.2.2 Επιθέσεις στον Πραγματικό Κόσμο

Οι επιθέσεις με adversarial examples μπορούν να τροφοδοτηθούν απευθείας στο ML σύστημα σαν είσοδοι, κάτι το οποίο προϋποθέτει κάποια άλλη εισβολή στο σύστημα που να επιτρέπει αυτήν την παραποίηση της εισόδου, αλλά μπορούν και πολύ απλά να δημιουργηθούν και σε συνθήκες του φυσικού κόσμου, τα οποία θα καταγραφούν από τους αισθητήρες και τις κάμερες τους συστήματος (βλ. σχήμα 6.4) και θα προκαλέσουν σφάλμα στο ML σύστημα εφόσον είναι ευάλωτο σε τέτοια δείγματα, όπως είχε αναδειχθεί στο [46].

Παρά το ότι τα βαθιά νευρωνικά δίκτυα είναι ευάλωτα σε ανταγωνιστικές επιθέσεις, μέσω μεταλλάξεων των εικόνων, οι επιβεβαιωμένες επιδείξεις επιτυχημένων επιθέσεων end-to-end, οι οποίες δηλαδή χειραγωγούν το φυσικό περιβάλλον και οδηγούν σε φυσικές επιπτώσεις, είναι σπάνιες, καθώς συνήθως περιλαμβάνουν προσεκτικά κατασκευασμένα adversarial examples σε επίπεδο pixel [189]. Αυτό όμως δε σημαίνει ότι δεν έχουν γίνει επιτυχημένες επιδείξεις, ή ότι το πρόβλημα δεν υπάρχει.

6.1.2.3 Παραδείγματα Μη-Κακόβουλων Επιθέσεων

Από τις αρχές τις αυτόνομης οδήγησης υπήρχαν ατυχήματα τα οποία δεν είχαν προκληθεί από κάποιον εξωτερικό κακόβουλο αντίπαλο, αλλά λόγω **ακούσιων ανταγωνιστικών**



Σχήμα 6.4: *Σχηματική απεικόνιση των φυσικών ανταγωνιστικών επιθέσεων, από την οπτική γωνία των επιτιθεμένων, καθώς συνήθως δεν έχουν τον πλήρη έλεγχο του computer vision συστήματος [188].*

συνθηκών. Τα μοντέλα επειδή δεν ήταν αρκετά ανθεκτικά σε αντίξοες συνθήκες, προκαλούσαν ατυχήματα λόγο κάποιας δυσλειτουργίας του μοντέλου [180]. (i) Για παράδειγμα το 2014, κατά τη διάρκεια ενός διαγωνισμού της Hyundai, ένα αυτόνομο όχημα συνετρίβη λόγω βλάβης του αισθητήρα ο οποίος λόγω της γωνίας που είχε το αυτοκίνητο είχε μετατοπιστεί προς την κατεύθυνση του ήλιου, με αποτέλεσμα να μην μπορέσει να αναγνωρίσει τις γραμμές του δρόμου και να τρακάρει[7]. (ii) Ένα άλλο περιστατικό αναφέρθηκε το 2016 όπου το autopilot της Tesla δεν μπόρεσε να ξεχωρίσει των φωτεινό ουρανό και ένα άσπρο φορτηγό με αποτέλεσμα να συγκρουστεί το όχημα με το φορτηγό κάτι είχε ως αποτέλεσμα τον θάνατο του οδηγού[8]. (iii) Ένα αντίστοιχο παράδειγμα συνέβη το 2016 με το αυτοκίνητο της Google (Waymo) όπου το αυτοκίνητο δεν μπόρεσε να υπολογίσει τη σχετική ταχύτητα που είχε, με αποτέλεσμα τη σύγκρουση με ένα λεωφορείο, χωρίς να προκαλέσει σοβαρό τραυματισμό[9].

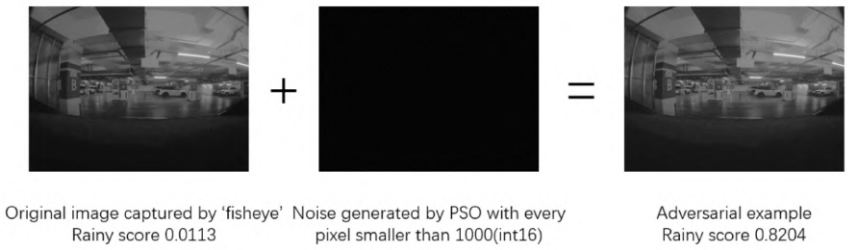
6.1.2.4 Παραδείγματα Κακόβουλων Επιθέσεων

Στις περιπτώσεις των κακόβουλων επιθέσεων, όπου κάποιος adversary έχει σκοπό να προκαλέσει δυσλειτουργία στο σύστημα, υπάρχουν πολλά παραδείγματα στη βιβλιογραφία και στις ειδήσεις σχετικά με τέτοιες επιθέσεις σε μοντέλα αυτόνομης οδήγησης.

Το 2019 από πειραματική έρευνα ασφαλείας του **Tencent Keen Security Lab πάνω στο Autopilot της Tesla** [190], κατάφεραν να επιδείξουν μια επιτυχημένη ανταγωνιστική επίθεση δημιουργώντας adversarial examples (βλ. εικόνα 6.5), τα οποία εμφανίζουν σε κάποια ηλεκτρονική οθόνη (π.χ. τηλεόραση, tablet) η οποία είναι στην εμβέλεια της κάμερας του αυτοκινήτου, για να ενεργοποιήσουν τους υαλοκαθαριστήρες ακόμα και σε συνθήκες που δεν υπάρχει βροχή, το οποίο από μόνο του δεν έχει κάποιες σοβαρές συνέπειες, όμως αποδεικνύει ότι οι φυσικές επιθέσεις πάνω στους αλγορίθμους αναγνώρισης εικόνας είναι πιθανές.

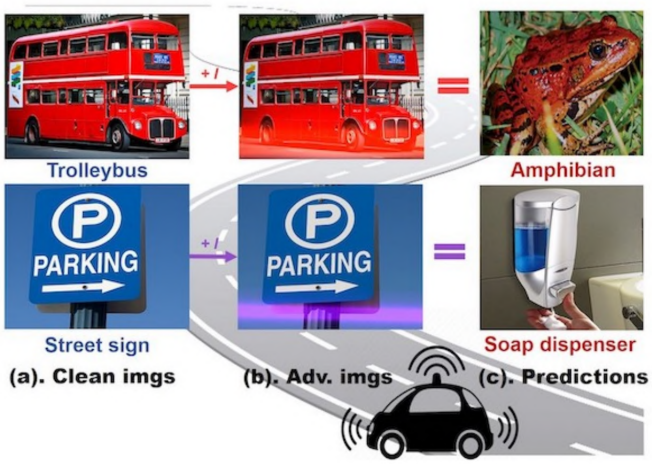
Ένα ακόμα παράδειγμα ανταγωνιστικών επιθέσεων στον φυσικό κόσμο, είναι και η επίθεση **Adversarial Laser Beam** [191], στην οποία χρησιμοποιώντας μια δέσμη laser (είτε φυσικά ή ψηφιακά) πάνω σε κάποιο αντικείμενο, τα DNN μοντέλα ξεγελιούνται εύκολα και παράγουν

[7] https://jalopnik.com/this-is-how-bad-self-driving-cars-suck-in-the-rain-1666268433
[8] https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk
[9] https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report



Σχήμα 6.5: *Παράδειγμα εφαρμογής θορύβου με τη μέθοδο Particle Swarm Optimization algorithm (PSO) για την παραγωγή adversarial example το οποίο ανεβάζει δραματικά το σκορ βροχής της αρχικής εικόνας [190].*

εντελώς διαφορετικές προβλέψεις (π.χ. ένα τρόλεϊ που προβλέπεται ως αμφίβιο ή μια πινακίδα του δρόμου ως διανομέας σαπουνιού όπως φαίνεται και στην εικόνα 6.6).



Σχήμα 6.6: *Παράδειγμα της επίθεσης Adversarial Laser Beam όπου η κάμερα αυτόνομης οδήγησης ενός οχήματος συλλαμβάνει αντικείμενο που πυροβολείται από μια συγκεκριμένη δέσμη λέιζερ, αλλάζοντας τελείως τις προβλέψεις του μοντέλου [191].*

Ένα παρόμοιο παράδειγμα, είναι και οι επιθέσεις με χρήση ανταγωνιστικών αφισών ή αυτοκόλλητων (adversarial poster/sticker) [192], στις οποίες οι ερευνητές δημιουργώντας κατάλληλα perturbed εικόνες πάνω σε φυσικά αντικείμενα τα οποία είτε αγνοούνται είτε επισημαίνονται εσφαλμένα από μοντέλα ανίχνευσης αντικειμένων. Πιο συγκεκριμένα, υλοποιείται μια επίθεση λεγόμενη **Disappearance Attack**, κατά την οποία προκαλείται η «εξαφάνιση» μιας stop πινακίδας σύμφωνα με τον ανιχνευτή, είτε καλύπτοντας την πινακίδα με μια ανταγωνιστική αφίσα είτε προσθέτοντας ανταγωνιστικά αυτοκόλλητα πάνω στην πινακίδα (όπως φαίνεται και στην εικόνα 6.7), τα οποία έχουν δημιουργεί με χρήση white-box τεχνικών. Από θέμα απόδοσης, σε βίντεο που καταγράφηκε σε ελεγχόμενο εργαστηριακό περιβάλλον, ο υπερσύγχρονος YOLO v2 ανιχνευτής απέτυχε να αναγνωρίσει αυτές τις adversarial stop πινακίδας με σε πάνω από το 85% των καρέ του βίντεο και με τις δύο μεθόδους, ενώ σε εξωτερικό χώρο ξεγελάστηκε από τις επιθέσεις με αφίσα και αυτοκόλλητο στο 72,5% και 63,5% των καρέ βίντεο αντίστοιχα. Ένα ακόμα σημαντικό αποτέλεσμα είναι και το transferability (δυνατότητα μεταφοράς) της επίθεσης, καθώς στο Faster R-CNN μοντέλο κατάφερε να το ξεγελάσει στο 85,9% των καρέ του βίντεο σε ελεγχόμενο εργαστηριακό περιβάλλον, και στο 40,2% σε

εξωτερικό περιβάλλον, για τις επιθέσεις με την αφίσα.

(α′) *Χρήση adversarial poster*

(β′) *Χρήση adversarial sticker*

Σχήμα 6.7: *Παράδειγμα των Disappearance Attack επιθέσεων σε εξωτερικό χώρο, όπου η stop πινακίδα δεν αναγνωρίζεται καθόλου [192].*

Υπάρχουν όμως και κάποια άλλα παραδείγματα, όχι απαραίτητα επιθέσεων, αλλά περισσότερο καταχρήσεως του συστήματος αυτόνομης οδήγησης, που έχει να κάνει με χρήστες που προσπαθούν να ξεγελάσουν τα (driver monitoring) συστήματα. Το 2021 σε 2 διαφορετικά καταγεγραμμένα περιστατικά οι οδηγοί ξεγέλασαν το Autopilot της Tesla, ενώ καθόντουσαν στο πίσω κάθισμα, αφήνοντας τη θέση του οδηγού κενή[10] [11]. Βάζοντας ένα βάρος στο τιμόνι για να ξεγελάσουν τον αισθητήρα ροπής[12] και βάζοντας τη ζώνη ασφαλείας κλειδωμένη πίσω από την πλάτη, ώστε να μπορούν να βγουν από το κάθισμα ενώ το όχημα οδηγεί είναι μερικές από τις τεχνικές για την κατάχρηση του συστήματος οδήγησης.

6.1.2.5 Κατηγορίες Επιθέσεων

Αφού έχουμε αναφέρει τα κενά ασφαλείας που υπάρχουν στα ML συστήματα στο πεδίο της αυτόνομης οδήγησης και έχουμε περιγράψει κάποια παραδείγματα ανταγωνιστικών επιθέσεων, σε αυτήν την ενότητα γίνεται μια προσπάθεια ταξινόμησης αυτών με βάση την ταξινόμηση και το μοντέλο απειλής που έχουμε κάνει στην ενότητα 3.4.1 για τις ανταγωνιστικές επιθέσεις γενικότερα.

Αρχικά, μπορούμε να κατηγοριοποιήσουμε τις επιθέσεις στα συστήματα αυτόματης οδήγησης ανάλογα με τα πιθανά σημεία επίθεσης:

- **Φυσικά Συστήματα**: Αισθητήρες (LiDAR, Radar, Κάμερες, GPS), Σύστημα Οχήματος (OBD, CAN-bus), Σύστημα τροφοδοσίας, κ.λπ.
- **Λογισμικό**: Εγκατεστημένες Εφαρμογές, Σύστημα Entertainment, Σύστημα επεξεργασίας δεδομένων, Σύστημα Πλοήγησης, κ.λπ.
- **Δεδομένα**: Τοπικά δεδομένα (προσωπικά δεδομένα χρηστών, vehicle ID), Δεδομένα ανταλλαγής (Ταχύτητα, Κατάσταση πέδησης)

[10] https://jalopnik.com/another-video-shows-a-driver-abusing-tesla-autopilot-on-1846851450
[11] https://electrek.co/2021/01/20/tiktok-star-criminally-tesla-autopilot-posts-video-evidence
[12] https://www.consumerreports.org/autonomous-driving/cr-engineers-show-tesla-will-drive-with-no-one-in-drivers-seat

- **Κανάλια Επικοινωνίας**: Vehicle-to-Infrastructure (V2I), Vehicle-to-Vehicle (V2V), Vehicle-to-Cloud (V2C) και Vehicle-to-Everything (V2X)

Με την αυξανόμενη ποσότητα των λειτουργιών αυτονομίας και συνδεσιμότητας, υπάρχουν ήδη και θα υπάρξουν περισσότερες ευπάθειες (vulnerabilities) ή σημεία επίθεσης. Για τις ανταγωνιστικές επιθέσεις, το πιο συχνό σημείο εφαρμογής είναι είτε τα φυσικά συστήματα όπως οι αισθητήρες τα οποία μπορούν να καταγράψουν τα adversarial examples είτε το ίδιο το λογισμικό στο οποίο μπορούν να τα τροφοδοτήσουν απευθείας [193]. Τέλος, επιθέσεις στα δεδομένα είναι πιθανές πηγές παραβίασης της ιδιωτικότητας, για παράδειγμα μέσω εξαγωγής προσωπικών στοιχείων, είτε πιθανώς στόχος για poisoning attacks με σκοπό την εισαγωγή 'μολυσμένων' δειγμάτων τα οποία θα αξιοποιηθούν για μεταγενέστερη εκπαίδευση των μοντέλων.

Όσον αφορά το περιβάλλον εφαρμογής αυτών των επιθέσεων, υπάρχει μια διακριτή διαφοροποίηση μεταξύ των επιθέσεων που είναι διαθέσιμες στη βιβλιογραφία [55]:

- **Physical**: Επιθέσεις που συμβαίνουν στον πραγματικό (φυσικό) κόσμο.
- **Digital**: Επιθέσεις που συμβαίνουν στον ψηφιακό κόσμο, συνήθως σε περιβάλλον προσομοίωσης, είτε χρησιμοποιώντας δεδομένα από καταγραφή στον πραγματικό κόσμο, είτε χρησιμοποιώντας δεδομένα που έχουν προκύψει από προσομοιωμένα σενάρια.

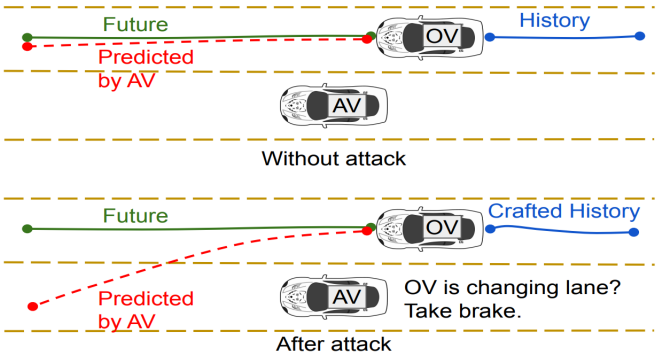
Οι Physical επιθέσεις είναι σαφώς πιο ισχυρές από αυτές που έχουν δημιουργηθεί και δοκιμαστεί μόνο ψηφιακά, καθώς πρέπει να έχουν αρκετά ανθεκτικές διαταραχές οι οποίες θα μπορέσουν να περάσουν μέσα από την εκτύπωση ή σχεδιασμό και να φανούν και σε πραγματικές συνθήκες, δηλαδή από διάφορες αποστάσεις, από πολλαπλές γωνίες θέασης και σε οποιαδήποτε κατάσταση (με βροχή ή ήλιο, με χαμηλό ή υψηλό φωτισμό) [192]. Υπάρχει και η άποψη όμως ότι ακόμα οι physical επιθέσεις δεν αποτελούν υπαρκτό κίνδυνο, αλλά περισσότερο ερευνητικό πεδίο με μικρή εφαρμογή, καθώς όντως χρειάζεται πολύ καλή υλοποίηση και προσπάθεια για να μπορέσουν οι αισθητήρες των οχημάτων να περάσουν τις διαταραχές στην είσοδο του μοντέλου [194].

Επίσης, δεν έχουν όλες οι επιθέσεις τον ίδιο στόχο σε ένα αυτόνομο όχημα ή ένα σύστημα αυτόματης οδήγησης. Μπορούμε να κατηγοριοποιήσουμε τις επιθέσεις ανάλογα με τον στόχο του επιτιθέμενου, δηλαδή το μέρους του συστήματος που έχει σκοπό να βλάψει [55], [180] (διαφορετικό από τον γενικότερο στόχο των ανταγωνιστικών επιθέσεων που αναφέραμε στην ενότητα 3.4.1):

- Συνολικό μοντέλο οδήγησης (**End-to-End driving model**): Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος οδήγησης.
- Ανίχνευση αντικειμένων (**Object detection**): Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος ανίχνευσης αντικειμένων, είτε για τη μη αναγνώριση κάποιον αντικειμένων, είτε για την αναγνώριση μη υπαρκτών αντικειμένων.
- Αναγνώριση οδικών σημάτων (**Traffic sign recognition**): Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος αναγνώρισης οδικών σημάτων, είτε για τη λάθος ερμηνεία των σημάτων ή και για τη μη αναγνώριση τους.
- Παρακολούθηση του οδηγού (**Driver Monitoring**): Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος παρακολούθησης του οδηγού, είτε για τη λάθος ανα-

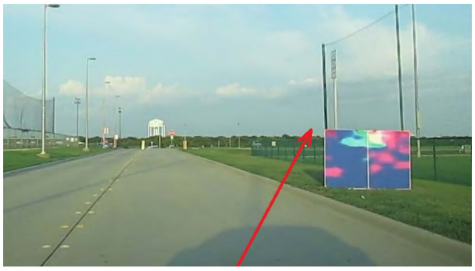
γνώριση της κατάστασης του οδηγού ή για τη μη αναγνώριση της παντελής απουσίας του οδηγού.

Οι επιθέσεις που στόχο έχουν το **End-to-End driving model**, ποικίλουν καθώς υπάρχουν παραδείγματα που προκαλούν δυσλειτουργία στην πρόβλεψη πορείας και την αλλαγή της γωνίας διεύθυνσης του αυτοκινήτου με χρήση (i) adversarial perturbation πάνω σε πινακίδες, ταμπέλες ή σήματα (βλ. DeepBillboard [195], PhysGAN [196]), είτε (ii) με την σχεδίαση μαύρων λωρίδων στο δρόμο (βλ. [197], [198], [189]). Μια υποκατηγορία του συνολικού μοντέλου οδήγησης είναι το υποσύστημα της πρόβλεψης τροχιάς (**Trajectory prediction**), και υπάρχουν αντίστοιχα επιθέσεις σε αυτό, όπου οι επιτιθέμενοι σε άλλα οχήματα υπολογίζουν μια adversarial τροχιά, οδηγούν πάνω σε αυτή κοντά στο άλλο όχημα με σκοπό να το κάνουν να νομίζει ότι οδηγάει σε τροχιά που θα καταλήξει σε πρόσκρουση, με αποτέλεσμα να αναγκάζει το όχημα σε απότομο φρενάρισμα ή ακόμα και πλήρης ακινητοποίηση [181] (βλ. σχήμα 6.8). Αντίστοιχα τα παραδείγματα των επιθέσεων με στόχο είτε το **Traffic sign recognition**, ή το **Object detection**, γίνονται κυρίως είτε με εφαρμογή (i) adversarial perturbation πάνω σε πινακίδες, ταμπέλες ή σήματα (βλ. CAMU [199], Rogue Signs [200], DARTS [201], Physical Examples [192]) είτε (ii) με χρήση αυτοκόλλητων με κατάλληλα σχέδια συνήθως για τον 'δηλητηριασμό' των δειγμάτων (βλ. ShapeShifter [188], [202], [203]). Τα παραδείγματα αυτά μπορεί να υπάρξουν είτε στον φυσικό είτε στον ψηφιακό κόσμο και συνήθως ο σχεδιασμός τους είναι αρκετά εύκολος και τα σημάδια αυτά είναι εύκολο να αγνοηθούν από τους ανθρώπους αλλά αρκετά πιθανό να ξεγελάσουν τα μοντέλα αυτόνομης οδήγησης. Για τις επιθέσεις με στόχο το **Driver Monitoring**, αυτές μπορεί να γίνουν είτε (i) με 'δηλητηριασμό' των δειγμάτων που συλλέγονται από τους αισθητήρες του οχήματος και ειδικότερα σε federated learning συστήματα που χρησιμοποιούνται για τη συλλογή δεδομένων εκπαίδευσης από πολλά οχήματα [204]), είτε (ii) με evasion τεχνικές για την εσφαλμένη ανίχνευση της κατάστασης του οδηγού [205].



Σχήμα 6.8: *Παράδειγμα σεναρίων επίθεσης εναντίων του συστήματος πρόβλεψης τροχιάς σε Autonomous Vehicles (AV). Στο πάνω σενάριο δεν υπάρχει επίθεση και στο κάτω σενάριο το Other Vehicle (OV) οδηγεί πάνω σε adversarial τροχιά. [181].*

Τώρα για την κατηγοριοποίηση των επιθέσεων με βάση την επιρροή ή την ικανότητα του επιτιθέμενου, μπορούμε να έχουμε Poisoning attacks και Evasion attacks. Αν συνυπολογίσουμε και τις γνώσεις του επιτιθέμενου για το σύστημα, δηλαδή White-box και Black-box επιθέσεις, τότε έχουμε τις εξής κατηγορίες:



(α′) *Χρήση adversarial billboards στον δρόμο (DeepBillboard [195])*

(β′) *Χρήση adversarial billboards στον δρόμο (PhysGAN [196])*

Σχήμα 6.9: *Παραδείγματα evasion attack με στόχο το* ***End-to-End driving model*** *σύστημα για την αλλαγή της πορείας του οχήματος.*

- **Evasion Attacks (White-box)**: Επιθέσεις με στόχο τη δημιουργία σφάλματος στο ML σύστημα, με κατάλληλο χειρισμό των δεδομένων εισόδου, έχοντας πλήρη γνώση του στοχευμένου συστήματος.
- **Evasion Attacks (Black-box)**: Επιθέσεις με στόχο την δημιουργία σφάλματος στο ML σύστημα, με κατάλληλο χειρισμό των δεδομένων εισόδου, μην έχοντας καμία γνώση του στοχευμένου συστήματος.
- **Poisoning Attacks**: Επιθέσεις με στόχο την παραποίηση των δεδομένων εκπαίδευσης, για την επίτευξη χειρότερης απόδοσης του εκπαιδευμένου ML συστήματος.

Οι **evasion** επιθέσεις που μπορούν να εφαρμοστούν στα μοντέλα αυτόνομης οδήγησης ποικίλουν, καθώς οι περισσότερες επιθέσεις που αφορούν μοντέλα ταξινομητών μπορούν να χρησιμοποιηθούν [7]. Για τις **White-box** επιθέσεις, υπάρχουν παραδείγματα με χρήση των κλασσικών μεθόδων για την παραγωγή adversarial perturbations, που έχουμε αναφέρει στην ενότητα 3.4, όπως (i) Fast Gradient Sign Method [1] ή Iterative Targeted Fast Gradient Sign Method [36], (ii) Universal Adversarial Perturbation [54], (iii) Optimization based επιθέσεις [31] και με χρήση (iv) GANs [206] όπως το AdvGAN [63]. Οι **Black-box** επιθέσεις, είναι πιο ρεαλιστικές καθώς στον πραγματικό κόσμο δε θα υπάρχει απαραίτητα γνώση του μοντέλου που χρησιμοποιεί το κάθε όχημα για αυτόνομη οδήγηση, και η ύπαρξη τους ενέχει σημαντικούς κινδύνους για πραγματικές επιθέσεις [55].

Οι **poisoning** επιθέσεις που μπορούν να εφαρμοστούν στα μοντέλα αυτόνομης οδήγησης ποικίλουν και αυτές, καθώς υπάρχουν παραδείγματα με χρήση (i) Backdoor attacks [203], (ii) Trojan attacks [207] και (iii) Out-of-Distribution attacks [201]. Σε αυτού του είδους τις επιθέσεις, συνήθως χρειάζεται ένα μικρό δείγμα δηλητηριασμένων δειγμάτων τάξεως του 5% με 10% και είναι δύσκολο να προσδιοριστεί εάν ένα μοντέλο αντιμετωπίζει επιθέσεις δηλητηρίασης παρατηρώντας μόνο τα αποτελέσματα της ακρίβειας της δοκιμής, καθώς η συνολική ακρίβεια παραμένει υψηλή και μόνο στα backdoor δείγματα εμφανίζεται υψηλή ακρίβεια για λάθος όμως πρόβλεψη [203].

Στους πίνακες 6.2, 6.3 φαίνονται συγκεντρωτικά διάφορες ερευνητικές και πρακτικές adversarial **evasion** και **poisoning** επιθέσεις αντίστοιχα, κατηγοριοποιημένες σύμφωνα με το παραπάνω μοντέλο απειλής. Στην περιγραφή της κάθε επίθεσης φαίνεται συνοπτικά η μέθοδος που ακολουθήθηκε για τη δημιουργία της, και η μέγιστη αρνητική επίπτωση που είχε στα

(α′) *Με χρήση adversarial poster έχουμε λανθασμένη ταξινόμηση της πινακίδας κατά 100% ως stop (Rogue Signs [200])*

(β′) *Με χρήση adversarial poster έχουμε λανθασμένη ταξινόμηση της πινακίδας κατά 78% ως μπάλα (ShapeShifter [188])*

Σχήμα 6.10: *Παραδείγματα evasion attack με στόχο το* ***Traffic sign recognition*** *σύστημα για τη λανθασμένη ταξινόμηση των πινακίδων.*

μοντέλα αυτόνομης οδήγησης που δοκιμάστηκε.

6.1.3 Άμυνες - Μέτρα Ασφαλείας

Για την αντιμετώπιση όλων των παραπάνω ανταγωνιστικών επιθέσεων, υπάρχει μια πληθώρα τεχνικών για την αμυντική θωράκιση των ML συστημάτων ενάντια σε τέτοιες επιθέσεις. Πολλές από αυτές τις άμυνες είναι προσαρμοσμένες για αντιμετώπιση συγκεκριμένων επιθέσεων ενώ άλλες προσφέρουν ασφάλεια από πολλαπλά είδη επιθέσεων. Η κύρια λειτουργία των περισσοτέρων αμυντικών μηχανισμών είναι η δημιουργία εύρωστων ML μοντέλων, χωρίς όμως πρόσθετη επιβάρυνση στην κανονική απόδοση του μοντέλου.

Οι περισσότερες υπάρχουσες άμυνες για ανταγωνιστικές επιθέσεις επικεντρώνονται στις εφαρμογές ταξινόμησης πάνω σε εικόνες, κάτι το οποίο μπορεί να έχει σχετικά εύκολη προσαρμογή και στα μοντέλα αυτόνομης οδήγησης, καθώς πρόκειται για εργασίες που επιτελούν τα περισσότερα μοντέλα ή για ιδέες που μπορούν να αξιοποιηθούν και σε άλλες εργασίες, όπως regression αντί για classification ή για εργασίες πάνω σε βίντεο αποτελούμενο από καρέ αντί για στατικές εικόνες [55]. Χρειάζεται όμως πολλές φορές προσοχή και έρευνα για το ποιες είναι κατάλληλες, καθώς ενώ μπορεί να οδηγήσουν σε ποιο εύρωστα μοντέλα, μπορεί να έχουν ανεπιθύμητες παρενέργειες όπως μείωση της κανονική τους απόδοσης, ή αύξηση της πολυπλοκότητας των μοντέλων κάτι που μπορεί να δημιουργήσει καθυστέρηση στις απαντήσεις του μοντέλου ή κατανάλωση περισσότερων υπολογιστικών και ενεργειακών πόρων [88].

6.1.3.1 Κατηγορίες Αμυνών

Σε αυτήν την ενότητα κάνουμε αναφορά στις τεχνικές αμυνών που μπορούν να χρησιμοποιηθούν για να αναπτυχθούν πιο εύρωστα ML **συστήματα αυτόνομης οδήγησης**, ενάντια σε ανταγωνιστικές επιθέσεις, και γίνεται μια προσπάθεια ταξινόμησης αυτών με βάση και την ταξινόμηση που έχουμε κάνει στην ενότητα 3.6.1 για τις άμυνες ενάντια σε ανταγωνιστικές επιθέσεις γενικότερα.

Αρχικά, μπορούμε να κατηγοριοποιήσουμε τις άμυνες στα συστήματα αυτόνομης οδήγησης, ανάλογα με το κύριο στόχο αντιμετώπισης των επιθέσεων, στις δύο γενικές τους κατηγορίες:

Στόχος	**Γνώση**	**Περιβάλλον**	**Περιγραφή**	**Αποτελέσματα**
E2E driving model	White-Box	Digital	DeepBillboard [195]: Δημιουργία adversarial διαφημιστικών πινακίδων μέσω επίλυσης προβλήματος βελτιστοποίησης	Δυσλειτουργία έως και 23 μοιρών σε γωνία διεύθυνσης του αυτόνομου οχήματος
E2E driving model	White-Box	Physical	PhysGAN [196]: Δημιουργία adversarial διαφημιστικών πινακίδων με χρήση GANs	Δυσλειτουργία έως και 19.17 μοιρών σε γωνία διεύθυνσης του αυτόνομου οχήματος
E2E driving model	Black-Box	Digital	[197]: Σχεδίαση μαύρων λωρίδων στο δρόμο μέσω Bayesian μεθόδου βελτιστοποίησης	Αλλαγή πορείας από δεξιά ή αριστερή στροφή σε ευθεία στο 25% των περιπτώσεων
E2E driving model	Black-Box	Digital	[198]: Σχεδίαση μαύρων λωρίδων στο δρόμο μέσω Gradient-based μεθόδου βελτιστοποίησης	Αλλαγή πορείας (δεξιά, αριστερή, ευθεία) πάνω από το 70% των περιπτώσεων
E2E driving model	White-Box	Digital	[189]: Σχεδίαση μαύρων λωρίδων στο δρόμο μέσω παραμετροποίησης των σχημάτων τους με στόχο τη βέλτιστη αλλαγή πορείας	Επιτυχημένος εξαναγκασμός τρακαρίσματος
Object Detection	Black-Box	Digital	CAMU [199]: Χρήση adversarial pattern καμουφλάζ πάνω σε αμάξια για την αποφυγή ανίχνευσης τους από Mask R-CNN μοντέλα	Μείωση της ακρίβειας ανίχνευσης κατά περίπου 40%
Traffic sign recognition	White-Box	Physical, Digital	[192]: Δημιουργία adversarial poster/sticker για αδυναμία ανίχνευσης οδικών σημάτων σε καρέ βίντεο από YOLO, R-CNN μοντέλα	Αδυναμία αναγνώρισης οδικών σημάτων σχεδόν στο 86% των καρέ του βίντεο
Traffic sign recognition	White-Box	Physical, Digital	Rogue Signs [200]: Δημιουργία adversarial διαφημιστικών ή οδικών πινακίδων για ξεγέλασμα μοντέλων ανίχνευσης βασισμένων στα CNN	Δυσλειτουργία με επιτυχία έως και 95% σε ψηφιακό ή φυσικό περιβάλλον
Traffic sign recognition	White-Box, Black Box	Physical, Digital	[202]: Δημιουργία φυσικών adversarial pertubration υπό μορφή ασπρόμαυρων αυτοκόλλητων πάνω σε οδικά σήματα για εσφαλμένη ταξινόμηση σε CNN μοντέλα	Σε κάποιες περιπτώσεις έως και 100% επιτυχία λανθασμένης ταξινόμησης σημάτων
Traffic sign recognition	White-Box	Physical, Digital	ShapeShifter [188]: Δημιουργία adversarial οδικών σημάτων μέσω επίλυσης προβλήματος βελτιστοποίησης για επιθέσεις σε Faster R-CNN μοντέλα	Δυσλειτουργία με επιτυχία έως και 93% σε μη αναγνώριση stop σημάτων

Πίνακας 6.2: *Πίνακας με state-of-the-art adversarial* ***Evasion*** *επιθέσεις σε συστήματα αυτόνομης οδήγησης.*

Στόχος	**Γνώση**	**Περιβάλλον**	**Περιγραφή**	**Αποτελέσματα**
Traffic sign recognition	White-Box, Black Box	Physical, Digital	DARTS [201]: Δημιουργία adversarial οδικών σημάτων με χρήση Out-of-Distribution και Lenticular Printing επιθέσεων	Επιτυχές ξεγέλασμα του μοντέλου σε όλα τα σενάρια
Traffic sign recognition	White-Box	Digital	[203]: Προσθήκη poisoning εικόνων οδικών σημάτων που φέρουν μικρά σχέδια, για πρόκληση backdoor επιθέσεων σε CNN μοντέλα	Το μοντέλο ξεγελιέται με ακρίβεια πάνω από 95% με χρήση πάνω από 5% backdoor εικόνων
Raindrop remonal	White-Box	Digital	[207]: Προσθήκη poisoning ζεύγη εικόνων, για πρόκληση trojan επιθέσεων σε GAN μοντέλα	Το GAN όταν αφαιρεί τις σταγόνες βροχής, ταυτόχρονα μεταμορφώνει το κόκκινο φανάρι σε πράσινο ή αλλάζει τον αριθμό στο σήμα ορίου ταχύτητας

Πίνακας 6.3: *Πίνακας με state-of-the-art adversarial* ***Poisoning*** *επιθέσεις σε συστήματα αυτόνομης οδήγησης.*

- **Reactive Defenses**: Άμυνες που εφαρμόζονται μετά την εκπαίδευση του μοντέλου και στόχο έχουν κυρίως να ανιχνεύσουν επιθέσεις.
- **Proactive Defenses**: Άμυνες που εφαρμόζονται κατά την εκπαίδευση του μοντέλου και στόχο έχουν κυρίως στο να κάνουν το μοντέλο πιο εύρωστο ενάντια σε ανταγωνιστικές επιθέσεις.

Όσον αφορά την αντιμετώπιση των διαφορετικών ειδών επιθέσεων μπορούμε να κατηγοριοποιήσουμε τις τεχνικές αμυνών στα συστήματα αυτόνομης οδήγησης, ανάλογα με τη φάση εφαρμογής τους:

- Άμυνες ενάντια σε **inference-time attacks**: Άμυνες που εφαρμόζονται κατά το τρέξιμο του μοντέλου, για την αντιμετώπιση των **evasion** επιθέσεων.
- Άμυνες ενάντια σε **training-time attacks**: Άμυνες που εφαρμόζονται κατά την εκπαίδευση του μοντέλου, για την αντιμετώπιση των **poisoning** επιθέσεων.

Ανάλογα τώρα με την προσέγγιση αντιμετώπισης, μπορούμε να κατηγοριοποιήσουμε τις άμυνες στα συστήματα αυτόνομης οδήγησης στις εξής κατηγορίες:

- **Data preprocessing**: Άμυνες που τροποποιούν τα δεδομένα εκπαίδευσης ή δοκιμής και των χαρακτηριστικών τους, για την ελαχιστοποίηση ή την αποφυγή των adversarial examples.
- **Model hardening**: Άμυνες που τροποποιούν κυρίως τα χαρακτηριστικά του μοντέλου, εκπαιδεύοντας το με σκοπό την ευρωστία ενάντια σε adversarial examples.
- **Auxiliary models**: Άμυνες που χρησιμοποιούν επιπλέον ML (ή και άλλα είδη) μοντέλα που κάνουν εξειδικευμένες εργασίες, για να ενισχύσουν την ευρωστία του κύριου μοντέλου.

6.1.3.2 Τεχνικές Αμυνών

Στις **Data preprocessing** άμυνες κατατάσσονται αρκετές τεχνικές όπως:

1. *Defense Distillation*: Όπως έχουμε αναφέρει στην ενότητα 3.6.2.3, πρόκειται για ένα μηχανισμό που έχει σχεδιαστεί για να συμπιέζει μεγάλα μοντέλα σε μικρότερα διατηρώντας παράλληλα την ακρίβεια πρόβλεψης. Έτσι το καινούριο μοντέλο είναι λιγότερο ευαίσθητο σε αλλαγές των κλίσεων οπότε και πιο εύρωστο σε adversarial examples. Όμως, μεταγενέστερη έρευνα έδειξε ότι αυτή η μέθοδος είναι ευάλωτη σε νέα επίθεση βασιζόμενη στη μέθοδο βελτιστοποίησης [26].
2. Feature Squeezing: Όπως έχουμε αναφέρει στην ενότητα 3.6.2.4, πρόκειται για μια τεχνική που μειώνει τον διαθέσιμο χώρο των χαρακτηριστικών εισόδου, καθώς είναι συνήθως άσκοπα μεγάλος και παρέχει χώρο για να κατασκευαστούν adversarial perturbations. Όμως, και για αυτήν τη μέθοδο μεταγενέστερη έρευνα έχει δείξει ότι δεν είναι τόσο αποτελεσματική [84].
3. *Feature Denoising*: Πρόκειται για μέθοδο που επίσης προσπαθεί να μειώσει τον αριθμό των χαρακτηριστικών εισόδου, για τη βελτίωση της ευρωστίας ενάντια σε adversarial examples. Αυτό μπορεί να γίνει με χρήση ειδικών denoising μπλοκ, τα οποία λειτουργούν σαν φίλτρα, και οδηγούν σε αύξηση της ευρωστίας ενάντια σε white-box και black-box επιθέσεις [96].

4. *Input transformation*: Όπως έχουμε αναφέρει στην ενότητα 3.6.2.7, πρόκειται για τεχνικές που χρησιμοποιούν τυχαίους μετασχηματισμούς ειδικά για εικόνες όπου μπορεί να γίνουν μετασχηματισμοί όπως περικοπές, αλλαγή ανάλυσης, μείωση του bit-depth, ελαχιστοποίηση της συνολικής διακύμανσης [75], ή και JPEG συμπίεση [86], για την αφαίρεση ή μείωση του οποιουδήποτε adversarial perturbation πριν την τροφοδότησή τους στα μοντέλα. Αυτές μπορούν να εφαρμοστούν είτε proactively στα δεδομένα εκπαίδευσης του μοντέλου, είτε reactively κατά το inference του μοντέλου, σαν ενδιάμεσο σύστημα άμυνας ανάμεσα από τις εισόδους και το μοντέλο.
5. *Adversarial transformation*: Σε αυτήν την κατηγορία τεχνικών συγκαταλέγονται όλες αυτές οι τεχνικές που έχουν σκοπό να 'καθαρίσουν' τις ανταγωνιστικές εισόδους σε καθαρές εισόδους, μέσω μετασχηματισμών. Όπως και στον απλό μετασχηματισμό εισόδων (Input transformation), αυτές οι τεχνικές μπορούν να εφαρμοστούν είτε proactively είτε reactively. Για παράδειγμα, πολλές τεχνικές χρησιμοποιούν GANs για αυτήν τη λειτουργία, όπως τα MagNet [76], APE-GAN [50], DefanseGAN [77] που έχουμε αναλύσει στην ενότητα 3.6.2.9. Αυτά τα μοντέλα συνήθως μαθαίνουν την υποκείμενη κατανομή του συνόλου δεδομένων εικόνων (manifold) και μπορεί να μετασχηματίσουν ανταγωνιστικές εικόνες πίσω σε καθαρές, που εμπίπτουν στη μαθημένη κατανομή. Υπάρχουν όμως και άλλες τεχνικές που χρησιμοποιούν άλλα μοντέλα για τον 'καθαρισμό' των adversarial perturbations, όπως με χρήση auto-encoders [81].

Στις **Model hardening** άμυνες κατατάσσονται αρκετές τεχνικές όπως:

1. *Adversarial Training*: Όπως έχουμε αναφέρει στην ενότητα 3.3, είναι η μέθοδος εκπαίδευσης του νευρωνικού με σύνολο δεδομένων τα αρχικά μαζί με adversarial examples για τη δημιουργία εύρωστων νευρωνικών απέναντι σε adversarial examples, αλλά και πιο γενικευμένων μοντέλων απέναντι σε κανονικά δείγματα. Αυτή η μέθοδος προστατεύει από επιθέσεις δειγμάτων με τα οποία έχει εκπαιδευτεί, αλλά αδυνατεί να προστατεύσει από δείγματα που δεν έχει ξαναδεί [36].
2. *Certified robustness*: Όπως έχουμε αναφέρει στην ενότητα 4.3.2, πρόκειται για τεχνικές που προσφέρουν αποδείξιμη άμυνα ενάντια σε επιθέσεις με adversarial examples μέχρι ένα συγκεκριμένο threshold διαταραχής. Όμως, για πολύπλοκα και μεγάλα σύνολα δεδομένων εικόνων (π.χ. ImageNet), ακόμα και οι state-of-the-art τεχνικές, έχουν πολύ χαμηλά επίπεδα certified robustness accuracy [102].
3. *Network regularization*: Όπως έχουμε αναφέρει στην ενότητα 3.6.2.2, πρόκειται για τεχνικές που προσθέτουν ένα επιπλέον επίπεδο ενός regularizer βασισμένου σε adversarial perturbations, σε υπάρχοντα μοντέλα για την αντίσταση και αποφυγή από adversarial attacks. Αυτές οι τεχνικές προσφέρουν ικανοποιητικά ποσοστά robustness accuracy [80].

Στις άμυνες που κάνουν χρήση **Auxiliary models** ή γενικότερα τρίτων συστημάτων για την ισχυροποίηση του μοντέλου κατατάσσονται αρκετές τεχνικές όπως:

1. *Adversarial Detection*: Όπως έχουμε αναφέρει στην ενότητα 3.6.2.11, πρόκειται για τεχνικές που ένα επιπλέον μοντέλο χρησιμοποιείται αποκλειστικά για την ανίχνευση adversarial examples, με σκοπό τον αποκλεισμό τους. Υπάρχουν πολλά παραδείγματα

τέτοιων ανιχνευτών στη βιβλιογραφία, τα οποία συνήθως αποτελούνται από πιο απλά ML μοντέλα όπως binary ταξινομητές [67], RBF-kernel SVMs (SafetyNet) [94], ή και unsupervised μοντέλα (I-defender) [69], τα οποία συνήθως κατατάσσουν τα adversarial examples ως outliers συγκρίνοντάς την κατανομή των χαρακτηριστικών τους και βρίσκοντας τα out-of-distribution δείγματα [68]. Αυτές οι τεχνικές δεν τροποποιούν το υπάρχον μοντέλο και μπορούν να συνδυαστούν με άλλες πιο proactive άμυνες.

2. *Ensemble Defenses*: Όπως έχουμε αναφέρει στην ενότητα 3.6.2.10, πρόκειται για μια κατηγορία αμυνών στις οποίες παραπάνω από μία τεχνική χρησιμοποιείται για την προστασία των μοντέλων. Πολλαπλές άμυνες (δηλαδή μοντέλα/ταξινομητές) συνδυάζονται (παράλληλη ή διαδοχικά) για την καλύτερη προστασία του συστήματος από πολλά και διαφορετικά είδη επιθέσεων [208]. Ένα χαρακτηριστικό παράδειγμα τέτοιας μεθόδου είναι το PixelDefend [91], στο οποίο ένας ανιχνευτής adversarial δειγμάτων και ένας 'ανακατασκευαστής εισόδου' είναι ενσωματωμένοι για να περιορίσουν τα adversarial examples, το οποίο αποδείχθηκε αρκετά αποτελεσματικό σε μεγάλη ποικιλία επιθέσεων. Όμως, χρειάζεται προσεκτική επιλογή των αμυνών που θα συνδυαστούν, καθώς μεταγενέστερη έρευνα έχει δείξει ότι ο συνδυασμός αδύναμων αμυνών δεν μπορεί να παρέχει ισχυρή άμυνα απέναντι σε adversarial examples [84].
3. *Anomaly Detection*: Πρόκειται για τεχνικές ανίχνευσης ανωμαλιών που βασίζονται σε τριτογενείς παράγοντες, όπως την καθυστέρησης πρόβλεψης του μοντέλου ή της χρήσης των πόρων του συστήματος (CPU, GPU, Memory). Η επεξεργασία των adversarial examples από ένα ML σύστημα είναι μια χρονοβόρα διαδικασία που μπορεί να προκαλέσει απότομη αύξηση χρήσης των πόρων του συστήματος, αλλά κάποιες τεχνικές επιθέσεων όπως οι Universal perturbation, μπορούν να ξεφύγουν ανίχνευσης καθώς προκαλούν ελάχιστη επιπλέον αύξηση πόρων [7].

Πέρα από τις παραπάνω γενικές τεχνικές άμυνας που μπορούν να εφαρμοστούν στα περισσότερα νευρωνικά δίκτυα υπάρχουν και τεχνικές ενίσχυσης της ευρωστίας των μοντέλων αυτόνομης οδήγησης, για συγκεκριμένες επιθέσεις που στοχεύουν στη δυσλειτουργία κάποιου υποσυστήματος του αυτόνομου οχήματος. Πιο συγκεκριμένα για τις επιθέσεις στο υποσύστημα της πρόβλεψης τροχιάς (**Trajectory prediction**), υπάρχουν ειδικές τεχνικές που μπορούν να χρησιμοποιηθούν για να ενισχυθεί η ευρωστία του μοντέλου, όπως:

- Ενίσχυση δεδομένων (*Data augmentation*): Προσθήκη δεδομένων με τυχαίες διαταραχές (υπό περιορισμούς) σε επιλεγμένες τροχιές για την ενίσχυση των δεδομένων εκπαίδευσης που ως επί το πλείστον έχουν ομαλές τροχιές με σταθερή επιτάχυνση [181]. Πρόκειται για τεχνική παρόμοια με το input transformation, αλλά εδώ δεν εισάγεται τυχαίος θόρυβος ή μετασχηματισμοί, αλλά συγκεκριμένες διαταραχές οι οποίες καταλήγουν να έχουν διαφορετική κατανομή δεδομένων, προσομοιάζοντας περισσότερο την τεχνική του adversarial training.
- Εξομάλυνση τροχιάς κατά την εκπαίδευση (*Train-time trajectory smoothing*): Δεδομένου ότι η ασταθής ταχύτητα ή επιτάχυνση είναι ένα βασικό μοτίβο των ανταγωνιστικών τροχιών, αυτό το ανταγωνιστικό αποτέλεσμα μπορεί να αφαιρεθεί με χρήση πολλαπλών διαθέσιμων αλγορίθμων εξομάλυνση τροχιάς στα δεδομένα εκπαίδευσης και δοκιμής [181].

- Ανίχνευση και εξομάλυνση τροχιάς κατά τη δοκιμή (*Test-time detection and trajectory smoothing*): Για να αποφευχθεί η επανεκπαίδευση του μοντέλου, μπορούν να χρησιμοποιηθούν τεχνικές ανίχνευσης adversarial τροχιών (π.χ. με χρήση SVM ταξινομητών) και εφαρμογή εξομάλυνσης μόνο τότε, δηλαδή κατά το inference του μοντέλου [181]. Πρόκειται επίσης για τεχνική παρόμοια με adversarial detection, η οποία όμως ειδικεύεται στο να ανιχνεύει ανταγωνιστικές τροχιές με βάση το μέγεθος και την κατεύθυνση της επιτάχυνσης.

Στους πίνακες 6.4, 6.5 φαίνονται συγκεντρωτικά η ταξινόμηση των αμυνών ενάντια σε ανταγωνιστικές επιθέσεις σε συστήματα αυτόνομης οδήγησης, χωρισμένες σε **Proactive** και **Reactive** τεχνικές και ενδεικτικά παραδείγματα για κάθε είδος άμυνας μαζί με συνοπτική ανάλυση και σχόλια για την κάθε κατηγορία.

Προσέγγιση	**Άμυνα**	**Περιγραφή/Παραδείγματα**	**Ανάλυση/Σχόλια**
Model Hardening	Adversarial Training	Εκπαίδευση νέου εύρωστου μοντέλου με νέο σύνολο δεδομένων που περιλαμβάνει adversarial examples (π.χ. [36], [1], [208])	Αύξηση χρόνου και πόρων για εκπαίδευση μοντέλου [55] και αδυναμία προστασίας από καινούρια adversarial examples [36]
Model Hardening	Certified Robustness	Εκπαίδευση νέου αποδεδειγμένα εύρωστου μοντέλου ενάντια σε ανταγωνιστικές επιθέσεις με συγκεκριμένο threshold διαταραχής (π.χ. [102], [103], [95], [106])	Αύξηση χρόνου και πόρων για εκπαίδευση μοντέλου [55], χαμηλά επίπεδα certified robustness accuracy για μεγάλα dataset [102]
Model Hardening	Network Regularization	Εκπαίδευση νέου εύρωστου μοντέλου με προσθήκη επιπλέον επίπεδου ενός regularizer βασισμένου σε adversarial perturbations (π.χ. [80], [81], [82])	Αύξηση χρόνου και πόρων για εκπαίδευση μοντέλου και συνήθως αποτελεσματική μόνο για απλές επιθέσεις [55]
Data Preprocessing	Defense Distillation	Εκπαίδευση νέου εύρωστου μοντέλου με τη μέθοδο του distillation, όπου γίνεται απόσταξη των κρυμμένων πληροφοριών των στρωμάτων από το αρχικό μοντέλο [51]	Όχι τόσο αποτελεσματική όσο αρχικά, καθώς έχουν αναπτυχθεί adaptive επιθέσεις ενάντίων της μεθόδου [26]
Data Preprocessing	Feature Squeezing	Μείωση του διαθέσιμου χώρου των χαρακτηριστικών εισόδου, καθώς είναι συνήθως άσκοπα μεγάλος και παρέχει χώρο για να κατασκευαστούν adversarial perturbations [70]	Όχι τόσο αποτελεσματική όσο αρχικά, καθώς έχουν αναπτυχθεί adaptive επιθέσεις ενάντίων της μεθόδου [84]
Data Preprocessing	Feature Denoising	Μείωση του αριθμού των χαρακτηριστικών εισόδου για μείωση του επιπλέον θορύβου και βελτίωση ευρωστίας ενάντια σε adversarial examples (π.χ. [96], [209])	Ανίχνευση white-box επιθέσεων με μέγιστη ακρίβεια 55% και black-box επιθέσεων με μέγιστη ακρίβεια 49.5% [96]
Data Preprocessing	Trajectory Smoothing	Εφαρμογή Data augmentation και Trajectory smoothing [181]	Μείωση σφάλματος πρόβλεψης στις επιθέσεις κατά 26%. Αύξηση σφάλματος πρόβλεψης στην κανονική λειτουργία κατά 11%

Πίνακας 6.4: *Πίνακας με ταξινόμηση* ***Proactive*** *αμυνών ενάντια σε ανταγωνιστικές επιθέσεις σε συστήματα αυτόνομης οδήγησης και ενδεικτικά παραδείγματα.*

6.1.3.3 Προστασία ιδιωτικότητας

Σε αυτήν την ενότητα, γίνεται μια αναφορά σε τεχνικές που στόχο έχουν στην προστασία ενάντια σε επιθέσεις κατά τις ιδιωτικότητας (privacy), καθώς πέρα από τη δημιουργία εύρωστων ML μοντέλων, η προστασία της ιδιωτικότητας αποτελεί ένα αρκετά σημαντικό κομμάτι στη γενικότερη ασφάλεια των συστημάτων. Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων (τα οποία μπορεί να έχουν προκύψει από δεδομένα που μαζεύουν άλλα οχήματα τα οποία είναι ήδη σε λειτουργία) όπως και αυτά τα οποία ανταλλάσσονται μεταξύ των

Προσέγγιση	**Άμυνα**	**Περιγραφή/Παραδείγματα**	**Ανάλυση/Σχόλια**
Data Preprocessing	Image Transformation	Μετασχηματισμοί κυρίως εικόνων (περικοπές, συμπίεση, αλλαγή ανάλυσης, κ.α.), για υπεράσπιση από ανταγωνιστικές επιθέσεις (π.χ. [75], [86])	Για εκπαίδευση με χρήση μετασχηματισμένων εικόνων, 70% μέση επιτυχία προστασίας ενάντια σε adversarial examples [75]
Data Preprocessing	Adversarial Transformation	Μετασχηματισμός εισόδων για μετατροπή adversarial examples σε καθαρές εισόδους (π.χ. MagNet [76], APE-GAN [50], DefanseGAN [77], PixelDefend [91], [81])	Πιθανότητα μείωσης της απόδοσης των μοντέλων υπό κανονικές συνθήκες [55]
Auxiliary Model	Adversarial Detection	Ανίχνευση adversarial examples με χρήση ανιχνευτή ή με επαλήθευση της αναπαράστασης των χαρακτηριστικών των εισόδων (π.χ. SafetyNet [94], I-defender) [69], [67], [68]	Πολλές φορές μπορεί να απαιτεί επιπλέον και μη διαθέσιμους υπολογιστικούς πόρους [55]
Auxiliary Model	Ensembling Defenses	Χρήση πολλαπλών αμυνών/μοντέλων/ταξινομητών (π.χ. PixelDefend [91])	Συνδυασμός αδύναμων αμυνών δεν μπορεί να παρέχει ισχυρή άμυνα απέναντι σε adversarial examples [84]
Auxiliary Model	Anomaly Detection	Παρακολούθηση του χρόνου εκτέλεσης δειγμάτων σε αυτόνομα οχήματα για ανίχνευση ανώμαλης αύξησης χρήση πόρων του συστήματος [7]	Σε 2/5 επιθέσεις εμφανίζεται επιπλέον χρήση CPU Memory κατά 50% και GPU Memory 35%, ενώ σε 2/5 επιθέσεις σχεδόν καμία επιπλέον χρήση (κάτω από 1%)
Auxiliary Model	Trajectory Smoothing	Εφαρμογή Adversarial trajectory detection και Trajectory smoothing [181]	Μείωση σφάλματος πρόβλεψης στις επιθέσεις κατά 12%. Αύξηση σφάλματος πρόβλεψης στην κανονική λειτουργία κατά 6%

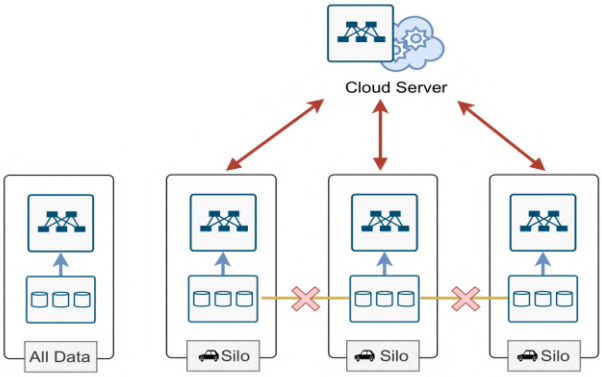
Πίνακας 6.5: *Πίνακας με ταξινόμηση **Reactive** αμυνών ενάντια σε ανταγωνιστικές επιθέσεις σε συστήματα αυτόνομης οδήγησης και ενδεικτικά παραδείγματα.*

συστημάτων των οχημάτων αλλά και απομακρυσμένα (π.χ. στο cloud) πρέπει να είναι ασφαλή και τα ML συστήματα να μην είναι ευαίσθητα σε ανταγωνιστικές επιθέσεις που στόχο έχουν στην παραβίαση της ιδιωτικότητας [180].

Για την αποφυγή λοιπόν επιθέσεων που στόχο έχουν στο να αποκτήσουν πρόσβαση στα δεδομένα, μπορούν να αξιοποιηθούν τεχνικές **Federated learning** για την εκπαίδευση των μοντέλων αυτόνομης οδήγησης. Ένα όχημα με σύστημα αυτόνομης οδήγησης το οποίο βρίσκεται ήδη στο δρόμο, επιτελεί και έναν επιπλέον σκοπό, τη συλλογή καινούριων δεδομένων από τον δρόμο και τη συμπεριφορά των οδηγών, για να τα αξιοποιήσει μελλοντικά για τη βελτίωση της εκπαίδευσης των καινούριων μοντέλων [210]. Για να γίνει όμως κεντρικά αυτή η εκπαίδευση, χρειάζεται να γίνει συλλογή όλων των δεδομένων από τον στόλο οχημάτων που είναι σε λειτουργία, κάτι το οποίο μπορεί να φέρει σε κίνδυνο την ιδιωτικότητα των οδηγών, καθώς το υλικό που μεταφέρεται μπορεί να περιέχει προσωπικά δεδομένα ή και πρόσωπα [174]. Ο κίνδυνος ενέχει σε δύο σημεία κυρίως, αρχικά τον κίνδυνο υποκλοπής αυτών σε οποιοδήποτε στάδιο της μεταφοράς και αποθήκευσης, καθώς και στο ότι τα δεδομένα αυτά θα κατέχονται κεντρικά από την οποιαδήποτε εταιρία. Το Federated learning όπως αναφέραμε και στην ενότητα 5.3.6, έχει τη δυνατότητα να λύσει αυτά τα θέματα ιδιωτικότητας, καθώς τα δεδομένα χρησιμοποιούνται για την εκπαίδευση τοπικών μοντέλων σε κάθε όχημα, και δε μεταφέρονται πουθενά, παρά μόνο οι αλλαγές στις παραμέτρους των τοπικά εκπαιδευμένων μοντέλων μεταφέρονται σε κάποιο κεντρικό μοντέλο (βλ. σχήμα 6.11).

Πέρα από την προστασία των δεδομένων που προσφέρει αυτή η μέθοδος όμως, αποτελεί και μια μέθοδος κατανεμημένης εκπαίδευσης, που προσφέρει μοντέλα με καλύτερες επιδόσεις

και πιο αποδοτική εκπαίδευση των μοντέλων αυτόνομης οδήγησης με περισσότερα δεδομένα, καθώς είναι πολύ πιο εύκολο και αποδοτικό υπολογιστικά και ενεργειακά να εκπαιδευτούν πολλά μοντέλα με διαφορετικά δεδομένα τα οποία θα μοιραστούν τις βελτιώσεις στις παραμέτρους τους, παρά ένα μεγάλο κεντρικό μοντέλο το οποίο θα πρέπει να εκπαιδευτεί με ένα τεράστιο όγκο δεδομένων [157].



Σχήμα 6.11: *Σχηματικό παράδειγμα με σύγκριση μεθόδων εκπαίδευσης για συστήματα αυτόνομης οδήγησης. (a) Κεντρική τοπική εκπαίδευση (b) Federated learning με κεντρικό μοντέλο [211].*

6.1.4 Αξιολόγηση Αμυνών και Ευρωστίας μοντέλων

6.1.4.1 Εκπαίδευση και Λειτουργία

Η εκπαίδευση μοντέλων αυτόνομης οδήγησης απαιτεί γενικά μεγάλα σύνολα δεδομένων και απαιτεί σημαντικό χρόνο εκπαίδευσης. Το adversarial training προσθέτει επιπλέον υπολογιστικό και χρονικό κόστος [55], οπότε θα πρέπει να αξιοποιείται εφόσον υπάρχει ο κίνδυνος συγκεκριμένων επιθέσεων και όχι για να προσφέρει γενική προστασία. Άλλες τεχνικές όπως image και adversarial transformation, μπορούν να βελτιώσουν την ευρωστία, των μοντέλων χωρίς μεγάλο υπολογιστικό κόστος και επιπλέον χρόνο εκπαίδευσης, καθώς μετασχηματίζουν τις εισόδους είτε κατά την εκπαίδευση είτε κατά το inference [55].

6.1.4.2 Επίπτωση στην ακρίβεια

Μια επίπτωση των περισσότερων μεθόδων άμυνας, είναι η πτώση της ακρίβειας του μοντέλου σε κανονικές εισόδους, κάτι το οποίο ανάλογα με την πτώση μπορεί να μην είναι αποδεκτό σε κρίσιμα συστήματα όπως τα ADS [55]. Για παράδειγμα, οι απλοί μετασχηματισμοί στις εικόνες εισόδου (image transformations) όπως περικοπή και περιστροφή [75], ενώ μειώνουν το ASR, μπορεί να προκαλέσουν μεγάλο σφάλμα πρόβλεψης σε regression μοντέλα, όπως τα μοντέλα οδήγησης. Επίσης, οι περισσότερες τεχνικές αφορούν μοντέλα που ειδικεύονται σε εργασίας classification, και μπορεί να μην έχουν τα ίδια ποσοστά επιτυχίας και ακρίβειας σε εργασίες regression που εκτελούν πολλά από τα μοντέλα αυτόνομης οδήγησης [7].

6.1.4.3 Χρόνος απόκρισης

Η κάθε άμυνα που προσθέτει κάποιο επιπλέον στάδιο στο μοντέλο, η ακόμα και η χρήση κάποιου επιπλέον μοντέλου για ανίχνευση επιθέσεων, συνήθως προκαλεί και αύξηση του χρόνου απόκρισης του μοντέλου σε κάθε είσοδο. Αυτό θα πρέπει να ελέγχεται καθώς, ο χρόνος για την απόφαση σε ένα μοντέλο αυτόνομης οδήγησης είναι κρίσιμος, και μια μεγάλη καθυστέρηση σε τέτοια συστήματα πραγματικού χρόνου μπορεί να προκαλέσει καταστροφικές συνέπειες [55].

Όμως γενικότερα, επειδή τα ADS είναι συστήματα πραγματικού χρόνου, τεχνικές και άμυνες που βασίζονται στην παρακολούθηση και ανίχνευση σε πραγματικό χρόνο, όπως adversarial detection και anomaly detection (resource monitoring), έχουν μεγάλη αξία εφόσον δεν καταναλώνουν πολλούς πόρους ή δεν προσθέτουν μεγάλη καθυστέρηση στη λήψη αποφάσεων [55].

6.1.4.4 Μεταφορά επιθέσεων

Από τις δοκιμές επιθέσεων σε διάφορα μοντέλα, φαίνεται ότι η δυνατότητα μεταφοράς (transferability) των επιθέσεων σε άλλα μοντέλα αυτόνομης οδήγησης δεν έχει μεγάλο attack success rate (ASR). Γενικά οι περισσότερες επιθέσεις έχουν μεγαλύτερη επιτυχία σε White-box συνθήκες, και πολύ λιγότερο σε Black-box συνθήκες. Πιο συγκεκριμένα στο [7], σε δοκιμή 5 διαφορετικών επιθέσεων (π.χ. IT-FGSM, Optimization, AdvGAN) το ASR σε Black-box συνθήκη ήταν 4% κατά μέσο όρο σε όλες τις επιθέσεις. Οπότε μεγάλο κομμάτι της ασφάλειας από ανταγωνιστικές επιθέσεις είναι και η απόκρυψη πληροφοριών για το μοντέλο και τους παραμέτρους τους, αλλά και η χρήση τεχνικών για το απόκρυψη (obfuscation) αυτών ή για την προστασία από model extraction επιθέσεις που στόχο έχουν να εξάγουν λεπτομέρειες για το μοντέλο, όπως το PRADA [135] για την ανίχνευση τέτοιων επιθέσεων.

6.1.4.5 Συνδυασμός αμυνών

Επίσης, υπάρχει η ανάγκη για χρήση πολλαπλών αμυντικών τεχνικών, καθώς καμία άμυνα από μόνη της δεν μπορεί να προσφέρει προστασία από όλους τους τύπους επιθέσεων. Συνήθως οι τεχνικές που δουλεύουν πολύ καλά για μόνο ένα είδος επιθέσεων, όπως adversarial training και defensive distillation οι οποίες είναι αποτελεσματικές μόνο στις πιο απλές επιθέσεις π.χ. FGSM, ή όταν έχουν δοθεί τα κατάλληλα δεδομένα εκπαίδευσης, δεν μπορούν να βοηθήσουν για τη γενικότερη ευρωστία του μοντέλου απέναντι σε ένα σύνολο επιθέσεων [181]. Αντίθετα, οι τεχνικές που μπορεί να παρέχουν αποτελεσματική προστασία σε μια γκάμα επιθέσεων, είτε θα έχουν μέτρια ακρίβεια ή θα ενισχύσουν τον ρυθμό των false positives, όπως συμβαίνει στο feature squeezing [7].

6.1.4.6 Δείκτες αξιολόγησης ευρωστίας

Όπως έχουμε αναφερθεί και στην ενότητα 4.4 υπάρχουν κάποια γενικά μετρήσεις οι οποίες μπορούν να μας δώσουν μια εικόνα για την ευρωστία ενός μοντέλου. Για τα μοντέλα αυτόνομης οδήγησης, μετρήσεις οι οποίες μπορούν να χρησιμοποιηθούν είναι: (i) **Empirical robustness** για να αξιολογηθεί η μικρότερη δυνατή διαταραχή το οποίο είναι ικανό να οδηγήσει σε σφάλμα το μοντέλο, και (ii) **Local loss sensitivity** για να αξιολογηθεί το smoothness του μοντέλου.

Πιο συγκεκριμένα όμως για τα μοντέλα αυτόνομης οδήγησης μπορούμε να χρησιμοποι-

ήσουμε και κάποιες άλλες μεθόδους για να αξιολογήσουμε τις ανταγωνιστικές άμυνες. Για παράδειγμα, (i) *ο χρόνος που απαιτείται από το σύστημα για τον εντοπισμό ενός ανταγωνιστικού δείγματος*, ο οποίος θέλουμε να είναι όσο το δυνατόν μικρότερος, καθώς και (ii) *το ποσοστό ανίχνευσης adversarial examples*, για να δειχθεί το πόσο αποτελεσματικά ένα σύστημα μπορεί να αναγνωρίσει ανταγωνιστικές εισόδους.

Πέρα από αυτές τις μετρήσεις οι οποίες είναι χρήσιμες για την αξιολόγηση των προτεινομένων αμυνών, είναι χρήσιμο να αξιολογηθούν και με τις μεθόδους που έχουμε περιγράψει στην ενότητα 4.5, καθώς οι μετρήσεις από μόνες τους δεν μπορούν να δώσουν ένα συνολικό συμπέρασμα για την ευρωστία του μοντέλου.

Ωστόσο, στα μοντέλα αυτόνομης οδήγησης είναι σημαντικό να αξιολογηθούν για την ευρωστία τους σε ένα γενικότερο πλαίσιο όσο αναφορά την ασφάλεια οδήγησης, και όχι μόνο για την ακρίβεια του μοντέλου, καθώς ο τελικός στόχος είναι η ασφάλεια του οχήματος και των οδηγών και όχι η καλύτερη δυνατή ακρίβεια πρόβλεψης των μοντέλων. Όπως αναφέραμε και στην αρχή της ενότητας 6.1.2, δεν υπάρχει ακόμα σύνδεση μεταξύ των δύο, άρα είναι πιθανό σε κάποιες περιπτώσεις ενώ αυξάνεται η ευρωστία και άρα η ακρίβειας ενός μοντέλου, να υπονομεύεται κάποιος άλλος τομέας με σκοπό να μειώνεται η συνολική ασφάλεια [187].

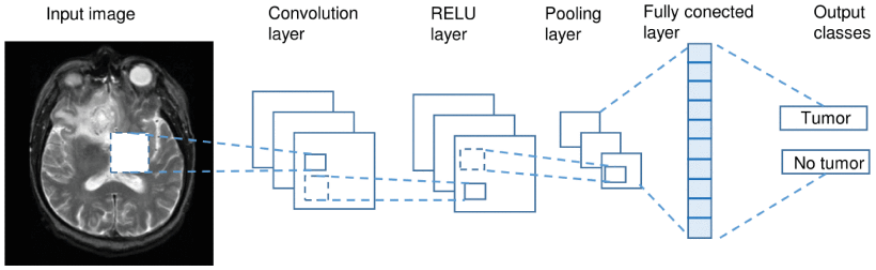
6.2 Φροντίδα Υγείας και Διάγνωση

Ο κλάδος που σχετίζεται με τη φροντίδα της υγείας περιλαμβάνει ένα μεγάλο εύρος αντικειμένων, ειδικοτήτων και εφαρμογών. Από ιατρούς, ασθενείς, νοσοκομεία, εξετάσεις, διαγνώσεις μέχρι και ιατρικές συσκευές. Όπως έχει συμβεί και με άλλους τομείς τα τελευταία χρόνια, η τεχνητή νοημοσύνη έχει αρχίσει να χρησιμοποιείται και για τέτοιες ιατρικές εφαρμογές, ώστε να διευκολύνει όλα τα άτομα που συμπεριλαμβάνονται σε αυτό το πεδίο, όπως τον ασθενή, ιατρό, νοσηλευτή κ.λπ., αλλά και να βελτιώνει τις υπηρεσίες υγείας. Κάποιες από τις υπηρεσίες στις οποίες έχει ήδη βρει εφαρμογή το AI και συγκεκριμένα η μηχανική μάθηση και τα DNNs είναι π.χ. η αναγνώριση οργάνων από ιατρικές εικόνες, η ταξινόμηση ασθενειών και η ανίχνευση όγκων (βλέπε παράδειγμα στο σχήμα 6.12) [6].

Παρά τις επιδόσεις του AI, η χρήση του στην ιατρική ενέχει κινδύνους, κυρίως λόγο ζητημάτων ασφαλείας και ιδιωτικότητας. Τα θέματα ασφαλείας προκύπτουν κυρίως από τη χρήση του σε ιατρικές διαγνώσεις και αποφάσεις, οι οποίες μπορεί να μπορεί να θέσουν σε κίνδυνο την υγεία ακόμα και τη ζωή των ασθενών [212]. Τα θέματα ιδιωτικότητας προκύπτουν από τη χρήση ιατρικών δεδομένων τα οποία είναι ευαίσθητα δεδομένα, άρα είναι και δύσκολο να αποκτηθούν αλλά ακόμα και μετά την απόκτηση τους, η χρήση τους με απλές τεχνικές ανωνυμοποίησης δεν είναι αρκετή για να διατηρήσει το απόρρητο των ασθενών [171].

6.2.1 Εφαρμογές Τεχνητής Νοημοσύνης στο πεδίο

Όπως ήδη αναφερθήκαμε η χρήση AI/ML στο πεδίο της υγείας έχει ήδη βρει εφαρμογή, καθώς έχει βοηθήσει η ψηφιοποίηση των δεδομένων υγείας, η ετερογένεια των δεδομένων και η μεγαλύτερη διαθεσιμότητα τους, ώστε να μπορέσουν να αναπτυχθούν αξιόπιστα μοντέλα. Ένα παράδειγμα από την πραγματική ζωή είναι και αυτό της Digital Diganostics που το προϊόν της ήταν η πρώτη ιατρική συσκευή που κάνει χρήση AI, που ενέκρινε ο αμερικανικός οργανισμός



Σχήμα 6.12: *Παράδειγμα χρήσης ενός νευρωνικού (CNN) για την ταξινόμηση ασθενειών με βάση μιας εικόνας εισόδου που περιέχει την αξονικής τομή ενός εγκεφάλου μέσω MRI. Στο διάγραμμα φαίνονται ενδεικτικά τα στρώματα που περνάει η είσοδος μέσα από το νευρωνικό για καταλήξει στην τελική ταξινόμηση [213].*

τροφίμων και φαρμάκων (FDA) να βγει στην αγορά[13]. Συγκεκριμένα πρόκειται για το προϊόν της με όνομα IDx-DR[14], το οποίο είναι μια αυτόνομη διαγνωστική συσκευή που κάνει χρήση AI αλγορίθμων για τη διάγνωση διαβητικής αμφιβληστροειδοπάθειας (diabetic retinopathy), μιας αρκετά κοινής αιτίας απώλειας όρασης για διαβητικούς, χρησιμοποιώντας εικόνες από τον αμφιβληστροειδή του ματιού της οποίες τραβάει και μπορεί να κάνει να κάνει τη διάγνωση χωρίς τη χρήση κάποιου ειδικού ιατρού, κάνοντας πιο εύκολη και προσβάσιμη την έγκαιρη διάγνωση για τέτοιες παθήσεις.

Γενικότερα όμως, τις εφαρμογές της τεχνητής νοημοσύνης στο πεδίο της υγείας, μπορούμε να τις ταξινομήσουμε στις εξής κατηγορίες: (i) **Πρόγνωση**, (ii) **Διάγνωση**, και (iii) **Θεραπεία**.

(i) Πρόγνωση είναι η διαδικασία πρόβλεψης της αναμενόμενης εξέλιξης μιας νόσου, και σε αυτήν την κατηγορία ανήκουν η αναγνώριση ασθενειών, η αναγνώριση όγκων και ταξινόμηση καρκινώματος [6].

(ii) Διάγνωση είναι η αναγνώριση της αιτίας και της ασθένειας με βάση των συμπτωμάτων και ενδείξεων, και αυτήν την κατηγορία μπορούμε να τη διαχωρίσουμε περαιτέρω στη (α) χρήση ML σε **ηλεκτρονικό ιατρικό ιστορικό ασθενών**, και σε (β) χρήση ML για **ανάλυση ιατρικών εικόνων**. Η ύπαρξη πολλών δεδομένων υπό τη μορφή ιατρικού ιστορικού μπορεί να είναι πολύ χρήσιμη για τη διάγνωση των ασθενών για διάφορες ασθένειες όπως ζαχαρώδη διαβήτη. Οι ιατρικές εικόνες μπορεί να είναι διαφόρων ειδών, όπως αξονικές και μαγνητικές τομογραφίες, ακτινογραφίες, υπέρηχοι κ.λπ. Η χρήση ML για την ανάλυση αυτών είναι πολύ σημαντική, καθώς μπορεί να βοηθήσει στη βελτιστοποίηση των εικόνων με προεπεξεργασία τους για αφαίρεση θορύβου και ενίσχυσης της ανάλυσης τους, στην αναγνώριση και ανίχνευση ανωμαλιών (π.χ. όγκος, καρκίνωμα), και στην ταξινόμηση ασθενειών [6].

(iii) Θεραπεία είναι το βήμα μετά τη διάγνωση για την αντιμετώπιση της ασθένειας, και αυτήν την κατηγορία μπορούμε να τη διαχωρίσουμε περαιτέρω στη (α) χρήση ML για την **ερμηνεία των ιατρικών εικόνων**, και (β) στη χρήση ML για την **παρακολούθηση ασθενών σε πραγματικό χρόνο**. Οι εξετάσεις που γίνονται συνήθως συνοδεύονται και από μια αναλυτική ερμηνεία που γίνεται από κάποιον ιατρό ή ραδιολόγο. Η τεχνητή νοημοσύνη μπορεί να βοηθήσει σε αυτές τις διεργασίες για εξοικονόμηση χρόνου και βελτίωση

[13] https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye
[14] https://www.digitaldiagnostics.com/products/eye-disease/idx-dr-eu/

των υπηρεσιών. Για την παρακολούθηση των ασθενών, υπάρχουν φορητές IoT συσκευές οι οποίες συγκεντρώνουν βασικά δεδομένα υγείας, τα οποία μπορούν να σταλθούν σε κάποιο απομακρυσμένο υπολογιστικό σύστημα ή cloud για να γίνει ερμηνεία τους με χρήση ML, όπως για παράδειγμα η παρακολούθηση των καρδιακών παλμών [6].

Μερικές ακόμα εφαρμογές οι οποίες δεν έχουν άμεση συσχέτιση με τους ασθενείς, αλλά ανήκουν στο πεδίο της υγείας είναι και η χρήση ML για οργάνωση των ιατρικών αρχείων καθώς και για τη βελτιστοποίηση της διαχείρισης των νοσοκομείων. Ακόμα χρήση ML μπορεί να γίνει για την ανακάλυψη και ανάπτυξη νέων φαρμάκων [214].

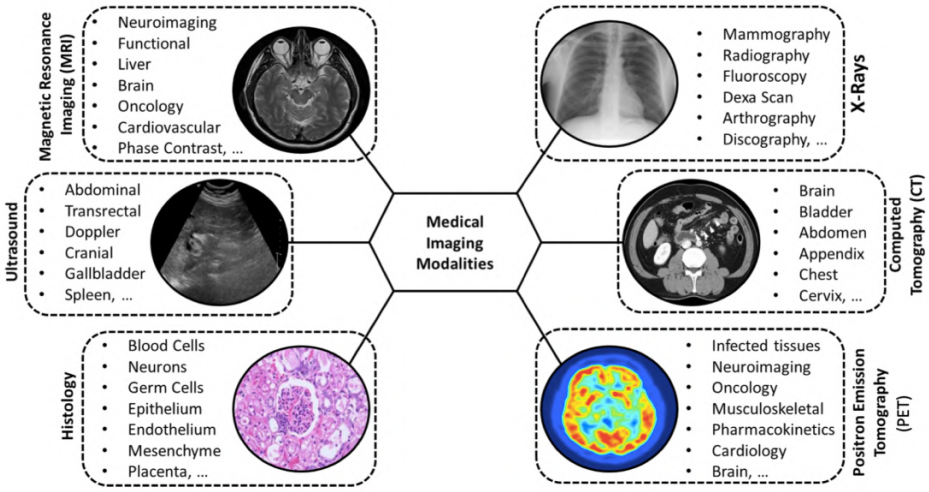
6.2.1.1 Ιατρικές Εικόνες

Μια από τις σημαντικότερες χρήσης του AI και πιο συγκεκριμένα των ML και DL τεχνικών είναι στο πεδίο των ιατρικών εικόνων (medical imaging), όπου χρησιμοποιούνται DNNs από εφαρμογές όραση υπολογιστών για την ανάλυση ιατρικών εικόνων, οι οποίες είναι αρκετά πολύπλοκες, και με υψηλή διάσταση δεδομένων. Με τη χρήση των ML μοντέλου, μπορεί να βελτιωθεί η διάγνωση, η θεραπεία και η παρακολούθηση της υγείας, καθώς γίνεται πιο εύκολη, ακριβής και γρήγορη, οδηγώντας σε υψηλότερα ποσοστά επιτυχίας και μειωμένα ποσοστά θνησιμότητας στην ιατρική [215]. Τα ML μοντέλα μπορεί να ενσωματωθούν σε υπάρχον υπολογιστικά διαγνωστικά συστήματα (Computer-aided detection - CADe) ή υπολογιστικά ανιχνευτικά συστήματα (Computer-aided diagnosis - CADx) τα οποία χρησιμοποιούνται σε κλινικές και νοσοκομεία για να αυτοματοποιήσουν τους ελέγχους των ιατρικών εικόνων [6].

Στην ανάλυση ιατρικών εικόνων, οι ML τεχνικές χρησιμοποιούνται για την αποτελεσματική και αποτελεσματική εξαγωγή πληροφοριών από ιατρικές εικόνες που λαμβάνονται με χρήση διαφορετικών μεθόδων όπως η μαγνητική τομογραφία (MRI), η αξονική τομογραφία (CT), ο υπέρηχος (ultrasound) και η ποζιτρονική τομογραφία (PET) (βλ. σχήμα 6.13 για τις πιο κοινές ιατρικές απεικονίσεις) [6]. Μερικά παραδείγματα εφαρμογής ML σε ιατρικές εικόνες είναι: (i) διάγνωση διαβητικής αμφιβληστροειδοπάθειας ή αλλιώς diabetic retinopathy (DR), η οποία μπορεί να οδηγήσει σε τύφλωση, μέσω εικόνων της ίριδας και του αμφιβληστροειδής χιτώνα (iris/retina scans) (ii) ταξινόμηση κακώσεων του δέρματος (scin lesion classification) μέσω δερματικών ή υπέρυθρων εικόνων, (iii) ανίχνευση καρκίνου του μαστού μέσω εικόνων μαστογραφίας (mammography), (iv) ανίχνευση όγκων εγκεφάλου μέσω MRI εικόνων, (v) διάγνωση καρδιακών νόσων μέσω x-ray στήθους [216], [217].

Οι πιο σημαντικές εργασίες των ML μοντέλων πάνω σε ιατρικές εικόνες, μπορούν να διακριθούν σε: (i) ταξινόμηση ή διάγνωση (**classification/diagnosis**) όπου οι εικόνες δίνονται σαν είσοδοι και το αποτέλεσμα αφορά το αν ο ασθενής έχει μια συγκεκριμένη ασθένεια ή όχι, (ii) ανίχνευση (**detection**) όπου με βάση τις εικόνες εισόδου γίνεται ανίχνευση διαφόρων ασθενειών και κυρίως όγκων, και (iii) κατάτμηση (**segmentation**) όπου γίνεται εξαγωγή συγκεκριμένων κομματιών μιας ιατρικής εικόνας όπως κυττάρων, όγκων ή οργάνων για περαιτέρω ανάλυση [215].

Για την επιτέλεση αυτών των ιατρικών εργασιών τα ML μοντέλα τα οποία χρησιμοποιούνται στην πράξη για ιατρικές εφαρμογές, ποικίλουν ανάλογα με την εφαρμογή τους, αλλά μερικά παραδείγματα είναι: (i) YOLO: πρόκειται για ένα υπερσύγχρονο σύστημα, το οποίο χρησιμοποιείται για την ανίχνευση αντικειμένων σε πραγματικό χρόνο [218], (ii) GANs: τα παραγωγικά ανταγωνιστικά δίκτυα [219] μπορούν να χρησιμοποιηθούν για τη δημιουργία πολ-



Σχήμα 6.13: *Σχηματική ταξινόμηση των πιο κοινών χρησιμοποιούμενων μεθόδων απεικόνισης ιατρικών εικόνων [6].*

λών συνθετικών αλλά ρεαλιστικών δεδομένων [220], (iii) Transformers/RNNs: τα αναδρομικά νευρωνικά δίκτυα και ειδικά η αρχιτεκτονική των transformers [20] χρησιμοποιείται για την ανάλυση και κατανόηση κλινικών και ιατρικών κειμένων [220].

6.2.2 Επιθέσεις - Κενά Ασφαλείας

6.2.2.1 Εφαρμογές Επιθέσεων

Παρά τις εντυπωσιακές επιδόσεις των ML αλγορίθμων σε ιατρικές εφαρμογές, εξακολουθούν να υπάρχουν κίνδυνοι χρήσης τους λόγω ανησυχιών για την ασφάλεια και το απόρρητο των δεδομένων, αλλά και γιατί τα ML μοντέλα είναι ευάλωτα σε ανταγωνιστικές επιθέσεις, ειδικότερα σε εφαρμογές ιατρικών εικόνων, όπου χρησιμοποιούνται deep learning τεχνικές κατά κόρον. Επίσης, η έλλειψη, μεγάλων ιατρικών δεδομένων και ανοιχτών ιατρικών βάσεων δεδομένων, λόγω ζητημάτων απορρήτου, ασφάλειας και κόστους, ήδη περιορίζει τη μέγιστη ακρίβεια που μπορεί να επιτευχθεί από αυτά τα μοντέλα, και επειδή συνήθως η κανονική κλάση συνήθως υπερεκπροσωπείται κάνει τα μοντέλα να συγκλίνουν πιο αργά και να εμφανίζουν υπερπροσαρμογή κάνοντας πιο ευάλωτα σε ανταγωνιστικές επιθέσεις [215].

Οι περισσότερες έρευνες αφοράν την εφαρμογή adversarial examples σε ML μοντέλα που δουλεύουν πάνω σε ιατρικές εικόνες (**medical imaging**). Τα μοντέλα αυτά μπορεί να είναι πιο ευάλωτα σε adversarial examples από ότι στις φυσικές εικόνες, γιατί (i) σε σχέση με άλλους τομείς της όρασης υπολογιστών, οι ιατρικές εικόνες είναι εξαιρετικά τυποποιημένες με υψηλή ανάλυση και δε διαθέτουν μεγάλη μεταβολή στον φωτισμό ή τη θέση, οπότε και τα μοντέλα δε μαθαίνουν στις μικρές φυσικές διαταραχές [221], (ii) η χαρακτηριστικά βιολογική υφή των ιατρικών εικόνων περιέχει πολλές περιοχές από τις οποίες είναι εύκολα να ξεγελαστούν τα μοντέλα [215], (iii) τα μοντέρνα ML/DL μοντέλα που είναι σχεδιασμένα για χρήση σε φυσικές εικόνες είναι αρκετά βαθιά και μπορούν πολύ εύκολα να οδηγηθούν σε υπερπαραμετροποίηση στην ανάλυση των ιατρικών εικόνων αυξάνοντας αυτήν την ευπάθεια [215].

Πιο συγκεκριμένα στις περιπτώσεις των ιατρικών εικόνων, τα adversarial examples μπορούν να βρουν εφαρμογή σε διάφορες εφαρμογές όπως: (i) ψηφιακές εικόνων δερματοσκόπησης (***dermoscopy images***) οι οποίες χρησιμοποιούνται για τη διάγνωση μελανώματος, (ii) σε

ακτινολογικές εικόνες (***radiology images***), όπως X-ray, PET, CT, ή MRI scans οι οποίες μπορούν να χρησιμοποιηθούν για την ανίχνευση και μέτρηση όγκων, αλλά και (iii) σε οφθαλμολογικές εικόνες (***ophthalmology images***), όπως οπτική τομογραφία ή αλλιώς optical coherence tomography (OCT) και οφθαλμοσκοπικές εξετάσεις ή αλλιώς Fundoscopic exams οι οποίες μπορούν να χρησιμοποιηθούν για την ανίχνευση της διαβητικής αμφιβληστροειδοπάθειας (DR) [221]. Οι περισσότερες δημοσιευμένες επιθέσεις επικεντρώνονται σε *MRI, X-ray* και *dermoscopy images*, καθώς είναι αυτές που περιέχονται στα περισσότερα δωρεάν σύνολα δεδομένων ιατρικών εικόνων [215].

Επίσης, τα ML μοντέλα που μπορούν να βρουν εφαρμογή σε άλλους **διαχειριστικούς τομείς τις ιατρικής**, όπως στη διαχείριση των ηλεκτρονικών ιατρικών εγγραφών (electronic healthcare record - EHR), στην ανίχνευση ασφαλιστικής οικονομικής απάτης ή ακόμα και για την έγκριση φαρμάκων και ιατρικών συσκευών, μπορεί να είναι ευάλωτα στις κλασσικές ανταγωνιστικές επιθέσεις [221].

Ένας ακόμα κίνδυνος για τα ML μοντέλα που εκπαιδεύονται με ευαίσθητα ιατρικά δεδομένα είναι οι επιθέσεις που σκοπό έχουν στην παραβίαση της ιδιωτικότητας αυτών των δεδομένων (**data privacy attacks**). Τα ML μοντέλα όπως έχουμε δει μπορεί να είναι ευάλωτα σε επιθέσεις ιδιωτικότητας των δεδομένων εκπαίδευσης (βλ. 5.1.2), κάτι το οποίο μπορεί να αποτρέψει την εκπαίδευσης τους με πραγματικά δεδομένα ασθενών και τη χρήση τους σε πραγματικές συνθήκες, καθώς θέτει σε κίνδυνο την ιδιωτικότητα των ίδιων των ασθενών [6]. Η ανωνυμοποίηση (data anonymization) των δεδομένων πολλές φορές δεν αρκεί για να θεωρηθούν τα δεδομένα ασφαλή, καθώς πολλά ιατρικά χαρακτηριστικά και διαγνώσεις μπορεί να είναι αρκετά για να αναγνωρίσουν μία μοναδική εγγραφή [171]. Οι τεχνικές των model inversion και membership inference επιθέσεων μπορούν να εφαρμοστούν και σε ML μοντέλα που εκπαιδεύονται σε ιατρικές εικόνες ή ιστορικό ασθενών, για να εξαχθούν πληροφορίες για τα δεδομένα εκπαίδευσης (εξαγωγή χαρακτηριστικών από τα δείγματα του συνόλου δεδομένων) [222] ή για την αναγνώριση αν ένα δείγμα ανήκει στο σύνολο δεδομένων αντίστοιχα [173].

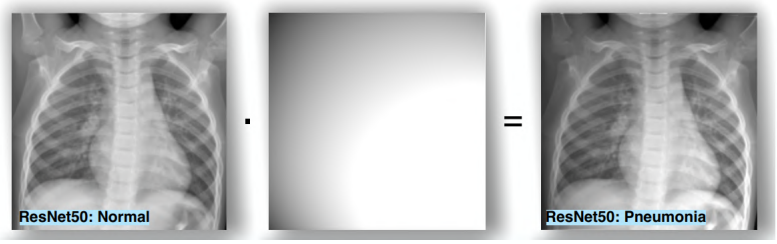
6.2.2.2 Παραδείγματα επιθέσεων

Ένα παράδειγμα ανταγωνιστικής επίθεσης σε x-ray ακτινογραφίες, φαίνεται στο σχήμα 6.14, όπου γίνεται εκμετάλλευση του πεδίου προκατάληψης (bias field) που προκαλείται από την ακατάλληλη διαδικασία λήψης ιατρικών εικόνων και υπάρχει ευρέως στις ακτινογραφίες θώρακα. Προτείνεται μια καινούρια επίθεση με όνομα **adversarial bias field attack**, στην οποία γίνεται αντικατάσταση του πρόσθετου θορύβου με ανταγωνιστικού θορύβου (adversarial noise), παραγόμενο από white-box επιθέσεις, δημιουργούνται εικόνες ακτινογραφίας οι οποίες έχουν μεγάλο attack success rate και ταυτόχρονα παρέχουν εγγυήσεις για ρεαλιστική απεικόνιση [223].

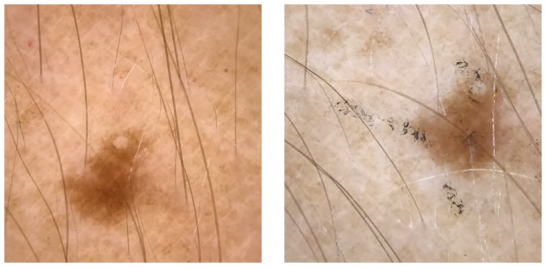
Σε ένα άλλο παράδειγμα, που φαίνεται στο σχήμα 6.14, εξετάζονται **physical adversarial** επιθέσεις σε dermoscopy images, όπου με την προσθήκη διαταραχών στις εικόνες, υπό τη μορφή κουκίδων ή γραμμών με ένα στυλό ή μαρκαδόρο, οι διαγνώσεις που κάνει το μοντέλο αλλάζουν σημαντικά (σε μελάνωμα, κηλίδες αίματος ή μώλωπες) [224].

6.2.2.3 Κατηγορίες Επιθέσεων

Αφού έχουμε αναφέρει τα κενά ασφαλείας που υπάρχουν στα ML συστήματα στο πεδίο της υγείας και των ιατρικών εικόνων και έχουμε περιγράψει κάποια παραδείγματα ανταγωνιστικών



Σχήμα 6.14: *Παράδειγμα χρήσης της adversarial bias field επίθεσης σε ακτινογραφίες θώρακος, όπου σε μια φυσιολογική εικόνα βγαίνει διάγνωση από το DNN ως πνευμονία [223].*



Σχήμα 6.15: *Παράδειγμα physical ανταγωνιστικής επίθεσης σε dermoscopy images, όπου με προσθήκη γραμμών με ένα στυλό (δεξιά εικόνα) γίνεται διαφορετική διάγνωση σε σχέση με την καθαρή (αριστερή εικόνα) [224].*

επιθέσεων, σε αυτήν την ενότητα γίνεται μια προσπάθεια ταξινόμησης αυτών με βάση την ταξινόμηση και το μοντέλο απειλής που έχουμε κάνει στην ενότητα 3.4.1 για τις ανταγωνιστικές επιθέσεις γενικότερα.

Όσον αφορά το περιβάλλον εφαρμογής αυτών των επιθέσεων, όπως είδαμε και από τα παραδείγματα μπορεί να είναι: (i) **Physical**, δηλαδή επιθέσεις που συμβαίνουν στον πραγματικό (φυσικό) κόσμο, με εισαγωγή διαταραχών σε πραγματικές εικόνες ή αντικείμενα στον πραγματικό χώρο, ή (ii) **Digital**, δηλαδή επιθέσεις που συμβαίνουν στον ψηφιακό κόσμο, με ψηφιακή επεξεργασία εικόνων. Οι Physical επιθέσεις συνήθως εμφανίζουν μικρότερη βεβαιότητα πρόβλεψης, διάγνωσης ή ταξινόμησης στα adversarial examples σε σχέση με τις αντίστοιχες καθαρές εικόνες από ότι εμφανίζουν με τις αντίστοιχες ψηφιακές διαταραχές σε σχέση με τις αντίστοιχες καθαρές εικόνες [224].

Όπως έχουμε αναφέρει και γενικότερα (βλ. 3.4.1) οι ανταγωνιστικές επιθέσεις σε ιατρικές εφαρμογές, μπορούν να κατηγοριοποιηθούν με βάση την ικανότητα του επιτιθέμενου, οπότε έχουμε: (i) **Poisoning Attacks**, και (ii) **Evasion Attacks**. Οι poisoning επιθέσεις είναι πιο σχετικές στις ιατρικές εφαρμογές, καθώς ο άμεσος χειρισμός των δεδομένων είναι αρκετά δύσκολος, σε αντίθεση με την εισαγωγή νέων δειγμάτων σε σύνολα δεδομένων εκπαίδευσης. Οι evasion επιθέσεις είναι πιθανές και μπορεί να προκληθούν, αν και συνήθως συναντούνται υπό τη μορφή μη ηθελημένων επιθέσεων, όπου μικρές φυσικές διαταραχές μπορεί να προκαλέσουν λάθος ταξινόμηση. Όμως γενικότερα, έχουν μελετηθεί και σενάρια κακόβουλων ασθενών ή εργαζόμενων υγείας, στα οποία προστίθεται ανταγωνιστικός θόρυβος για να προκληθεί λανθασμένη διάγνωση με σκοπό το οικονομικό κέρδος [221], για παράδειγμα κλινικές ή ιατροί μπορεί να διαταράσσουν τις ιατρικές εικόνες ώστε να οδηγήσουν σε άσκοπες εγχειρήσεις για επιπλέον κέρδος [215].

Επίσης, όσον αφορά τη γνώση του επιτιθέμενου, έχουν δημοσιευθεί ανταγωνιστικές επιθέσεις σε **White-box** και σε **Black-box** σενάριο επίθεσης, σε διάφορες εφαρμογές [6].

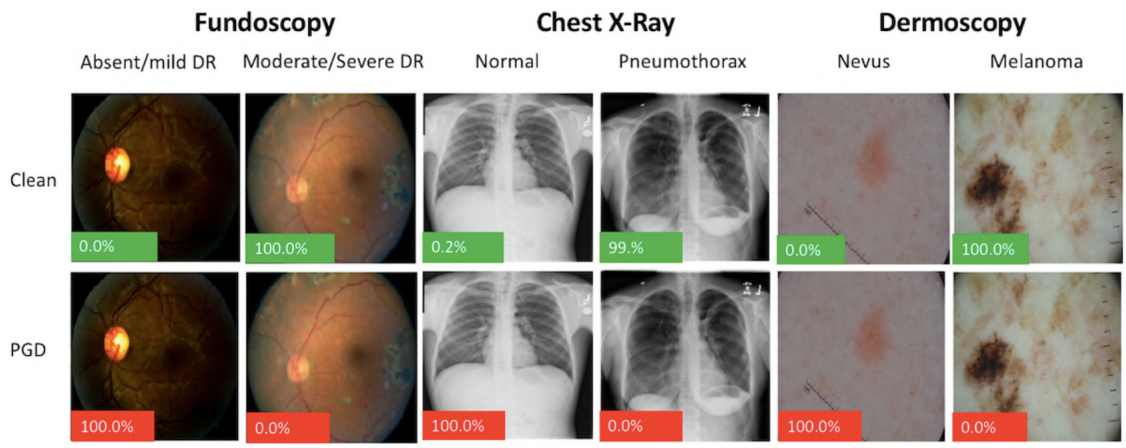
Για τις ανταγωνιστικές επιθέσεις τις οποίες μελετάμε, μπορούμε να τις ταξινομήσουμε επίσης, ανάλογα την εφαρμογή των ML μοντέλων στις ιατρικές εικόνες, όπως αναλύσαμε και προηγουμένως και αντιστοιχούν στον στόχο του επιτιθέμενου, δηλαδή το είδος εικόνων που θέλει να επιτεθεί: **MRI**, **X-Ray**, **PET**, **Ultrasound**, **CT**, **Histology**, **Fundoscopy**.

Για παράδειγμα, στο [225], εξετάζονται ανταγωνιστικές επιθέσεις για δύο διαφορετικές εφαρμογές. Για Classification δερματικών κακώσεων με χρήση **Dermoscopy** εικόνων, και για Segmentation ολόκληρου του εγκεφάλου με χρήση **MRI** εικόνων. Στην πρώτη περίπτωση, αξιολογούνται οι επιθέσεις FGSM, DeepFool, JSMA σε Black-box σενάριο πάνω σε 3 state-of-the-art deep learning μοντέλα: InceptionV3 (IV3), InceptionV4 (IV4), MobileNet (MN), ενώ στη δεύτερη μια ειδική επίθεση που μοιάζει με τον FGSM, αλλά προσαρμοσμένο στο να δημιουργεί adversarial examples ανά τμήμα ονομαζόμενη Dense Adversarial Generation (DAG) πάνω σε 3 δημοφιλή CNN μοντέλα: SegNet (SN), UNet (UN), DenseNet(DN). Για τις **Classification** εργασίες, στην FGSM επίθεση η ακρίβεια έχει πτώση 26% στο IV3, 40% στο IV4, 59% στο MN, στην DeepFool επίθεση η ακρίβεια έχει πτώση 0% στο IV3, 5% στο IV4, 13% στο MN, ενώ στην JSMA επίθεση η ακρίβεια έχει πτώση 0% στο IV3, 4% στο IV4, 14% στο MN. Για τις **Segmentation** εργασίες, στην καλύτερη έκδοση της DAG επίθεσης η ακρίβεια έχει πτώση 57% στο SN, 43% στο UN, 45% στο DN.

Στο [221], χρησιμοποιείται PGD επίθεση υπό White-box και Black-box σενάριο (transfer-based), σε εικόνες **Fundoscopy**, **X-ray**, **Dermoscopy** για Classification εργασίες, χρησιμοποιώντας ένα προεκπαιδευμένο ResNet50 μοντέλο, όπου και παρατηρείται δραματική πτώση της ακρίβειας. Επίσης, χρησιμοποιήθηκαν και επιθέσεις με adversarial patches χρησιμοποιώντας φυσικές εικόνες. Συγκεκριμένα, η επίθεση ρίχνει την ακρίβεια στο 0% σε όλες τις εικόνες σε White-box σενάριο (βλ. εικόνα 6.16). Σε Black-box σενάριο για τις Fundoscopy εικόνες η ακρίβεια πέφτει πάλι στο 0% (από 91% αρχικά), στις X-ray εικόνες στο 15.1% (από 94.9% αρχικά), ενώ στις Dermoscopy στο 37.9% (από 87.6% αρχικά). Τα adversarial patches με φυσικές εικόνες παρουσιάζουν πολύ καλύτερα αποτελέσματα αλλά και πάλι με μείωση ακρίβειας 10%-20%.

Στο [226], αξιολογούνται FGSM, PGD επιθέσεις υπό White-box σενάριο, σε **Dermoscopy** εικόνες για ανίχνευση καρκίνου του δέρματος και ταξινόμηση του σε 7 κατηγορίες (Classification), χρησιμοποιώντας δύο δημοφιλή deep learning μοντέλα: MobileNet, VGG16. Στην FGSM επίθεση η ακρίβεια έχει πτώση 74% στο MobileNet και πτώση 61% στο VGG16, και παρόμοια στην PGD επίθεση η ακρίβεια έχει πτώση 74% στο MobileNet και πτώση 63% στο VGG16.

Στο [227], χρησιμοποιείται FGSM επίθεση υπό White-box σενάριο για την ανίχνευση συμπτωμάτων COVID-19 (Classification) μέσω **X-Ray** ακτινογραφιών θώρακος και **CT** εικόνων, χρησιμοποιώντας δύο δημοφιλή προεκπαιδευμένα deep learning μοντέλα: VGG16, InceptionV3. Στις X-Ray εικόνες, η χρήση της FGSM επίθεσης με $\epsilon = 0.009$ έχει πτώση στην ακρίβεια σχεδόν 80% στο VGG16 και 40% στο InceptionV3. Στις CT εικόνες, η χρήση της FGSM επίθεσης έχει πτώση στην ακρίβεια σχεδόν 45% με $\epsilon = 0.003$ στο VGG16 και 40% με $\epsilon = 0.0007$ στο InceptionV3. Σε όλες τις περιπτώσεις η διαταραχή είναι πολύ μικρή ώστε να μην προκαλεί αναγνωρίσιμες παραμορφώσεις από ανθρώπινο μάτι, αλλά αρκετό για να



Σχήμα 6.16: *Αποτελέσματα White-box PGD επίθεσης σε classification εργασίες πάνω σε Fundoscopy, X-ray, Dermoscopy ιατρικές εικόνες. Στην πρώτη γραμμή φαίνονται οι καθαρές εικόνες, ενώ στη δεύτερη γραμμή οι εικόνες μετά την PGD επίθεση. Σε κάθε εικόνα φαίνεται το ποσοστό πιθανότητας για την κάθε διάγνωση που φαίνεται σε κάθε στήλη [221].*

ξεγελάσει τα μοντέλα.

Στο [228], χρησιμοποιείται Universal Adversarial Perturbations (UAP) επίθεση υπό Black-box σενάριο σε εικόνες **MRI** 4 διαφορετικών απεικονίσεων (T1, T2, T1ce, FLAIR) για την ανίχνευση όγκων εγκεφάλου (Segmentation) σε 3 παραλλαγές του U-net νευρωνικού δικτύου, για διάφορα μεγέθη διαταραχών. Τα καλύτερα αποτελέσματα εμφανίζονται όταν η επίθεση εφαρμόζεται και στις 4 διαφορετικές απεικονίσεις με πτώση ακρίβειας από 30% έως και 65%. Στο [229], για την εργασία ταξινόμησης όγκων εγκεφάλου (Classification) μέσω **MRI** εικόνων από CNN μοντέλο, δοκιμάζονται επιθέσεις FGSM, Noise-based, Virtual Adversarial Training (VAT) σε White-box σενάριο. Στην FGSM επίθεση η ακρίβεια έχει πτώση 24%, στην Noise-based επίθεση η ακρίβεια έχει πτώση 69%, και στην VAT επίθεση η ακρίβεια έχει πτώση 35%.

Πέρα όμως από τις κλασσικές ανταγωνιστικές επιθέσεις, στα δύο παραδείγματα επιθέσεων που παρουσιάσαμε στο 6.2.2.2, χρησιμοποιούνται δύο διαφορετικές επιθέσεις που έχουν σχεδιαστεί ειδικά για τις εφαρμογές που στοχεύουν. Στο [223] (βλ. εικόνα 6.14), χρησιμοποιείται η adversarial bias field attack (AdvSBF) επίθεση για τη διάγνωση παθήσεων του πνεύμονα (Classification) μέσω **X-Ray** ακτινογραφιών θώρακος, πάνω σε 3 δημοφιλή deep learning μοντέλα: ResNet50, Dense121, MobileNet. Η AdvSBF επίθεση έχει Attack Success Rate (ASR) 38.69% στο ResNet50, 34.49% στο Dense121, και 33.51% στο MobileNet, έχοντας επίσης πολύ καλύτερα ASR και σε transfer επιθέσεις σε άλλα μοντέλα, από τις κλασσικές adversarila επιθέσεις. Στο [224] (βλ. εικόνα 6.15), χρησιμοποιείται η ανεπτυγμένη Physical ανταγωνιστική επίθεση σε **Dermoscopy** εικόνες για το Classification δερματικών κακώσεων πάνω σε 5 δημοφιλή μοντέλα: ResNet, InceptionV3, InceptionResNetV2, MobileNet, Xception. Σε αυτήν την επίθεση δοκιμάζονται διάφορες φυσικές διαταραχές, όπου όλες είναι επιτυχημένες σε όλα τα μοντέλα με μέγιστη μείωση της ακρίβειας στο 60% για κάποιες και μέση πτώση ακρίβειας 30.8%.

Στον πίνακα 6.6, φαίνονται συνοπτικά όλες οι παραπάνω επιθέσεις. Γενικά οι FGSM, PGD επιθέσεις φαίνεται να είναι οι πιο αποδοτικές, και οι περισσότερες έρευνες εστιάζουν

σε εφαρμογές σε MRI, Dermoscopy και X-ray εικόνες, καθώς είναι και αυτές που είναι πιο εύκολα διαθέσιμες.

Στόχος	Γνώση	Μοντέλα	Επιθέσεις	Εργασίες	Αποτελέσματα
Dermoscopy, MRI	Black-Box	Inception, MobileNet, SegNet, U-Net, DenseNet	[225]: FGSM, DeepFool, JSMA, DAG	Classification, Segmentation	Πτώση ακρίβειας: (Classification): FGSM 26% IV3, 40% IV4, 59% MN, DeepFool 0% IV3, 5% IV4, 13% MN, JSMA 0% IV3, 4% IV4, 14% MN. (Segmentation) DAG 57% στο SN, 43% UN, 45% DN.
Dermoscopy, Fundoscopy, X-Ray	White-Box, Black-Box	ResNet50	[221]: PGD	Classification	Ακρίβεια: (White-box) 0% παντού. (Black-box) Fundoscopy 0%, στις X-ray εικόνες στο 15.1%, Dermoscopy 37.9%, Adv.Patches 10%-20%.
Dermoscopy	White-Box	MobileNet, VGG16	[226]: FGSM, PGD	Classification	Πτώση ακρίβειας: FGSM 74% MobileNet, 61% VGG16. PGD 74% MobileNet, 63% VGG16.
X-Ray, CT	White-Box	VGG16, InceptionV3	[227]: FGSM	Classification	Πτώση ακρίβειας: (X-Ray) FGSM 80% VGG16, 40% InceptionV3. (CT) FGSM 45% VGG16, 40% InceptionV3
MRI	Black-Box	U-net	[228]: UAP	Segmentation	Πτώση ακρίβειας 30%-65%.
MRI	White-Box	CNN	[229]: Noise-based, FGSM, VAT	Classification	Πτώση ακρίβειας: FGSM 24%, Noise-based 69%, VAT 35%.
X-Ray	White-Box, Black-Box	ResNet50, Dense121, MobileNet	[223]: AdvSBF	Classification	ASR: 38.69% ResNet50, 34.49% Dense121, 33.51% MobileNet
Dermoscopy	Black-Box	ResNet, InceptionV3, Inception-ResNetV2, MobileNet, Xception	[224]: Physical	Classification	Μέγιστη μείωση της ακρίβειας 60% και μέση πτώση ακρίβειας 30.8%

Πίνακας 6.6: *Πίνακας με ανταγωνιστικές επιθέσεις σε ML μοντέλα ιατρικών εικόνων.*

6.2.3 Άμυνες - Μέτρα Ασφαλείας

6.2.3.1 Κατηγορίες Αμυνών

Σε αυτήν την ενότητα κάνουμε αναφορά στις τεχνικές αμυνών που μπορούν να χρησιμοποιηθούν και έχουν ερευνηθεί για να αναπτυχθούν πιο εύρωστα ML μοντέλα στο πεδίο της υγείας και των medical images, ενάντια σε ανταγωνιστικές επιθέσεις, και τις ταξινομούμε με βάση και την κατηγοριοποίηση στην ενότητα 3.6.1 για τις άμυνες ενάντια σε ανταγωνιστικές επιθέσεις γενικότερα.

Αρχικά, ανάλογα τώρα με την προσέγγιση αντιμετώπισης, μπορούμε να κατηγοριοποιήσουμε τις άμυνες, όπως και στη γενική περίπτωση σε:

- **Data preprocessing**: Άμυνες που τροποποιούν τα δεδομένα εκπαίδευσης ή δοκιμής και των χαρακτηριστικών τους, για την ελαχιστοποίηση ή την αποφυγή των adversarial examples.
- **Model hardening**: Άμυνες που τροποποιούν κυρίως τα χαρακτηριστικά του μοντέλου, εκπαιδεύοντας το με σκοπό την ευρωστία ενάντια σε adversarial examples.

- **Auxiliary models**: Άμυνες που χρησιμοποιούν επιπλέον ML (ή και άλλα είδη) μοντέλα που κάνουν εξειδικευμένες εργασίες, για να ενισχύσουν την ευρωστία του κύριου μοντέλου.

6.2.3.2 Τεχνικές Αμυνών

Στις *Model Hardening* άμυνες, το **Adversarial Training** είναι από τις πιο συνηθισμένες άμυνες. Στο [230], χρησιμοποιείται adversarial training με χρήση του PGD αλγόριθμου, για την εκπαίδευση ενός εύρωστου μοντέλου DR (diabetic retinopathy) Classification από retina images. Μετά την εκπαίδευση το μοντέλο (ResNet32), παρουσιάζει αύξηση του adversarial accuracy από 43% σε 83%. Στο [231], γίνεται επίσης adversarial training με χρήση του PGD αλγόριθμου, για την εκπαίδευση ενός εύρωστου 3D ResUNets μοντέλου για ανίχνευση πνευμονικών οζιδίων μέσω CT εικόνων. Η PGD επίθεση χρησιμοποιείται για την ανίχνευση των μοτίβων που οδηγούν σε λάθος ταξινόμηση με υψηλή βεβαιότητα, και χρησιμοποιήθηκαν για την εκπαίδευση του εύρωστου μοντέλου, το οποίο μπορεί και τα ανιχνεύει με ακρίβεια 87%, αλλά και δείγματα τα οποία περιέχουν απλό και όχι ανταγωνιστικό θόρυβο. Στο, [229] γίνεται adversarial training με χρήση 3 διαφορετικών αλγορίθμων FGSM, Noise-based, Virtual Adversarial Training (VAT), για την εκπαίδευση ενός εύρωστου CNN μοντέλου που επιτελεί εργασία ταξινόμησης όγκων εγκεφάλου (Classification) μέσω MRI εικόνων. Μετά την εκπαίδευση το μοντέλο παρουσιάζει αυξήσεις στο adversarial accuracy για κάθε μέθοδο, οι οποίες είναι πολύ κοντά και στο clean accuracy, με λιγότερη αποτελεσματική τη VAT. Στην FGSM επίθεση η ακρίβεια έχει αύξηση 24.6% (συν. 86.79%), στην Noise-based επίθεση η ακρίβεια έχει αύξηση 77% (συν. 89.07%), και στην VAT επίθεση η ακρίβεια έχει αύξηση 22% (συν. 75.20%).

Στο [232], γίνεται adversarial training με χρήση 2 διαφορετικών αλγόριθμων (FGSM, JSMA), αλλά και φυσικών διαταραχών υπό τη μορφή Gaussian Noise (**Data Augmentation** τεχνική), για την εκπαίδευση εύρωστων CNN μοντέλων για ανίχνευση καρκίνου του πνεύμονα μέσω CT εικόνων του πνεύμονα ή για ανίχνευση εγκεφαλικών όγκων μέσω MRI εικόνων του εγκεφάλου. Τα μοντέλα εκπαιδευμένα με τα FGSM, JSMA δείγματα έχουν την καλύτερη επίδοση, από τη μέθοδο προσθήκης κανονικού θορύβου, καθώς χρειάζονται πολύ μεγαλύτερες διαταραχές για να μη είναι αποτελεσματικές, σε σημείο όμως που είναι εμφανή με το ανθρώπινο μάτι. Αντίστοιχα, στο [233] χρησιμοποιείται **Data Augmentation** τεχνική για αύξηση του συνόλου δεδομένων εκπαίδευσης με δείγματα τα οποία περιέχουν διαταραχές που είναι εγγενής στις MRI ιατρικές εικόνες λόγω της ανομοιογένειας παραγωγής τους (bias field). Αυτή η μέθοδος ενισχύει την ευρωστία των μοντέλων και δοκιμάζεται σε U-Net μοντέλο που εκτελεί Segmentation MRI εικόνων καρδιάς. Συγκεκριμένα, συγκρίνεται με άλλες μεθόδους όπως Random Augmentation, Mixup και VAT στο οποίο βασίζεται αυτή η μέθοδος (AdvBias), και έχει τις καλύτερες επιδόσεις.

Στα [234], [235], χρησιμοποιείται μια ειδική τεχνική **Robust Training**, που εφαρμόζεται σε μοντέλα U-Net, I-RIM που εκτελεί εργασία MRI reconstruction. Αυτή λειτουργεί βρίσκοντας τις χειρότερες περιπτώσεις false negatives δειγμάτων (false-negative adversarial feature ή FNAF), και χρησιμοποιώντας τα σε ανταγωνιστική εκπαίδευση, για τη δημιουργία ενός εύρωστου μοντέλου, με λιγότερα false negatives. Αυτή η μέθοδος δείχνει να βελτιώνει σημαντικά την αποτελεσματικότητα του μοντέλου, και οι ανακατασκευασμένες εικόνες

αξιολογούνται με τιμές δομικής ομοιότητας SSIM 0.7197 ± 0.2613.

Στο [236], προτείνεται μια τεχνική που ενσωματώνει ένα επίπεδο auto-encoder σε ένα CNN νευρωνικό (**Modify Network**), για την αφαίρεση θορύβου, η οποία μπορεί να συνδυαστεί και με άλλες αμυντικές μεθόδους για να κάνει τις διαγνώσεις πιο ανθεκτικές. Η μέθοδος αξιολογείται σε X-Ray εικόνες και Dermoscopy εικόνες υπό FGSM, IFGSM, C&W επιθέσεις και σε κάθε περίπτωση αυξάνει σημαντικά το adversarial accuracy. Συγκεκριμένα, στις X-Ray εικόνες, για $\epsilon = 4$ στην FGSM επίθεση η ακρίβεια έχει αύξηση 50% (συν. 71.21%), στην IFGSM έχει αύξηση 44% (συν. 72.35%), και στην C&W έχει αύξηση 48% (συν. 60.87%). Αντίστοιχα, στις Dermoscopy εικόνες, για $\epsilon = 4$ στην FGSM επίθεση η ακρίβεια έχει αύξηση 31% (συν. 70.21%), στην IFGSM έχει αύξηση 23% (συν. 74%), και στην C&W έχει αύξηση 40% (συν. 54.73%).

Στις *Data preprocessing* άμυνες, μια από τις τεχνικές που έχουν χρησιμοποιηθεί είναι αυτή στο [237], η οποία προσομοιάζει σε λειτουργία το **MagNet** (βλ. 3.6.2.9), δηλαδή ο μηχανισμός μετατρέπει μια εικόνα στον τομέα συχνότητας με διακριτό μετασχηματισμό Fourier, το οποίο βοηθάει στον διαχωρισμό καθαρών εικόνων από ανταγωνιστικές εικόνες, σε εργασίες Segmentation ιατρικών εικόνων. Η διαφορά με το MagNet είναι ότι δε χρησιμοποιούνται auto-encoders, αλλά deep semantic segmentation models, όπως U-Net, DenseNet, και χρησιμοποιείται το πεδίο της συχνότητας για την ανίχνευση των adversarial examples, αντί του χώρου. Αυτή η μέθοδος δε χρειάζεται γνώση σχετικά με την αρχιτεκτονική του μοντέλου ή τα adversarial examples, οπότε είναι εύκολα εφαρμόσιμη. Η καλύτερη εκδοχή αυτού του μηχανισμού ανιχνεύει τα adversarial examples με μέση ακρίβεια 98%. Στο [238], προτείνεται η Fuzzy Unique Image Transformation (**FUIT**) μέθοδος για προστασία από adversarial examples για διαγνωστικά μοντέλα για τον COVID-19 μέσω X-Ray και CT εικόνες. Αυτή η μέθοδος, απεικονίζει τα pixel μιας εικόνας σε ένα συγκεκριμένο διάστημα (downsampling) και τροφοδοτεί τη μετασχηματισμένη είσοδο στο διαγνωστικό CNN μοντέλο, το οποίο κάνει προβλέψεις με πολύ μεγάλη ακρίβεια και σε clean και σε adversarial δείγματα, χωρίς καμία αλλαγή του διαγνωστικού μοντέλου, με ένα μικρό αντίκτυπο στην αύξηση του χρόνου πρόβλεψης, λόγο του επιπλέον βήματος μετασχηματισμού. Σε 6 μη στοχευμένες επιθέσεις (FGSM, BIM, PGD, PGD-r, Deep Fool, C&W), το adversarial accuracy είναι πάνω από 95%, ενώ η clean ακρίβεια παραμένει σχεδόν ίδια με την αρχική (μέγιστη μείωση 2%).

Στις άμυνες που κάνουν χρήση *Auxiliary models* άμυνες, η πιο κοινή τεχνική είναι η **Adversarial Example Detection**, για την ανίχνευση των κακόβουλων δειγμάτων πριν καν εισέλθουν στο μοντέλο, έχοντας και το πρόσθετο πλεονέκτημα, ότι δεν επηρεάζουν την ακρίβεια των μοντέλων, εφόσον δεν επανεκπαιδεύονται, κάτι το οποίο είναι πολύ σημαντικό στα μοντέλο που χρησιμοποιούν ιατρικά δεδομένα, καθώς είναι πολύ δύσκολη η απόκτηση πολλών και διαφορετικών δεδομένων για την αύξηση της ακρίβειας. Στο [239], προτείνεται αμυντική μέθοδος ανίχνευσης adversarial examples για διαγνωστικά deep learning μοντέλα που επιτελούν εργασίες Classification, Segmentation, Object detection πάνω σε X-Ray και Dermoscopy εικόνες. Η τεχνική αυτή χρησιμοποιεί μια μη γραμμική ακτινική βάση συνελικτικών χαρτογραφημένων χαρακτηριστικών που μαθαίνονται από μια Mahalanobis συνάρτηση απόστασης. Με αυτόν τον τρόπο τα δείγματα τα οποία βρίσκονται μακριά από το όριο απόφασης χαρακτηρίζονται ως adversarial, παρόμοια με τις out-of-distribution τεχνικές ανίχνευσης που αναλύσαμε στην ενότητα 3.6.2.11. Η τεχνική αυτή δεν επηρεάζει την πολυπλοκότητα των

μοντέλων, και μπορεί να εφαρμοστεί σε οποιοδήποτε αλλάζοντας απλά τη συνάρτηση ενεργοποίησης. Για παράδειγμα, στις Classification εργασίες, για ℓ_2 εμφανίζει adversarial accuracy από 60.58% (C&W) έως και 98.79% (C&W), ενώ για ℓ_∞ από 13% (BIM) έως και 91.31% (FGSM) σε διάφορα μοντέλα. Αντίστοιχα, στο [240] χρησιμοποιείται η ίδια μετρική απόστασης Mahalanobis για ανίχνευση των **out-of-distribution** δειγμάτων. Η τεχνική αυτή αξιολογείται σε εικόνες μικροσκοπίου (Microscopy) από δείγματα αίματος για την ανίχνευση ελονοσίας από VGG-19, ResNet-18 μοντέλα, και δοκιμάζεται ενάντια των FGSM, BIM, C&W, DeepFool επιθέσεων, παρουσιάζοντας state-of-the-art επιδόσεις, με ακρίβεια ανίχνευσης από 61.95% (DeepFool) έως και 99.95% (FGSM).

Γενικά, παρατηρούμε ότι η πιο συχνή μέθοδος άμυνας είναι το adversarial training ενός εύρωστου μοντέλου, καθώς το κόστος εκπαίδευσης δεν είναι μεγάλο σε σχέση με άλλα πεδία εφαρμογής λόγω των λιγότερων διαθέσιμων δεδομένων και γιατί είναι από τις πιο αποτελεσματικές άμυνες. Το αρνητικό είναι όμως η τάση της πτώσης της ακρίβειας στα καθαρά δεδομένα, η οποία σε αυτά το μοντέλα είναι συνήθως υψηλή λόγω υπερπροσαρμογής, οπότε το adversarial training προσφέρει ένα είδος regularization το οποίο μπορεί να βοηθήσει τη γενικότητα των μοντέλων [215]. Η ανίχνευση των adversarial examples είναι επίσης αρκετά σημαντική μέθοδος για τη δημιουργία εύρωστων ML μοντέλων, καθώς ανιχνεύει τα δείγματα με πολύ μεγάλη ακρίβεια, λόγω το ότι οι διαταραχές είναι πιο εύκολο να ανιχνευθούν σε ιατρικές εικόνες από ότι σε φυσικές εικόνες, δε μειώνει την ακρίβεια του αρχικού μοντέλου και μπορούν να χρησιμοποιηθούν και για την ανίχνευση poisoning ανταγωνιστικών επιθέσεων [6]. Στον πίνακα 6.7 φαίνονται συγκεντρωτικά οι άμυνες που αναλύσαμε ενάντια σε ανταγωνιστικές επιθέσεις σε ML διαγνωστικά συστήματα υγείας ιατρικών εικόνων.

6.2.3.3 Προστασία Ιδιωτικότητας

Η διατήρηση της ιδιωτικότητας των χρηστών στο πεδίο της υγείας είναι εξαιρετικά σημαντική, καθώς περιλαμβάνει τη συλλογή προσωπικών δεδομένων και οποιαδήποτε παραβίαση μπορεί να οδηγήσει σε αναπόφευκτες συνέπειες. Τα μοντέλα μηχανικής μάθησης (ML) δεν πρέπει να αποκαλύπτουν πρόσθετες πληροφορίες για οποιαδήποτε πληροφορία έχει χρησιμοποιηθεί κατά την εκπαίδευσης τους [6]. Για την προστασία λοιπόν, των διαγνωστικών ML συστημάτων και των ευαίσθητων ιατρικών δεδομένων, μπορούν να αξιοποιηθούν οι γενικότεροι μέθοδοι προστασίας των δεδομένων, όπως αυτούς που παρουσιάσαμε στην ενότητα 5.3.2, οι οποίοι προσαρμόζονται στο πεδίο της ιατρικής, για την προστασία των δεδομένων ενός ML μοντέλου, δημιουργώντας έτσι μοντέλα που διασφαλίζουν την ασφάλεια και την ιδιωτικότητα των δεδομένων (PPML) [173]. Οι παρακάτω μέθοδοι μπορούν να προσφέρουν εγγυήσεις για την ιδιωτικότητα και την ακεραιότητα των μοντέλων και των δεδομένων, ή μπορούν να εξασφαλίζουν ότι ακόμα και στην περίπτωση εξαγωγής δεδομένων οι επιπτώσεις θα είναι οι ελάχιστες δυνατές.

Μια από τις πιο αποτελεσματικές μεθόδους για την εγγύηση της ιδιωτικότητας είναι η **Differential Privacy (DP)** (βλ. 5.3.3), όπου διασφαλίζεται ότι επιτιθέμενοι δεν μπορούν να αναγνωρίσουν εάν ένα δείγμα υπάρχει στο σύνολο δεδομένων εκπαίδευσης, εξασφαλίζοντας εξ ορισμού προστασία από membership inference επιθέσεις. Η λειτουργία του στηρίζεται στην εισαγωγή στατιστικού θορύβου στα δεδομένα, προσφέροντας ισχυρές εγγυήσεις ιδιωτικότητας (privacy guarantees), θυσιάζοντας όμως χρησιμότητα λόγω αυτού του εισαγομένου

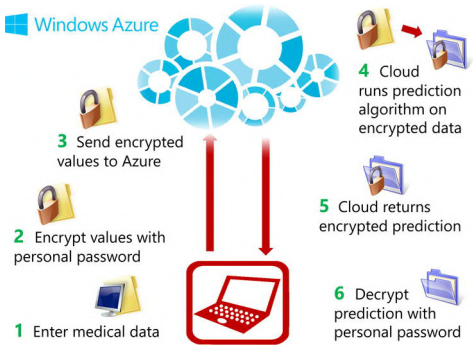
Στόχος	Προσέγγιση	Άμυνα	Εργασίες	Αποτελέσματα
Fundoscopy	Model Hardening	[230]: Adversarial Training (PGD)	Classification	Αύξηση adversarial accuracy από 43% σε 83%
X-Ray	Model Hardening	[231]: Adversarial Training (PGD)	Classification	Ανίχνευση adversarial example με ακρίβεια 87%
MRI	Model Hardening	[229]: Adversarial Training (FGSM, Noise, VAT)	Classification	Αύξηση ακρίβειας: FGSM 24.6% (συν. 86.79%), Noise-based 77% (συν. 89.07%), VAT 22% (συν. 75.20%)
CT	Model Hardening	[232]: Adversarial Training / Data Augmentation (Gaussian Noise)	Classification, Segmentation	Εκπαίδευση με FGSM, JSMA δείγματα, καλύτερα από θόρυβο Gauss
MRI	Model Hardening	[233]: Data Augmentation (Bias Field)	Segmentation	AdvBias καλύτερες επιδόσεις από Random Augmentation, Mixup, VAT μεθόδους
MRI	Model Hardening	[234], [235]: Robust Training (FNAF)	Reconstruction	Βελτίωση ανακατασκευασμένων εικόνων με SSIM 0.7197±0.2613
X-Ray, Dermoscopy	Model Hardening	[236]: Embedded denoising auto-encoder	Classification	Αύξηση ακρίβειας ($\epsilon = 4$): (X-Ray) FGSM 50%, IFGSM 44%, C&W 48%. (Dermoscopy) FGSM 31%, IFGSM 23%, C&W 40%.
MRI	Data Preprocessing	[237]: MagNet with Fourier	Segmentation	Ακρίβεια ανίχνευσης adversarial example: 98% (average)
X-Ray, CT	Data Preprocessing	[238]: FUIT (image downsampling)	Classification	Adversarial Accuracy είναι πάνω από 95%
X-Ray, Dermoscopy	Auxiliary Model	[239]: Detection Out-of-Distribution	Classification, Segmentation	Ακρίβεια ανίχνευσης (Classification): Για ℓ_2 από 60.58% (C&W) έως 98.79% (C&W). Για ℓ_∞ από 13% (BIM) έως 91.31% (FGSM)
Microscopy	Auxiliary Model	[240]: Detection Out-of-Distribution	Classification	Ακρίβεια ανίχνευσης από 61.95% (DeepFool) έως και 99.95% (FGSM).

Πίνακας 6.7: *Πίνακας με ανταγωνιστικές άμυνες σε ML μοντέλα ιατρικών εικόνων.*

θορύβου, ανάλογα και με το επιλεγμένο privacy budget με βάση τις απαιτήσεις ασφαλείας [6]. Η εφαρμογή του DP σε ML μοντέλα, μπορεί να γίνει με διάφορες τεχνικές, όπως DP-SGD [162], ή PATE [241], για εφαρμογές σε φυσικές εικόνες. Η χρήση του DP έχει ερευνηθεί σε εφαρμογές υγείας, όπως στο [242] όπου εφαρμόζεται σε ιατρικά δεδομένα ασθενών για την εκπαίδευση διαφορίσιμα ιδιωτικού δέντρου αποφάσεων για πρόβλεψη ζαχαρώδη διαβήτη. Επίσης, έχει ερευνηθεί και η εφαρμογή του σε δεδομένα ιατρικών εικόνων, όπως στο [243] όπου χρησιμοποιείται ο DP-SGD αλγόριθμος για την εκπαίδευση ενός διαφορίσιμα ιδιωτικού νευρωνικού δικτύου που εκτελεί εργασίες ταξινόμησης παιδικής πνευμονίας από X-Ray ακτινογραφίες θώρακα και κατάτμησης CT εικόνων ήπατος. Ενώ πρόκειται για μια πολύ ισχυρή τεχνική που έχει άμεση εφαρμογή σε απλά δεδομένα, όπως ιατρικά αρχεία και ιστορικό ασθενών, η εφαρμογή του σε δεδομένα εικόνας δεν είναι τόσο ξεκάθαρη και δεν έχει ακόμα ευρεία εφαρμογή σε εργασίες ιατρικών εικόνων [222].

Μια άλλη μέθοδος για την προστασία της ιδιωτικότητας είναι μέσω της χρήσης **Homomorphic Encryption (HE)** (βλ. 5.3.5), όπου τα δεδομένα κρυπτογραφούνται με τέτοιο τρόπο ώστε να είναι δυνατόν να γίνουν οι υπολογισμοί που χρειάζονται για τις τεχνικές μηχανικής μάθησης. Με αυτόν τον τρόπο είναι αδύνατη η εξαγωγή των πραγματικών δεδομένων, καθώς θα πρέπει να αποκρυπτογραφηθούν από δημόσια κλειδιά τα οποία έχουν μόνο οι κάτοχοι των δεδομένων [6]. Η εφαρμογή του HE σε ML μοντέλα, μπορεί να γίνει με διάφορες τεχνικές.

Ένα από τα πιο δημοφιλή νευρωνικά δίκτυα που δουλεύει πάνω σε κρυπτογραφημένα δεδομένα είναι το Crypto-Nets [244], το οποίο έχει μελετηθεί με βάση εφαρμογές όπου τα δεδομένα είναι πολύ ευαίσθητα, όπως στην ιατρική. Γενικότερα, η χρήση του HE σε εφαρμογές της υγείας έχει ήδη ερευνηθεί, κυρίως για την *κρυπτογράφηση των ιατρικών δεδομένων των ασθενών*, όπως στο [245] όπου τα ιατρικά δεδομένα ασθενών κρυπτογραφούνται με HE, στέλνονται στο Cloud για εκτέλεση προγνωστικής ανάλυσης (predictive analysis), με σκοπό την πρόβλεψη καρδιαγγειακών νοσημάτων, όπως καρδιακή προσβολή, χωρίς να μαθαίνεται ποτέ κάποια πληροφορία για οποιαδήποτε ιατρική εγγραφή (βλ. σχήμα 6.17). Ακόμη η εφαρμογή της HE έχει μελετηθεί και σε *εφαρμογές βιοπληροφορικής*, για την εκπαίδευση privacy-preserving μοντέλων μηχανικής μάθησης για γενετικές διεργασίες, όπου τα δεδομένα είναι ανθρώπινα γονιδιώματα [246].

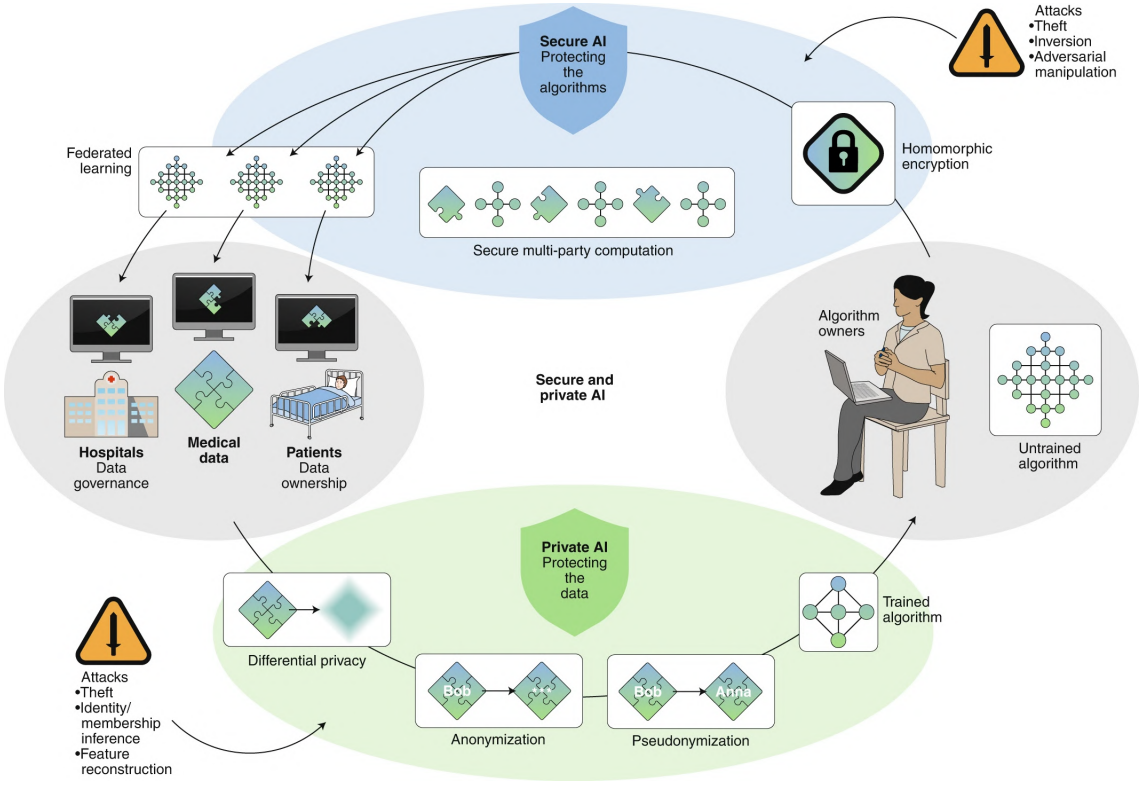


Σχήμα 6.17: *Σχηματική αναπαράσταση υπηρεσίας στο Cloud όπου εκτελείται προγνωστική ανάλυση για πρόβλεψη καρδιαγγειακών νοσημάτων πάνω σε κρυπτογραφημένα ιατρικά δεδομένα ασθενών με χρήση Homomorphic Encryption [245].*

Για την ασφαλή κοινοποίηση δεδομένων από πολλαπλά συνεργαζόμενα μέλη, μπορεί να αξιοποιηθεί η τεχνική του **Secure Multi-party Computation (MPC)** (βλ. 5.3.4), όπου τα δεδομένα είναι κρυπτογραφημένα με τέτοιο τρόπο ώστε κανένα μέλος ατομικά να μπορεί να δει τα πραγματικά δεδομένα, αλλά να μπορούν να χρησιμοποιηθούν συνολικά για την επίτευξη ενός σκοπού, όπως για παράδειγμα την εκπαίδευση κάποιου μοντέλου, απαιτώντας κοινή συναίνεση από όλα τα μέλη [222]. Για την εφαρμογή του MPC σε ML μοντέλα, έχουν αναπτυχθεί διάφορα μοντέλα, όπως το CrypTen framework [166], το οποίο προσφέρει τις κατάλληλες διεπαφές για εκπαίδευση μοντέλων ανεξάρτητα εργασίας (π.χ. ταξινόμηση εικόνων ή φωνής, αναγνώριση φωνής) με ασφαλή επικοινωνία και υπολογισμούς στα δεδομένα εκπαίδευσης. Η χρήση του MPC σε εφαρμογές υγείας, μπορεί να συνδυαστεί με την τεχνική του HE για την υλοποίηση υπολογισμών στα κρυπτογραφημένα δεδομένα, όπως παρουσιάζεται στο [247] όπου αναπτύσσεται ένα διαγνωστικό σύστημα το οποίο βρίσκει την ασθένεια που ταιριάζει περισσότερο με βάση τα ιατρικά δεδομένα υγείας ενός ασθενή, βασισμένο σε αυτές τις τεχνικές για την προστασία της ιδιωτικότητας αυτών των δεδομένων. Επίσης, η εφαρμογή του MPC έχει ερευνηθεί και για χρήση με ιατρικά δεδομένα χρηστών που καταγράφονται από ασύρματες υγειονομικές φορητές συσκευές, για την ασφαλή κοινοποίηση και επεξεργασία αυτών των δεδομένων [248].

Η τεχνική του **Federated Learning (FL)** (βλ. 5.3.6) μπορεί να χρησιμοποιηθεί για την κατανεμημένη εκπαίδευση μοντέλων σε τοπικούς κόμβους με σκοπό την εκπαίδευση ε-

νός κεντρικού μοντέλου, χωρίς την κοινοποίηση δεδομένων των χρηστών. Βέβαια, θα πρέπει να συνδυαστεί και με κάποια από τις παραπάνω μεθόδους για να προσφέρει την εγγυήσεις ιδιωτικότητας, όπως με προσθήκη θορύβου στα τοπικά δεδομένα μέσω DP ή με κρυπτογράφηση αυτών μέσω HE, καθώς τα δεδομένα στους τοπικούς κόμβους μπορεί να κλαπούν από κακόβουλους χρήστες ή ακόμα και να εξαχθούν από το κεντρικό μοντέλο με model inversion επιθέσεις, λόγο απομνημόνευσης των δεδομένων από τα μοντέλα, ακόμα και μόνο από διαμοιρασμό των βαρών [249]. Με την εφαρμογή DP στα τοπικά δεδομένα επιτυγχάνεται εγγύηση ιδιωτικότητας, ενώ με την εφαρμογή HE στα τοπικά δεδομένα επιτυγχάνεται ασφαλής απόκρυψη των δεδομένων, από το κεντρικό μοντέλο, αλλά και από επιτιθέμενους στους τοπικούς κόμβους, δημιουργώντας μια ασφαλή και ιδιωτική διαδικασία εκπαίδευσης (βλ. σχήμα 6.18).



Σχήμα 6.18: *Σχηματική αναπαράσταση όλων των αλληλεπιδράσεων μεταξύ δεδομένων, αλγορίθμων, και τεχνικών για την ασφάλεια και την ιδιωτικότητα της τεχνητής νοημοσύνης σε ιατρικές εφαρμογές [222].*

Δεδομένου ότι το FL είναι μια γενική μέθοδος μάθησης, η εφαρμογή στο πεδίο της υγείας έχει μελετηθεί σημαντικά σε όλο το εύρος εφαρμογών του, όπως για την πρόβλεψη καρδιακών νοσημάτων με χρήση των ηλεκτρονικών δεδομένα των ασθενών (EHR), εκπαιδεύοντας ένα δυαδικό SVM ταξινομητή [250]. Επίσης, έχει ερευνηθεί και η χρήση του σε εφαρμογές ιατρικών εικόνων, όπως στο [251] όπου 7 κλινικά ινστιτούτα συνεργατικά εκπαίδευσαν με χρήση FL, ένα μοντέλο για την ταξινόμηση της πυκνότητας του μαστού, με χρήση εικόνων μαστογραφίας, το οποίο αποδίδει κατά μέσο όρο 6.3% καλύτερα από αντίστοιχα μοντέλα εκπαιδευμένα μόνο με τοπικά δεδομένα ενός ινστιτούτου, και με σχετική βελτίωση 45.8% στη γενίκευση μετά από αξιολόγηση με δεδομένα από άλλα ινστιτούτα. Χρήση FL μαζί με κάποια

άλλη τεχνική για ενίσχυση της ιδιωτικότητας, έχει ερευνηθεί στο [252], όπου γίνεται χρήση του DP-SGD αλγορίθμου για εφαρμογή του DP σε εικόνες ιστοπαθολογίας δίνοντας αυστηρές εγγυήσεις ιδιωτικότητας (ϵ, δ) και χρησιμοποιώντας FL για κατανεμημένη εκπαίδευση ενός μοντέλου για ανίχνευση καρκίνου του πνεύμονα, το οποίο έχει παρόμοια απόδοση με αντίστοιχο μοντέλο εκπαιδευμένο κεντρικά, αποδεικνύοντας πειραματικά ότι είναι εφικτή η ανάλυση ιατρικών εικόνων χωρίς κοινή χρήση δεδομένων. Στο [253] παρουσιάζεται το Unified CT-COVID AI Diagnostic Initiative (UCADI), ένα διαγνωστικό ML μοντέλο για τον COVID-19, εκπαιδευμένο με FL από ακτινογραφίες θώρακα συγκεντρωμένες από 23 διαφορετικά νοσοκομεία, με χρήση HE τεχνικών για την ασφαλή μετάδοση των παραμέτρων του μοντέλου, δημιουργώντας ένα κεντρικό μοντέλο το οποίο παρουσιάζει καλύτερη επίδοση από κάθε άλλο μοντέλο εκπαιδευμένο μόνο με τα τοπικά δεδομένα του κάθε νοσοκομείου. Επίσης, στο [173], παρουσιάζεται το PriMIA (Privacy-preserving Medical Image Analysis), ένα ανοιχτού κώδικα προγραμματιστικό πλαίσιο εκπαίδευσης για εκπαίδευση ιδιωτικών μοντέλων με χρήση FL και DP σε δεδομένα ιατρικών εικόνων, το οποίο με εφαρμογή του σε ένα CNN μοντέλο για ταξινόμηση παιδιατρικών ακτινογραφιών θώρακος, εκπαιδεύει ένα μοντέλο με εγγυήσεις ιδιωτικότητας το οποίο αποδεικνύεται εμπειρικά ότι προστατεύει από model inversion επιθέσεις και έχει απόδοση εφάμιλλη με τοπικά, μη ασφαλώς εκπαιδευμένα μοντέλα.

Τεχνική	**Εφαρμογή**	**Σύστημα**	**Περιγραφή**	**Σχολιασμός**
DP	Medical Data	Decision Tree	[242]: Εφαρμογή DP σε δέντρα αποφάσεων για πρόβλεψη ζαχαρώδη διαβήτη	Εκπαίδευση δέντρων με Privacy Guarantees και 93% ακρίβεια
DP	X-Ray, CT	VGG-11	[243]: Εφαρμογή DP-SGD για ταξινόμηση παιδικής πνευμονίας	Framework για εκπαίδευση μοντέλων με Privacy Guarantees
HE	Medical Data	Predictive Analysis	[245]: Πρόβλεψη καρδιακών νοσημάτων στο Cloud	Κρυπτογράφηση ιατρικών δεδομένων
HE	Bioinformatics	*N/A*	[246]: Εφαρμογή HE σε γενετικά δεδομένα	Κρυπτογράφηση γενετικών δεδομένων
MPC, HE	Medical Data	Euclidean distance	[247]: Προσωπικές διαγνώσεις ασθενών με MPC, HE τεχνικές	Κρυπτογράφηση ιατρικών δεδομένων και συναίνεση νοσοκομείου, γιατρών, ασθενή
MPC	Medical Data	*N/A*	[248]: Ανάλυση και επεξεργασία ιατρικών δεδομένων από πολλαπλές πηγές	Ασφαλής, εμπιστευτική και κατανεμημένη ανάλυση δεδομένων
FL	EHR	Binary SVM	[250]: Πρόβλεψη καρδιακών νοσημάτων με FL σε SVM	Πιο γρήγορη σύγκλιση από κεντρικές μεθόδους
FL	Mammography	DenseNet-121	[251]: Εφαρμογή FL για ταξινόμηση μαστογραφιών	Καλύτερη 6.3% μ.ο. απόδοση και 45.8% βελτίωση γενίκευσης από κάθε κεντρικό μοντέλο
FL, DP	Histopathology	DenseNet	[252]: Εφαρμογή FL, DP-SGD για ανίχνευση καρκίνου του πνεύμονα	Privacy Guarantees και παρόμοια απόδοση με κεντρικό μη ιδιωτικό μοντέλο
FL, HE	X-Ray	3D CNN	UCADI [253]: Εφαρμογή FL, και HE στα βάρη, για ανίχνευση COVID-19	Καλύτερη απόδοση από κάθε κεντρικό μοντέλο
FL, DP	X-Ray	ResNet18	PriMIA[173]: Εφαρμογή FL, DP-SGD open-source ML framework για ταξινόμηση παιδικής πνευμονίας	Προστασία από model inversion, Παρόμοια απόδοση με κεντρικό μη ιδιωτικό μοντέλο

Πίνακας 6.8: *Πίνακας με τεχνικές προστασίας της ιδιωτικότητας σε εφαρμογές στο πεδίο της υγείας.*

6.2.4 Αξιολόγηση Αμυνών και Ευρωστίας Μοντέλων

6.2.4.1 Επίπτωση στην ακρίβεια

Όπως αναφέραμε και στις τεχνικές αμυνών για ML συστήματα υγείας (βλ. 6.2.3), η πιο μελετημένη άμυνα για δημιουργία εύρωστων διαγνωστικών μοντέλων ενάντια σε adversarial example είναι το adversarial training, το οποίο όμως δεν προσφέρει την ίδια προστασία από άγνωστες επιθέσεις με δείγματα στα οποία δεν έχει εκπαιδευτεί [215]. Επίσης, μπορεί να έχει επιπτώσεις στην ακρίβεια του μοντέλου σε καθαρά δείγματα, κάτι το οποίο ισχύει γενικότερα και για μοντέλα άλλων εφαρμογών, καθιστώντας την εκπαίδευση εύρωστων διαγνωστικών μοντέλων με υψηλή ακρίβεια ένα ανοιχτό πρόβλημα που παραμένει ακόμα [212]. Η χρήση FL τα τελευταία χρόνια, έχει επιτρέψει τη δημιουργία μοντέλων με υψηλότερη ακρίβεια, λόγω της δυνατότητας χρήσης περισσότερων δεδομένων, από ότι θα είχε διαθέσιμα ένας οργανισμός, ινστιτούτο ή κλινική [171].

Όμως τα εύρωστα μοντέλα δείχνουν να παρουσιάζουν και κάποια μη υπολογισμένα πλεονεκτήματα, όπως τη βελτίωση της γενίκευσης σε συνδυασμό με regularization τεχνικών, μειώνοντας το χάσμα μεταξύ ακρίβειας και ευρωστίας και κάνοντας τα μοντέλα να είναι πιο ανθεκτικά σε ανταγωνιστικές ή μη διαταραχές [215]. Πιο συγκεκριμένα, στο [254] χρησιμοποιώντας adversarial training, έδειξαν ότι το μοντέλο έχει βελτιωμένη γενίκευση σε δεδομένα εκτός κατανομής, το οποίο είναι πολύ σημαντικό σε εφαρμογές ανάλυσης ιατρικών εικόνων.

6.2.4.2 Ιδιωτικότητα δεδομένων

Η ιδιωτικότητα των ιατρικών δεδομένων είναι κρίσιμης σημασίας, όμως η εφαρμογή τεχνικών προστασίας της μπορεί να έχει κάποιες αρνητικές συνέπειες, όπως στην ακρίβεια του μοντέλου, στην ευκολία ή στον χρόνο εκπαίδευσης. Συγκεκριμένα, η εφαρμογή της DP στη διαδικασία της εκπαίδευσης είναι γνωστό ότι επιφέρει κόστος στη χρησιμότητα των αποτελεσμάτων, ανάλογα με το επιλεγμένο privacy budget ϵ, όπου όσο μεγαλύτερο, τόσο μεγαλύτερη εφαρμογή θορύβου και άρα εγγύηση ιδιωτικότητας, αλλά αντίστοιχα τόσο λιγότερα εύχρηστα αποτελέσματα. Η εφαρμογή DP σε ML διαγνωστικά μοντέλα, ειδικά ιατρικών εικόνων, είναι ακόμα νέο πεδίο έρευνας και παραμένει ανοιχτό το πρόβλημα εφαρμογής αυστηρών εγγυήσεων ιδιωτικότητας, ώστε να μπορεί να γίνει πρακτική εφαρμογή της τεχνικής αυτής [173].

6.2.4.3 Εφαρμογή αμυνών σε πραγματικές συνθήκες

Αν και έχουν γίνει πολλές εργασίες και δοκιμές σε adversarial examples σε φυσικές εικόνες, υπάρχει πολύ λιγότερη έρευνα στο πεδίο των ιατρικών εικόνων. Πολλοί ερευνητές θεωρούν ότι η εφαρμογή adversarial examples σε ιατρικές εικόνες, είναι πολύ δύσκολο να εφαρμοστεί σε πραγματικές συνθήκες, αν και υπάρχουν υποθετικά σενάρια κακόβουλων χρηστών οι οποίο θα μπορούσαν να δημιουργήσουν τέτοια δείγματα [215], όπως αναφέραμε και στην κατηγοριοποίηση των επιθέσεων στα ML συστήματα υγείας (βλ. 6.2.2.3). Αντίστοιχα οι προτεινόμενες άμυνες είναι ακόμα σε ερευνητικό στάδιο, χωρίς να προσφέρουν ακόμα υψηλά επίπεδα ακρίβειας και ευρωστίας σε σύνθετα και πολύπλοκα σύνολα δεδομένων [221], παρουσιάζοντας όμως ελπιδοφόρα αποτελέσματα για την επίτευξη πλήρως εύρωστων ML μοντέλων [212].

6.2.4.4 Δείκτες αξιολόγησης ευρωστίας

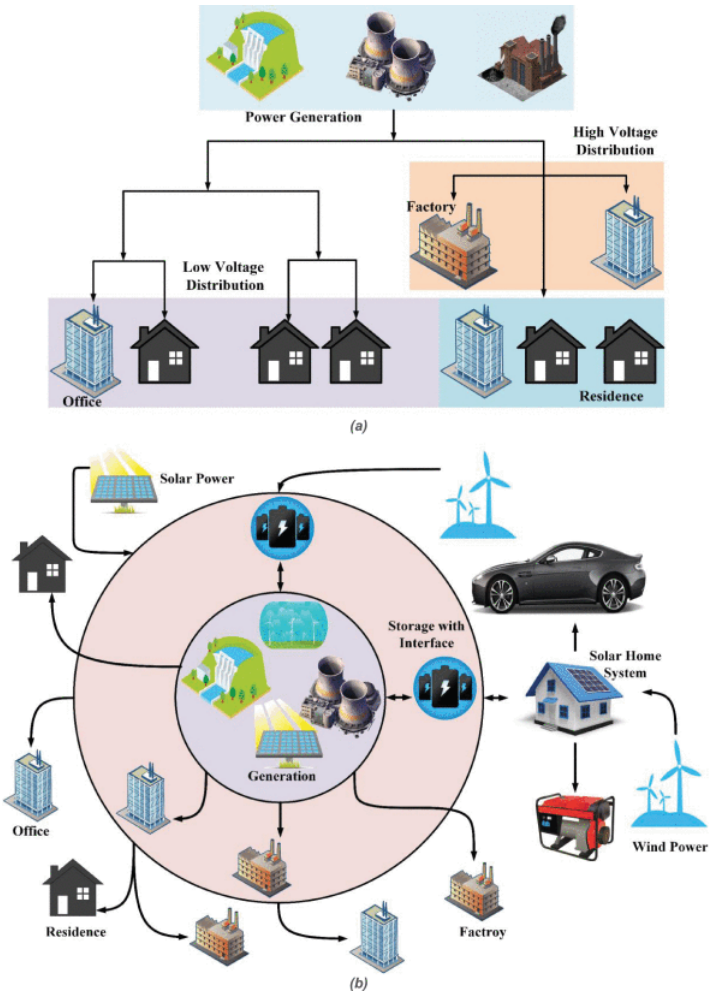
Όπως έχουμε αναφερθεί και στην ενότητα 4.4 υπάρχουν κάποιες γενικές μετρήσεις οι οποίες μπορούν να μας δώσουν μια εικόνα για την ευρωστία ενός μοντέλου. Για τα διαγνωστικά μοντέλα, οι μετρήσεις οι οποίες μπορούν να χρησιμοποιηθούν είναι: (i) **Empirical robustness** για να αξιολογηθεί η μικρότερη δυνατή διαταραχή το οποίο είναι ικανό να οδηγήσει σε σφάλμα το μοντέλο, και (ii) Adversarial Success Rate (**ASR**) για να αξιολογηθεί η επιτυχία των επίθεσης ή αντίστοιχα επιτυχία αποτροπής των επιθέσεων. Επίσης, για την αξιολόγηση της ιδιωτικότητας των δεδομένων, μπορεί να χρησιμοποιηθεί η τιμή του privacy budget (ϵ) ή (ϵ, δ) που έχει επιλεχθεί κατά την εκπαίδευση του μοντέλου, για τη σύγκριση της εγγύησης ιδιωτικότητας που προσφέρεται, με μικρότερες τιμές να μεταφράζονται σε μεγαλύτερο επίπεδο ασφάλεια ιδιωτικότητας [156].

6.3 Συστήματα Ηλεκτρικής Ενέργειας και Έξυπνα Δίκτυα

Τα **έξυπνα δίκτυα** (Smart Grids ή SG) είναι εκσυγχρονισμένα ηλεκτρικά δίκτυα που χρησιμοποιούν προηγμένες τεχνολογίες επικοινωνίας και πληροφοριών. Σε αντίθεση με τα παραδοσιακά δίκτυα, ενσωματώνουν αμφίδρομη επικοινωνία, αισθητήρες και αυτοματισμό για παρακολούθηση και έλεγχο της ροής ηλεκτρικής ενέργειας σε πραγματικό χρόνο. Αυτή η έξυπνη υποδομή βελτιώνει την αποδοτικότητα, την αξιοπιστία και τη βιωσιμότητα επιτρέποντας καλύτερη διαχείριση της ζήτησης ηλεκτρικής ενέργειας, ενσωμάτωση ανανεώσιμων πηγών ενέργειας και γρήγορη απόκριση σε διακοπές. Τα έξυπνα δίκτυα στοχεύουν στη δημιουργία ενός ανταποκρινόμενου και διασυνδεδεμένου ενεργειακού οικοσυστήματος, βελτιστοποιώντας τη χρήση των πόρων και βελτιώνοντας τη συνολική απόδοση του δικτύου [255].

Για να μπορούν να λειτουργήσουν τα SG, είναι επιτακτική η αντικατάσταση των παραδοσιακών ηλεκτρομηχανικών μετρητών κατανάλωσης ηλεκτρικής ενέργειας, με έξυπνους μετρητές (**smart meters**), οι οποίοι παρέχουν ταχύτερη αμφίδρομη επικοινωνία μεταξύ των παρόχων ενέργειας και των τελικών χρηστών, επιτρέποντας τον άμεσο έλεγχο φορτίου για πιο άμεση ανταπόκριση στη ζήτηση ρεύματος και για την εξοικονόμηση ενέργειας [256].

Για την ομαλή λειτουργία όλων των εμπλεκόμενων συστημάτων είναι αναγκαία η παρακολούθηση του δικτύου, ώστε π.χ. σε περιπτώσεις διακοπής να είναι όσο το δυνατόν πιο ακριβής η πρόβλεψη ζήτησης για τον ισοζυγισμό της παραγωγής. Το βασικότερο σύστημα για αυτήν τη διαδικασία είναι το Σύστημα Εποπτικού Ελέγχου και Συλλογής Πληροφοριών ή αλλιώς Supervisory Control and Data Acquisition system (**SCADA**), το οποίο εκτελεί παρακολούθηση σε πραγματικό χρόνο και είναι βαθιά συσχετισμένο με τους κρίσιμους τομείς της ενέργειας [257]. Το SCADA συλλέγει συνεχώς μετρήσεις από απομακρυσμένες τερματικές μονάδες ή αλλιώς Remote Terminal Units (**RTU**), οι οποίες χρησιμοποιούνται στη συνέχεια από τα συστήματα Εκτιμήσεως της Κατάστασης ή αλλιώς State Estimation (**SE**), για την εκτίμηση των μεταβλητών κατάστασης (state variables) της τρέχουσας τοπολογίας του συστήματος, π.χ. το διάνυσμα των μεγεθών και των γωνιών τάσης σε όλους τους διαύλους του δικτύου. Τα αποτελέσματα αυτά αξιοποιούνται περαιτέρω από το Σύστημα Διαχείρισης Ενέργειας ή αλλιώς Energy Management System (EMS) για την εκτέλεση διαφόρων βασικών λειτουργιών, όπως ανάλυση έκτακτης ανάγκης, βέλτιστη ροή ισχύος κ.λπ. [258].



Σχήμα 6.19: *Δίκτυα Ενέργειας: (α) Συμβατικό Δίκτυο (β) Smart Grid. Στο συμβατικό σύστημα η ισχύς ρέει από μία μόνο κατεύθυνση, αλλά για στο smart grid, δεν υπάρχει αυστηρή δομή. Η παραγωγή μπορεί να συμβεί και από την πλευρά του καταναλωτή, όπως από αιολικές ή ηλιακές πηγές. Η ροή ισχύος μπορεί επίσης να είναι αμφίδρομη όπως από τις αποθήκες ενέργειας και τα σπίτια των καταναλωτών [259].*

6.3.1 Εφαρμογές Τεχνητής Νοημοσύνης στο πεδίο

Τα τελευταία χρόνια, το ενεργειακό τοπίο έχει υποστεί μεγάλες αλλαγές λόγω των κατανεμημένων πόρων, των ανανεώσιμων πηγών ενέργειας, των σταθμών φόρτισης ηλεκτρικών αυτοκινήτων (EV), των συστημάτων αποθήκευσης ενέργειας και των συσκευών IoT (Internet of Things). Αυτά τα στοιχεία ενισχύουν το δίκτυο με νέα ευελιξία, επιτρέποντας υπηρεσίες ρύθμισης τάσης και διαχείρισης ροής ισχύος. Ωστόσο, αυτό απαιτεί την αξιοποίηση ανεξερεύνητων δεδομένων εντός του δικτύου διανομής για τη δημιουργία αποτελεσματικών αλγορίθμων βελτιστοποίησης. Αυτά τα δεδομένα μπορεί να προέρχονται από λειτουργικές πηγές του δικτύου όπως μετρητές, συσκευές ελέγχου ή μη λειτουργικές πηγές, όπως καιρικά μοτίβα, τάσεις της αγοράς ηλεκτρικής ενέργειας και πληροφορίες για τη συμπεριφορά πελατών.

Για την αποτελεσματική διαχείριση όλων αυτών των λειτουργιών έχει αρχίσει να γίνεται χρήση της τεχνητής νοημοσύνης για την πρόβλεψη κατανάλωσης και παραγωγής ενέργειας. Στα έξυπνα δίκτυα έχει γίνει καίριας σημασίας, για τη βελτιστοποίηση της διαχείρισης ενέργειας και τη βελτίωση της απόδοσης του δικτύου. Οι αλγόριθμοι τεχνητής νοημοσύνης

αναλύουν τεράστιες ποσότητες δεδομένων, συμπεριλαμβανομένων των ιστορικών προτύπων κατανάλωσης, των καιρικών συνθηκών και των διακυμάνσεων της ζήτησης σε πραγματικό χρόνο. Αυτό επιτρέπει ακριβείς προβλέψεις της χρήσης ενέργειας, διευκολύνοντας την προληπτική λήψη αποφάσεων για τους φορείς εκμετάλλευσης του δικτύου. Στα έξυπνα δίκτυα, η τεχνητή νοημοσύνη συμβάλλει στη δυναμική εξισορρόπηση φορτίου, στη βελτιστοποίηση απόκρισης της ζήτησης και στην προγνωστική συντήρηση. Τα μοντέλα μηχανικής μάθησης προσαρμόζονται συνεχώς στα εξελισσόμενα πρότυπα, επιτρέποντας πιο ανθεκτικά και προσαρμοστικά ενεργειακά συστήματα [260].

Πιο συγκεκριμένα κάποιες από τις εφαρμογές της τεχνητής νοημοσύνης σε αυτό το πεδίο είναι:

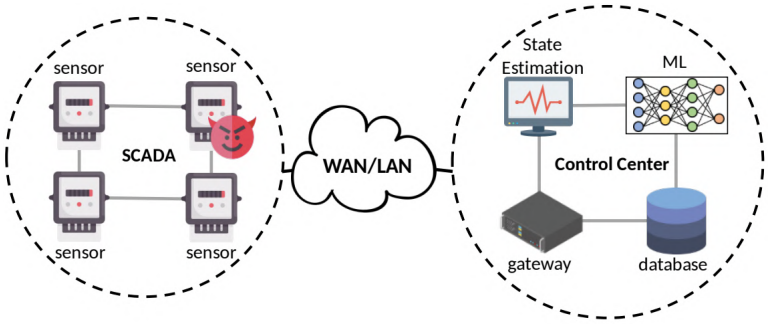
- Load Forecasting: Πρόβλεψη φορτίου (Βραχυπρόθεσμη/Μεσοπρόθεσμη/Μακροπρόθεσμη) [261]
- PV Power Forecasting: Πρόβλεψη παραγωγής ενέργειας Φωτοβολταϊκών [262]
- Wind Energy Forecasting: Πρόβλεψη παραγωγής ενέργειας Αιολικών πηγών [259]
- Stability Assessment: Αξιολόγηση ευστάθειας (Δικτύου ισχύος, Συχνότητας, Τάσεως) [260]
  - Power Quality Assessment: Αξιολόγηση ποιότητας ισχύος του δικτύου [263]
  - Voltage Stability Assessment: Αξιολόγηση ευστάθειας τάσης του δικτύου [264]
- Event Cause Analysis: Ανάλυσης αιτιών συμβάντων στα ηλεκτρικά δίκτυα [265]
- Fault Detection: Ανίχνευση σφαλμάτων σε δεδομένα του δικτύου [260]
  - FDIA Detection: Ανίχνευση εισαγωγής κακόβουλων δεδομένων στο δίκτυο [266]
  - Energy Theft Detection: Ανίχνευση κλοπής ρεύματος [267]
- Nonintrusive load monitoring (NILM): Μη παρεμβατική παρακολούθηση φορτίου [256]

Η πρόβλεψη φορτίου (**Load Forecasting**) παίζει πολύ σημαντικό ρόλο στη λειτουργία και προγραμματισμό των συστημάτων ισχύος. Χρησιμοποιώντας χαρακτηριστικά εισόδου, όπως ιστορικά φορτία και μετεωρολογικές προβλέψεις, οι διαχειριστές συστημάτων και οι επιχειρήσεις κοινής ωφέλειας δημιουργούν μοντέλα προβλέψεων για να καθοδηγούν τη λήψη αποφάσεων κατά τη δέσμευση και την αποστολή ενέργειας [268]. Η ακρίβεια αυτών τον προβλέψεων παίζει πολύ σημαντικό ρόλο, ειδικότερα των βραχυπρόθεσμων προβλέψεων (Short Term Load Forecasting) οι οποίες έχουν ορίζοντα ημερών ή βδομάδων και επηρεάζονται από πολλαπλούς οικονομικό κοινωνικούς και μετεωρολογικούς παράγοντες. Για την καλύτερη πρόβλεψη, έχουν αξιοποιηθεί πολλές στατιστικές και ML τεχνικές, όπως Support Vector Machines ή Neural Networks, αλλά και Deep learning τεχνικές που παρουσιάζουν από τις καλύτερες επιδόσεις στο πεδίο αυτό. Ειδικότερα, χρησιμοποιούνται απλά Feed-Forward Neural Networks (FNN), Recurrent Neural Networks (RNN), αλλά και Long Short-Term Memory (LSTM) networks, τα οποία είναι σχεδιασμένα να βελτιώσουν την απόδοση των RNN 'ξεχνώντας' παροδικές εξαρτήσεις, εστιάζοντας σε αυτές που είναι μακροχρόνιες αναγνωρίζοντας τα πραγματικά μοτίβα [269].

Η αξιολόγηση ευστάθειας (**Stability Assessment**) είναι μια πολύ σημαντική λειτουργία για τη δημιουργία αξιόπιστων δικτύων και τη σωστή αξιολόγηση της κατάστασης και λειτουργίας αυτών. Χωρίς ευστάθεια μπορεί να υπάρξουν καταστροφικές διακοπές ρεύματος που απειλούν τις περιουσίες και τις ζωές των ανθρώπων [264]. Οι αξιολογήσεις αυτές γίνονται για διάφορες συνιστώσες του δικτύου, όπως για την ισχύ, ρεύμα ή τάση. Τα τελευταία χρόνια έχει γίνει χρήση ML αλγορίθμων για αυτές τις εργασίες, και οι επιδόσεις αυτών είναι πολύ καλύτερες από παραδοσιακές μεθόδους και υπολογιστικά πιο αποδοτικοί [263]. Αντίστοιχα,

τα συστήματα ανάλυσης αιτιών σε συμβάντα του δικτύου (**Event Cause Analysis**), αξιοποιούν πλέον ML τεχνικές, εφόσον πλέον υπάρχουν πλέον πολλαπλές συσκευές και σημεία μέτρησης που προσφέρουν μεγάλο όγκο και ποικιλία δεδομένων, τα οποία χρησιμοποιούνται για την εκπαίδευσης τους [265].

Η ύπαρξη όμως τόσων μετρήσεων και δεδομένων για την εκτέλεση State Estimation, επιτρέπει σε κακόβουλους χρήστες να καταλάβουν (RTUs) ώστε να εισάγουν δεδομένα στο δίκτυο με σκοπό να επηρεάσουν της διαδικασίες πρόβλεψης, χωρίς να γίνουν αντιληπτοί, φέρνοντας το σύστημα σε λάθος κατάσταση και απειλώντας σοβαρά τη λειτουργία και αξιοπιστία των δικτύων. Αυτές οι επιθέσεις ονομάζονται **False Data Injection Attacks (FDIA)**, και υπάρχει διαρκής έρευνα για την αντιμετώπιση τους, για παράδειγμα με χρήση deep learning αλγορίθμων (όπως π.χ. CNNs, GANs [270]), οι οποίοι προσφέρουν από τις καλύτερες επιδόσεις και ακρίβεια ανίχνευσης [266]. Η βασική αρχιτεκτονική τέτοιων συστημάτων ανίχνευσης φαίνεται στο σχήμα 6.20. Στην κατηγορία των εσφαλμένων μετρήσεων, ανήκουν και οι ρευματοκλοπές, όπου κακόβουλοι χρήστες εισάγουν δεδομένα κατανάλωσης ρεύματος πολύ μικρότερα από τα πραγματικά, παίρνοντας υπό έλεγχο τους μετρητές ενέργειας (είτε smart είτε όχι), για τη μείωση του λογαριασμού ηλεκτρικού ρεύματος. Επίσης, αξιοποιούνται ML μοντέλα (π.χ. SVM, DNN, CNN, RNN) για την ανίχνευση τέτοιων φαινομένων (**Energy Theft Detection**), είτε (i) με βάση τις μετρήσεων των αισθητήρων είτε (ii) με βάση τα προφίλ κατανάλωσης, για την ανίχνευση περίεργων μετρήσεων ή συμπεριφορών [267].



Σχήμα 6.20: *Σχηματικό παράδειγμα λειτουργίας ενός DDN FDIA ανιχνευτή. Ο επιτιθέμενος καταλαμβάνει ένα μέρος των αισθητήρων για να εισάγει fault data. Το εκπαιδευμένο DNN βρίσκεται σε ένα κέντρο ελέγχου και ανιχνεύει κακόβουλες μετρήσεις [271].*

6.3.2 Επιθέσεις - Κενά Ασφαλείας

6.3.2.1 Εφαρμογές Επιθέσεων

Ωστόσο, η χρήση δεδομένων και AI για αυτές τις λειτουργίες αυξάνει και τον κίνδυνο κυβερνοεπιθέσεων και παραβιάσεων προσωπικών δεδομένων. Οι φορείς εκμετάλλευσης ενέργειας πρέπει πλέον να εστιάζουν στην ασφάλεια και την ευρωστία των AI μοντέλων στα συστήματά τους. Η χρήση ειδικότερα μοντέλων μηχανικής μάθησης (ML) σε έξυπνα δίκτυα, μπορεί να τα αφήσει ευάλωτα σε **ανταγωνιστικές επιθέσεις** οι οποίες είναι εφαρμόσιμες σε μοντέλα βαθιάς μάθησης. Αυτές οι επιθέσεις είναι εγγενώς συγκαλυμμένες και μπορεί να προκαλέσουν τυχαία ή στοχευμένα κακόβουλα αποτελέσματα, εφόσον αντικατασταθεί η είσοδος με adversarial examples [269] (βλ. σχήμα 6.21). Επίσης, παρά το ότι οι περισσότερες ανταγωνιστικές επιθέσεις είναι εστιασμένες σε εφαρμογές ταξινόμησης εικόνων, δεν είναι οι μοναδικές, και

μπορούν να μεταφερθούν εύκολα σε ML εφαρμογές στα ηλεκτρικά και έξυπνα δίκτυα. Σε αυτήν την ενότητα εστιάζουμε μόνο στις ανταγωνιστικές επιθέσεις σε ML συστήματα τα οποία χρησιμοποιούνται σε εφαρμογές ηλεκτρικών και έξυπνων δικτύων, και όχι σε όλο το φάσμα των πιθανών κυβερνοεπιθέσεων που μπορεί να επιτευχθούν.

Οι ανταγωνιστικές επιθέσεις μπορεί να στοχεύουν σε διάφορα υποσυστήματα των SG, όπως για παράδειγμα σε ML συστήματα τα οποία ανιχνεύουν επιθέσεις εισαγωγής εσφαλμένων δεδομένων (**false data injection adversarial attacks** ή **FDIA**), όπου επιτιθέμενοι με χρήση adversarial examples προσπαθούν να ξεγελάσουν τους ταξινομητές από το να ανιχνεύσουν εσφαλμένες μετρήσεις που έχουν σκοπό να διαταράξουν τους υπολογισμούς του δικτύου [257]. Αντίστοιχα, μπορεί να συμβούν και ανταγωνιστικές επιθέσεις ενάντια σε ML μοντέλα τα οποία ανιχνεύουν κλοπές ενέργειας και ρεύματος (**energy theft detection adversarial attacks**), όπου ένας επιτιθέμενος μπορεί να αναφέρει εξαιρετικά χαμηλές μετρήσεις κατανάλωσης, παρακάμπτοντας τον εντοπισμό από το ML σύστημα ανίχνευσης κλοπής [267].

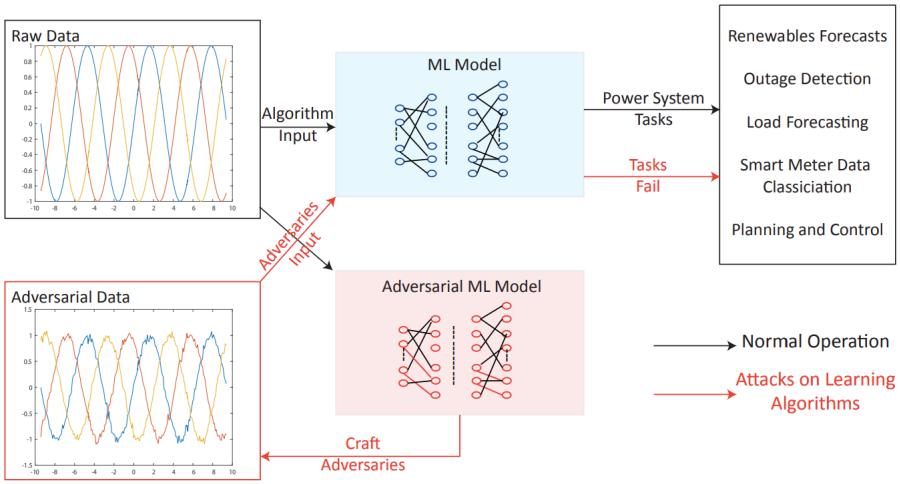
Άλλες ανταγωνιστικές επιθέσεις στοχεύουν στη δυσλειτουργία των ML μοντέλων που χρησιμοποιούνται για πρόβλεψη φορτίου (**load forecasting adversarial attacks**), όπου με κατάλληλη εισαγωγή perturbed δεδομένων, οι επιτιθέμενοι μπορούν να χειριστούν τις προβλέψεις για πρόκληση στοχευμένων ή μη ζημιών στη λειτουργία του συστήματος, χωρίς να υπάρχει απαραίτητα κάποια γνώση για το μοντέλο πρόβλεψης φορτίου ή το υποκείμενο σύστημα ισχύος [268]. Οι επιθέσεις μπορεί να στοχεύουν επίσης σε συστήματα αξιολόγησης του δικτύου, όπως αξιολόγησης της σταθερότητας της τάσης (**voltage stability attacks**) [263] ή της ποιότητας ισχύος (**power quality attacks**) [263].

Ακόμη, έχουν μελετηθεί επιθέσεις σε συστήματα ανάλυσης αιτιών συμβάντων (**event cause analysis frameworks attacks**) [265], στα οποία επιτιθέμενοι μπορούν να επιτεθούν με χειραγωγημένα δεδομένα για να τα παραπλανήσουν σχετικά με την αιτία ή την τοποθεσία των γεγονότων και κατά συνέπεια να προκαλέσουν τεχνικές ή οικονομικές ζημιές στο δίκτυο.

Μια άλλη εφαρμογή επιθέσεων έχει να κάνει σε συστήματα που παρακολουθούν τις καταναλώσεις τον νοικοκυριών, όπως ανταγωνιστικές επιθέσεις σε ML συστήματα ανίχνευσης πληρότητας μιας οικείας (**occupancy detection adversarial attacks**) με βάση της τιμές των μετρήσεων από έξυπνους smart meters, ώστε να διαπιστωθεί η ύπαρξη ανθρώπινης ενέργειας σε μια οικία και οι συνήθειες των καταναλωτών [256]. Μια άλλη παρόμοια εφαρμογή, έχει να κάνει σε συστήματα που εκτελούν μη παρεμβατική παρακολούθηση φορτίου (**non-intrusive load monitoring adversarial attacks**), τα οποία αναλύουν τις αλλαγές στην τάση και το ρεύμα που καταναλώνεται από ένα νοικοκυριό για να εξάγουν πληροφορίες για τις ηλεκτρικές συσκευές που χρησιμοποιούνται [272]. Αυτές οι επιθέσεις στοχεύουν στο να αποτρέψουν αυτήν την παρακολούθηση ή την ανίχνευση για τη διατάραξη της ακεραιότητας (integrity) των SG.

Τέλος, μια άλλη κατηγορία επιθέσεων που μπορεί να βρει εφαρμογή στα SG, είναι και οι επιθέσεις ιδιωτικότητας δεδομένων (**data privacy attacks**), καθώς τα δεδομένα των καταναλωτών, δηλαδή οι μετρήσεις κατανάλωσης ηλεκτρικού ρεύματος, είναι πολύ πιο εκτεθειμένα από ότι στα παραδοσιακά συστήματα ηλεκτρικής ενέργειας, και όλα τα μοτίβα και οι συμπεριφορές που είναι δυνατόν να ανιχνευθούν από τη χρήση ενέργειας ενός καταναλωτή, μπορεί να αξιοποιηθούν είτε για εμπορικούς είτε για κακόβουλους σκοπούς, δημιουργώντας παραβιάσεις

στην ιδιωτική ζωή και δεδομένα των καταναλωτών [260].



Σχήμα 6.21: *Σχηματικό παράδειγμα ανταγωνιστικών επιθέσεων σε συστήματα ηλεκτρικής ενέργειας στα οποία χρησιμοποιούνται ML μοντέλα για διάφορες εργασίες. Τα adversarial examples παράγονται από καθαρές τιμές του συστήματος ηλεκτρικής ενέργειας (τάση, ρεύμα, θερμοκρασία, κ.λπ.) που χρησιμοποιούνται ως είσοδοι για τα ML μοντέλα ανά περίπτωση [263].*

6.3.2.2 Κατηγορίες Επιθέσεων

Αφού έχουμε αναφέρει τα κενά ασφαλείας που υπάρχουν στα ML συστήματα στο πεδίο των SG και κάποιες εφαρμογές ανταγωνιστικών επιθέσεων, σε αυτήν την ενότητα γίνεται μια προσπάθεια ταξινόμησης αυτών με βάση την ταξινόμηση και το μοντέλο απειλής που έχουμε κάνει στην ενότητα 3.4.1 για τις ανταγωνιστικές επιθέσεις γενικότερα.

Αρχικά, όσο αναφορά την ικανότητα του επιτιθέμενου, όλες οι επιθέσεις που θα αναφερθούμε πρόκειται για **evasion attacks**, καθώς το πεδίο των poisoning attacks σε SG δεν έχει μελετηθεί αρκετά και δεν υπάρχουν αρκετές επιθέσεις στη βιβλιογραφία [269]. Επίσης, όσον αφορά τη γνώση του επιτιθέμενου, υπάρχουν και **White-box** και **Black-box** ανταγωνιστικές επιθέσεις.

Για τις υπάρχουσες evasion επιθέσεις μπορούμε να τις ταξινομήσουμε ανάλογα την εφαρμογή των ML μοντέλων στα SG και αντίστοιχα τον στόχο του επιτιθέμενου, δηλαδή το μέρος του συστήματος έχει σκοπό να βλάψει:

- **FDIA Detection**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος ανίχνευσης εσφαλμένων δεδομένων που εισάγονται στις μετρήσεις (FDIAD).
- **Energy Theft Detection**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος ανίχνευσης υποκλοπής ενέργειας και ρεύματος (ETD).
- **Load Forecasting**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος πρόβλεψης φορτίου (LF).
- **Voltage Stability Assessment**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος αξιολόγησης σταθερότητας της τάσης του δικτύου (VSA).
- **Power Quality Classification**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος ταξινόμησης της ποιότητας της ισχύος (PQC).
- **Event Cause Analysis**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήμα-

τος ανάλυσης αιτιών συμβάντων (ECA).

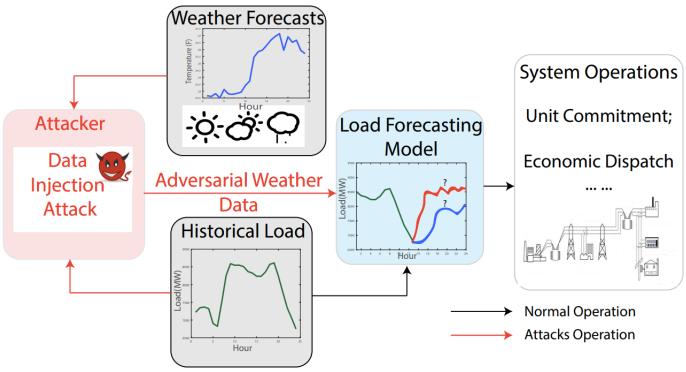
- **Occupancy Detection**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος ανίχνευσης πληρότητας [256].
- **Non-intrusive load monitoring**: Επιθέσεις που στόχο έχουν τη δυσλειτουργία του συστήματος μη παρεμβατικής παρακολούθησης φορτίου (NILM) [272].

Οι επιθέσεις που στόχο έχουν τα **FDIA Detection** μοντέλα, πρόκειται κυρίως για White-Box επιθέσεις σε Multilayer Perceptron (MLP), που στόχο έχουν να μειώσουν την ακρίβεια την ακρίβεια ανίχνευσης των deep learning αλγορίθμων που χρησιμοποιούνται για την ανίχνευση FDIAs. Παρά την επιτυχία των ML μοντέλων στην ανίχνευση τέτοιων επιθέσεων, αυτά παραμένουν ακόμα ευάλωτα σε ανταγωνιστικές επιθέσεις. Ένα τέτοιο παράδειγμα είναι στο [266], όπου το MLP ανιχνεύει FDIAs με ακρίβεια 99%, χρησιμοποιώντας ανταγωνιστικές επιθέσεις L-BFGS και JSMA, επιτυγχάνει μόλις 20% μετά από 15 επαναλήψεις και 10% αντίστοιχα, καθιστώντας το ML μοντέλο αναξιόπιστο. Στο [258], το MLP που χρησιμοποιείται για την ανίχνευση FDIAs έχει ακρίβεια 99% με απώλεια κοντά στο 0.25. Με εφαρμογή της Targeted Fast Gradient Sign Method (TFGSM) επίθεσης, η ακρίβειες του μοντέλου πέφτει δραματικά για διάφορες τιμές ϵ, π.χ. για $\epsilon = 0.1$ είναι κοντά στο 85%, για $\epsilon = 0.2$ είναι κοντά στο 40% και για $\epsilon = 0.3$ είναι κοντά στο 10%. Επίσης, στο [271], χρησιμοποιείται DNN για την ανίχνευση, το οποίο αξιολογείται με τη μέτρηση του recall, που αντιπροσωπεύει την πιθανότητα των ανταγωνιστικών επιθέσεων να ξεγελάσουν το μοντέλο. Με εφαρμογή επίθεσης τύπου BIM στο DNN, το recall παίρνει διάφορες τιμές ανάλογα το μέγεθος της ℓ_2-φραγμένης διαταραχής, έχοντας ελάχιστη τιμή 0.9% και μέγιστη τιμή 18.9%.

Για τις περιπτώσεις επιθέσεων που στόχο έχουν τα **Energy Theft Detection** μοντέλα, γίνεται γενικά η υπόθεση ότι ο επιτιθέμενος έχει πρόσβαση στον ενεργειακό μετρητή του, και μπορεί να πειράξει τις μετρήσεις που στέλνει στο υπόλοιπο δίκτυο. Οι επιθέσεις μπορεί ναι είναι είτε White-Box, ή Black-box επιθέσεις, αλλά οι δεύτερες είναι πιθανές σε πραγματικές συνθήκες, καθώς συνήθως δεν υπάρχει πρόσβαση ή γνώση για το ML μοντέλο. Μια αναπτυγμένη επίθεση που εκμεταλλεύεται την ευαλωτότητα των ML μοντέλων σε adversarial examples, για να στέλνει εξαιρετικά χαμηλές μετρήσεις κατανάλωσης ενέργειας αποφεύγοντας την ανίχνευση είναι η SearchFromFree [267], η οποία είναι gradient-based και είναι παρόμοια με τον DeepFool αλγόριθμο. Η αξιολόγηση της έγινε σε 3 είδη νευρωνικών δικτύων (DNN, RNN, CNN) πάνω σε σύνολα δεδομένων πραγματικών μετρήσεων από smart meters, με ℓ_1-φραγμένα adversarial examples για διάφορα μεγέθη, όπου και φαίνεται ότι μπορεί να μειωθεί η ακρίβεια ανίχνευσης των ML μοντέλων, σχεδόν στο 0% σε White-box σενάριο (από 86.9% στο DNN), αλλά ακόμα και στο 20% σε Black-box σενάριο (από 97.5% στο RNN).

Οι επιθέσεις που στόχο έχουν τα **Load Forecasting** μοντέλα, αναφέρονται συνήθως σε βραχυπρόθεσμης πρόβλεψης φορτίου (**STLF**), καθώς είναι πολύ πιο δύσκολο να επηρεαστούν οι πιο μακροπρόθεσμες προβλέψεις, λόγω και του μεγαλύτερου όγκου δεδομένων που απαιτούνται και αντίστοιχα της διάρκειας της επίθεσης. Οι επιθέσεις αυτές επιχειρούν να εισάγουν adversarial δεδομένα, τα οποία θα ξεγελάσουν τα ML μοντέλα load forecasting, ιδανικά σε Black-box σενάριο, για την πρόκληση ζημιών στα επιχειρησιακά συστήματα. Στο [268] προτείνεται μια Black-box επίθεση, η οποία εισάγει διαταραχές στις εισόδους θερμοκρασίας κάτω από συγκεκριμένα όρια για την αποφυγή ανίχνευσης (βλ. σχήμα 6.22). Με δοκιμές σε διαφο-

ρετικά ML μοντέλα (FNN, RNN, LSTM), σε White-box και Black-box σενάρια (με χρήση substitute model ή με ερωτήματα σε υπάρχον), με σχετικά μικρό βαθμό διαταραχών που εισάγονται στις τιμές των θερμοκρασιών, ο load forecasting αλγόριθμος, αποκλίνει δραστικά από τις αρχικές τιμές, με αποτέλεσμα είτε την αύξηση του λειτουργικού κόστους ή ακόμα και την απόρριψή φορτίου (load shedding). Πιο συγκεκριμένα, με μέγιστη απόκλιση 4 βαθμών Fahrenheit, και από τα δύο Black-box σενάρια, τα μοντέλα παρουσιάζουν μέγιστο σφάλμα πρόβλεψης έως και 13% (από περίπου 1.5%).



Σχήμα 6.22: *Σχηματικό παράδειγμα load forecasting ανταγωνιστικής επίθεσης. Χωρίς γνώση για τις παραμέτρους του αλγορίθμου πρόβλεψης, ο επιτιθέμενος εισάγει μικρές διαταραχές στις προβλέψεις του καιρού (θερμοκρασία) για την πρόκληση ζημιών [268].*

Αντίστοιχα, στις επιθέσεις με στόχο τα μοντέλα αξιολόγησης ποιότητας του δικτύου ή ανάλυσης συμπτωμάτων, γίνεται προσπάθεια επηρεασμού αυτών των συστημάτων αξιολόγησης, με προσθήκη adversarial examples από τις *τιμές τάσης* κατευθείαν στο σύστημα. Συγκεκριμένα στο [263] γίνεται παραγωγή adversarial examples με τεχνική παρόμοια με Fast Gradient Sign (FGS) επίθεση, όπου προστίθενται διαταραχές στις τιμές του σήματος τάσης του συστήματος, οι οποίες είναι αρκετά μικρές και κοντά στις αρχικές, για την αποφυγή ανίχνευσης. Η επίθεση εφαρμόζεται σε FNN μοντέλο σε Black-box σενάριο, και η ακρίβεια του **Power Quality Classifier**, πέφτει στο 67.5% (από 97.5% αρχικά) με $\gamma = 40\%$ αλλαγές στις εισόδους του σήματος εισόδου, και διαταραχές $\epsilon = 0.1$. Ακόμη, στο [264], υλοποιούνται και δοκιμάζονται δημοφιλείς επιθέσεις (FGSM, PGD, DeepFool, C&W, Universal Perturbations/Networks) σε White-box και Black-box σενάρια, σε CNN μοντέλα υπεύθυνα για **Voltage Stability Assessment**, όπου προστίθενται διαταραχές στην τιμή της bus voltage που στέλνεται ως είσοδος στο μοντέλο. Όλες οι επιθέσεις, χρειάζονται πρόσβαση ανάγνωσης των καθαρές μετρήσεων της τάσης (εκτός των universal), και αντίστοιχα δυνατότητα εγγραφής στο σύστημα μετάδοσης των τιμών bus voltage. Το CNN έχει ακρίβεια επικύρωσης σταθερότητας 99.5%, η οποία στην περίπτωση των input-specific White-box επιθέσεων πέφτει έως και 57.6% με $l = 20$ bus voltage σημεία μετρήσεων υπό επίθεση (PGD attack) και έως 15.5% με $l = 39$ (C&W). Στο Black-box σενάριο η πτώση ακρίβειας είναι πολύ μικρότερη, με χαμηλότερη τιμή 46.1% με $l = 35$ (PGD). Στην περίπτωση των universal επιθέσεων, έχουν αντίστοιχες επιδόσεις και στα δύο σενάρια, με πτώση ακρίβειας έως και 49.5% με $l = 20$. Στο [265] δοκιμάζονται επιθέσεις (FGSM, JSMA) σε CNN μοντέλο το οποίο επιτελεί **Event Cause Analysis**, αλλάζοντας τις τιμές τάσης σε White-box και Black-box (με χρήση substitute model) σενάρια. Για τις White-box επιθέσεις και ℓ_∞-φραγμένες διαταραχές, με $\epsilon = 0.07$

επιτυγχάνεται 76% ASR στο αρχικό μοντέλο με ακρίβεια 99.5%, με μεγαλύτερες τιμές ϵ να αυξάνουν την επιτυχία επίθεσης, ενώ αντίστοιχα για τις Black-box επιθέσεις, με $\epsilon = 0.07$ υπάρχει σχεδόν 80% transferability στο αρχικό μοντέλο από το substitue ακρίβειας 94%.

Στόχος	**Γνώση**	**Μοντέλα**	**Περιγραφή**	**Αποτελέσματα**
FDIAD	White-Box	MLP	[266]: L-BFGS, JSMA επιθέσεις σε FDIA Detection μοντέλα	L-BFGS: 20% accuracy, JSMA: 10% accuracy
FDIAD	White-Box	MLP	[258]: TFGSM επίθεση σε FDIA Detection μοντέλα	Πτώση accuracy: 85% για $\epsilon = 0.1$, 40% για $\epsilon = 0.2$, 10% για $\epsilon = 0.3$
FDIAD	White-Box	DNN	[271]: BIM επίθεση σε FDIA Detection μοντέλα	Min recall 0.9%, Max recall 18.9%
ETD	White-Box, Black-Box	DNN, RNN, CNN	[267]: SearchFromFree επίθεση για αποφυγή ανίχνευσης ρευματοκλοπής	White-box: 0% accuracy από 86.9% στο DNN, Black-box: 20% από 97.5% στο RNN
LF	Black-Box	FNN, RNN, CNN	[268]: Εισαγωγή perturbation στις τιμές θερμοκρασίας	Max prediction error: 13% από 1.5% με μέγιστη απόκλιση 4 βαθμών Fahrenheit
PQC	Black-Box	FNN	[263]: FGS για εισαγωγή perturbations στις τιμές τάσης	Πτώση accuracy: 67.5% από 97.5% αρχικά με $\gamma = 40\%$, $\epsilon = 0.1$
VSA	White-Box, Black-Box	CNN	[264]: FGSM, PGD, DeepFool, C&W, Universal για εισαγωγή perturbations στις τιμές τάσης	White-box: accuracy 57.6% με $l = 20$ (PGD attack), 15.5% με $l = 39$ (C&W), Black-box: accuracy 46.1% με $l = 35$ (PGD), 49.5% με $l = 20$ (Universal)
ECA	White-Box, Black-Box	CNN	[265]: FGSM, JSMA για εισαγωγή perturbations στις τιμές τάσης	White-box: 76% ASR με ℓ_∞, $\epsilon = 0.07$ (original accuracy 99.5%), Black-box: 80% transferability με $\epsilon = 0.07$ (substitue accuracy 94%)

Πίνακας 6.9: *Πίνακας με adversarial* ***Evasion*** *επιθέσεις σε παραδοσιακά και smart δίκτυα ηλεκτρικής ενέργειας.*

6.3.3 Άμυνες - Μέτρα Ασφαλείας

Για την αντιμετώπιση όλων των παραπάνω ανταγωνιστικών επιθέσεων, υπάρχει μια πληθώρα τεχνικών για την αμυντική θωράκιση των ML συστημάτων ενάντια σε τέτοιες επιθέσεις στα δίκτυα ηλεκτρικής ενέργειας. Πολλές από αυτές τις άμυνες είναι προσαρμοσμένες για αντιμετώπιση συγκεκριμένων επιθέσεων ενώ άλλες προσφέρουν ασφάλεια από πολλαπλά είδη επιθέσεων. Η κύρια λειτουργία των περισσοτέρων αμυντικών μηχανισμών είναι η δημιουργία εύρωστων μοντέλων, χωρίς όμως πρόσθετη επιβάρυνση στην κανονική απόδοση του μοντέλου.

Όπως αναφερθήκαμε, οι περισσότερες υπάρχουσες άμυνες για ανταγωνιστικές επιθέσεις επικεντρώνονται στις εφαρμογές ταξινόμησης εικόνων, το οποίο δεν έχει άμεση προσαρμογή στις εργασίες των ML μοντέλων που εκτελούνται στο πεδίο της ενέργειας, όμως πολλές από αυτές προσαρμόζονται εύκολα ή ακόμα και με μηδενικές τροποποιήσεις. Χρειάζεται όμως προσοχή και έρευνα για τις καταλληλότερες άμυνες, καθώς ενώ μπορεί να οδηγήσουν σε ποιο εύρωστα μοντέλα, μπορεί να έχουν ανεπιθύμητες παρενέργειες όπως μείωση της ακρίβειας τους απόδοσης.

6.3.3.1 Κατηγορίες Αμυνών

Σε αυτήν την ενότητα κάνουμε αναφορά στις τεχνικές αμυνών που μπορούν να χρησιμοποιηθούν και έχουν ερευνηθεί για να αναπτυχθούν πιο εύρωστα ML μοντέλα στο πεδίο της

ενέργειας, ενάντια σε ανταγωνιστικές επιθέσεις, και τις ταξινομούμε με βάση και την κατηγοριοποίηση στην ενότητα 3.6.1 για τις άμυνες ενάντια σε ανταγωνιστικές επιθέσεις γενικότερα.

Αρχικά, ανάλογα τώρα με την προσέγγιση αντιμετώπισης, μπορούμε να κατηγοριοποιήσουμε τις άμυνες στα συστήματα ηλεκτρικής ενέργειας και SG, όπως και στη γενική περίπτωση σε:

- **Data preprocessing**: Άμυνες που τροποποιούν τα δεδομένα εκπαίδευσης ή δοκιμής και των χαρακτηριστικών τους, για την ελαχιστοποίηση ή την αποφυγή των adversarial examples.
- **Model hardening**: Άμυνες που τροποποιούν κυρίως τα χαρακτηριστικά του μοντέλου, εκπαιδεύοντας το με σκοπό την ευρωστία ενάντια σε adversarial examples.
- **Auxiliary models**: Άμυνες που χρησιμοποιούν επιπλέον ML (ή και άλλα είδη) μοντέλα που κάνουν εξειδικευμένες εργασίες, για να ενισχύσουν την ευρωστία του κύριου μοντέλου.

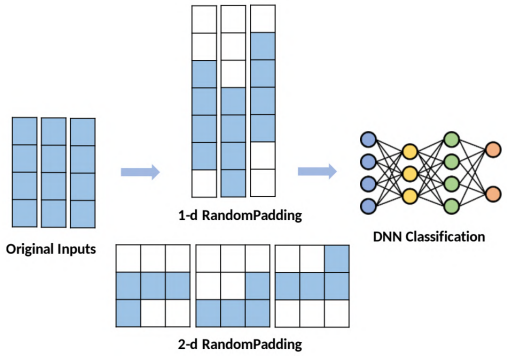
6.3.3.2 Τεχνικές Αμυνών

Στις *Data preprocessing* άμυνες κατατάσσονται αρκετές τεχνικές όπως: **Defensive Distillation** για την αντιμετώπιση FDIA Detection επιθέσεων, ή χρήση **APE-GAN** για την αντιμετώπιση VSA επιθέσεων [264]. Στις *Model hardening* άμυνες κατατάσσονται αρκετές τεχνικές όπως: **Adversarial Training** για την αντιμετώπιση FDIA Detection επιθέσεων [271], για την αντιμετώπιση VSA επιθέσεων [264] και για την αντιμετώπιση ECA επιθέσεων [265] ή **Gradient masking** για την παραγωγή μοντέλων με πιο ομαλές κλίσεις για παρεμπόδιση των ανταγωνιστικών επιθέσεων που βασίζονται στη βελτιστοποίηση [269]. Στις άμυνες που κάνουν χρήση *Auxiliary models* ή γενικότερα τρίτων συστημάτων για την ισχυροποίηση του μοντέλου κατατάσσονται αρκετές τεχνικές όπως: **Adversarial Detection** για την αντιμετώπιση FDIA Detection επιθέσεων [271], **Classifier Ensembles** για την ανίχνευση adversarial examples ενάντια σε ETD μοντέλα [273], και **Attack detection** τεχνικές για την ανίχνευση πιθανών επιθέσεων που μπορεί να εκτελούνται σε ένα SG [260].

Πέρα από την παραπάνω γενική ταξινόμηση βασικών τεχνικών άμυνας που μπορούν να εφαρμοστούν στα περισσότερα νευρωνικά δίκτυα υπάρχουν και τεχνικές ενίσχυσης της ευρωστίας των μοντέλων, για συγκεκριμένες επιθέσεις που στοχεύουν στη δυσλειτουργία κάποιου συστήματος του δικτύου, όπως κατηγοριοποιήσαμε στην ενότητα 6.3.2.2. Οπότε παρουσιάζουμε ενδεικτικά κατηγορίες και παραδείγματα αμυνών, ανάλογα και την εφαρμογή των ML μοντέλων στα SG και αντίστοιχα τον στόχο του επιτιθέμενου.

Πιο συγκεκριμένα, για την αντιμετώπιση **FDIA Detection** επιθέσεων, στο [271] ερευνώνται τρεις κλασσικές ανταγωνιστικές άμυνες για την αποτελεσματικότητα τους. Η *Defensive Distillation* (βλ. 3.6.2.3) μπορεί να εφαρμοστεί, όμως η αποτελεσματικότητα του είναι περιορισμένη, καθώς το νέο μοντέλο είναι λιγότερο ευαίσθητο σε διαταραχές εισόδου, στις FDIA οι διαταραχές δε χρειάζεται να είναι μικρές, όπως στην περίπτωση των εικόνων (για να είναι μην είναι διακριτά με το ανθρώπινο μάτι). Το *Adversarial Training* (βλ. 3.3) μπορεί να εφαρμοστεί, όμως η αποτελεσματικότητα της είναι επίσης περιορισμένη, καθώς δεν κλιμακώνεται σε συστήματα με πάρα πολλά δεδομένα, λόγο του χρόνου και το κόστους εκπαίδευσης που χρειάζεται (65x χρόνος εκπαίδευσης από την κανονική από τις δοκιμές). Όμως στο [274] η εφαρμογή *Adversarial Training*, ενισχύει σημαντικά την ευρωστία των DNN μοντέλων ε-

νάντια σε FGSM επιθέσεις, με 98.4% test accuracy και μέση τιμή διαταραχών 0.14. Τέλος, η *Adversarial Detection* (βλ. 3.6.2.11), έχει επίσης περιορισμένη αποτελεσματικότητα, καθώς υποθέτει ότι τα adversarial examples ακολουθούν διαφορετική κατανομή από τις κανονικές εισόδους, μια υπόθεση που εφαρμόζεται στις εφαρμογές της όρασης υπολογιστών, αλλά όχι στο FDIA Detection, οπότε και αυτή η τεχνική δεν μπορεί να διαχωρίσει μεταξύ κανονικών και ανταγωνιστικών μετρήσεων. Για τον καλύτερο μετριασμό των adversarial FDIA Detection επιθέσεων, προτείνεται το *Random Input Padding Framework*, το οποίο απαιτεί εκπαίδευση ενός νέου μοντέλου (Model Hardening τεχνική) προσθέτοντας padding μηδενικών τιμών στις αρχικές εισόδους με τυχαίο τρόπο (βλ. σχήμα 6.23). Σε δοκιμές το νέο αυτό αμυντικό σχήμα παρουσιάζει 96% ακρίβεια και 95% recall, σε ℓ_2-φραγμένες διαταραχές, αυξάνοντας κατά πολύ την ευρωστία του DNN μοντέλου.



Σχήμα 6.23: *Σχηματικό διάγραμμα της τεχνικής Random Input Padding Framework για προστασία από επιθέσεις που ξεφεύγουν από ML μοντέλα που εκτελούν FDIA Detection [256].*

Για αντιμετώπιση **Energy Theft Detection** επιθέσεων, στο [273] παρουσιάζεται ένας *Ensemble Detector* ο οποίος αποτελείται από auto-encoders με attention (AEA), gated recurrent units (GRUs) και feed forward neural networks (FNN) για τη δημιουργία ενός ανιχνευτή ο οποίος προστατεύει από επιθέσεις δηλητηρίασης (poisoning) των δεδομένων εκπαίδευσης με σκοπό τη ρευματοκλοπή. Διαφέρει από τις άλλες άμυνες, καθώς εστιάζει στην προστασία από adversarial τιμές που θα βρεθούν στα δεδομένα εκπαίδευσης των ανιχνευτών για ETD, οι οποίες θα παρθούν ιστορικά από τις μετρήσεις των smart meters των καταναλωτών. Οπότε αν τα δεδομένα είναι ήδη δηλητηριασμένα από ρευματοκλοπές, ο ανιχνευτής δε θα μπορέσει να εκπαιδευτεί και αξιολογηθεί σωστά. Ο enemble detector είναι ικανός να ανιχνεύσει ρευματοκλοπή με υψηλό detection rate (DR) 95.2% και false alarms (FA) μόλις 2.9%, με μεταβολή της απόδοσής του κατά 1-3%, σε περίπτωση poisoning attacks.

Για αντιμετώπιση **Voltage Stability Assessment** επιθέσεων, στο [264] εξετάζονται δύο κλασσικές ανταγωνιστικές άμυνες για την αποτελεσματικότητα τους. Χρησιμοποιείται *Adversarial Training*, με FGSM και PGD αλγορίθμους για την παραγωγή των adversarial examples, και εξετάζεται σε FGSM, PGD, DeepFool, C&W, Universal Perturbations/Networks επιθέσεις σε White-box και Black-box σενάρια, όπου ειδικότερη η PGD adv training εμφανίζεται αποτελεσματική (ακρίβεια πάνω από 80%) σε όλες τις περιπτώσεις, όταν ο αριθμός των σημείων bus voltage υπό επίθεση είναι $l \leq 30$. Χρησιμοποιείται επίσης το *APE-GAN* δίκτυο για άμυνα (βλ. 3.6.2.9), στα ίδια σενάρια επιθέσεων, η οποία όμως είναι αποτελεσματική μόνο στα White-box σενάρια των FGSM, PGD, C&W επιθέσεων ακόμα και όταν

όλα τα σημεία είναι υπό επίθεση ($l = 39$), καθώς είναι σχεδιασμένη για την προστασία από adversarial perturbations φτιαγμένα σε White-box σενάριο.

Για αντιμετώπιση **Event Cause Analysis** επιθέσεων, στο [265] εξετάζεται το *Adversarial Training* ως τεχνική για την δημιουργία ενός εύρωστου CNN ταξινομητή, ενάντια σε adversarial δείγματα τάσης. Το εκπαιδευμένο ανταγωνιστικά μοντέλο είναι πολύ αποτελεσματικό, έχοντας μόνο 5% ASR ακόμα και σε διαταραχές με $\epsilon = 0.3$, το οποίο είναι αρκετά μεγάλο για να μπορεί το δείγμα να ανιχνευθεί και από bad data detectors.

6.3.3.3 Προστασία Ιδιωτικότητας

Για την προστασία των SG και των συστημάτων ηλεκτρικής ενέργειας, μπορούν να αξιοποιηθούν και γενικότεροι μέθοδοι προστασίας των δεδομένων, όπως αυτούς που παρουσιάσαμε στην ενότητα 5.3.2, για την προστασία των δεδομένων ενός ML μοντέλου. Τέτοιες μέθοδοι, προσφέρουν πρώτον προστασία της ιδιωτικότητας (privacy) το οποίο αποτελεί ένα αρκετά σημαντικό κομμάτι στη γενικότερη ασφάλεια των συστημάτων και στη δημιουργία εύρωστων ML μοντέλων, αλλά μπορούν να προστατέψουν και από άλλες επιθέσεις στα συστήματα που παρουσιάσαμε στις προηγούμενες ενότητες, όπως στα FDIA detection συστήματα [270].

Αρχικά, τα δεδομένα ενεργειακών καταναλώσεων που συλλέγονται από τις διάφορες συσκευές ενός SG, μπορεί να υποκλαπούν από επιτιθέμενους ή κακόβουλες χρήστες μέσα στο δίκτυο. Για παράδειγμα, η ανίχνευση επιθέσεων εσφαλμένων δεδομένων, χωρίς την πρόσβαση σε προσωπικά δεδομένα κατανάλωσης είναι ένα ανοιχτό πεδίο έρευνας [257]. Μια πρόταση είναι η χρήση **Homomorphic Encryption** (βλ. 5.3.5), για την επίτευξη εμπιστευτικότητας, ακεραιότητας και προστασίας της ιδιωτικότητας των δεδομένων, σε CNN μοντέλα που χρησιμοποιούνται για την ανίχνευση μη φυσιολογικής συμπεριφοράς των δεδομένων μέτρησης από μακροπρόθεσμες παρατηρήσεις, κυρίως για ανίχνευση ρευματοκλοπής (ETD) [275]. Ο συγκεκριμένος ανιχνευτής επιτυγχάνει ακρίβεια ανίχνευσης μη φυσιολογικών συμπεριφορών έως 92.67%, εξασφαλίζοντας παράλληλα την ιδιωτικότητα των δεδομένων.

Μια ακόμα μέθοδος που χρησιμοποιείται για την προστασία της ιδιωτικότητας των δεδομένων, αλλά και για την κλιμάκωση των υπολογισμών του δικτύου είναι το **Federated Learning**. Η χρήση FL τεχνικών, επιλύει το πρόβλημα της ιδιωτικότητας, μη επιτρέποντας τον διαμοιρασμό δεδομένων, και το πρόβλημα της κλιμάκωσης, κάνοντας τους υπολογισμούς και την ανανέωση του μοντέλου τοπικά, χωρίς να εξαρτάται από ένα κεντρικό μοντέλο για όλα αυτά [257]. Ακόμη, στο [270] προτείνεται ένας **FL-based FDIA detector**, ο οποίος συνδυάζει FL με έναν transformer για να επιτελέσει ανίχνευση fault data injection επιθέσεων. Το σύστημα αυτό χρησιμοποιεί επίσης homomorphic encryption για να προστατέψει τα βάρη του νευρωνικού από υποκλοπές κατά τη μεταφορά τους στο κεντρικό (global) μοντέλο. Από τα πειραματικά αποτελέσματα, η άμυνα αυτή ξεπερνάει τις κλασσικές deep learning μεθόδους ανίχνευσης, και προσφέρει το πλεονέκτημα της προστασίας της ιδιωτικότητας των δεδομένων.

6.3.4 Αξιολόγηση Αμυνών και Ευρωστίας Μοντέλων

6.3.4.1 Επίπτωση στην ακρίβεια και χρόνο εκπαίδευσης

Η εκπαίδευση μοντέλων με δεδομένα που προκύπτουν από τα SG απαιτούν μεγάλα σύνολα δεδομένων, οπότε και αρκετό χρόνο εκπαίδευσης. Τεχνικές όπως το adversarial training, προσθέτουν επιπλέον χρόνο και υπολογιστικό κόστος [271], οπότε θα πρέπει να συνυπολο-

Στόχος	Προσέγγιση	Άμυνα	Περιγραφή	Αποτελέσματα
FDIAD	Model Hardening	Adversarial Training	[274]: Adversarial Training ενάντια FGSM για FDIA Detection μοντέλα	98.4% test accuracy και μέση τιμή perturbation 0.14 σε FGSM
FDIAD	Model Hardening	Random Input Padding Framework	[271]: Επανεκπαίδευση μοντέλου με προσθήκη zero padding στις αρχικές εισόδους με τυχαίο τρόπο	96% accuracy και 95% recall, σε ℓ_2
ETD	Auxiliary Models	Ensemble Detector	[273]: Ανιχνευτής που αποτελείται από auto-encoders με AEA, GRUs, FNN για προστασία poisoning attacks ρευματοκλοπής	95.2% DR και FA 2.9%, με 1-3% μεταβολή απόδοσής σε poisoning attacks
VSA	Model Hardening	Adversarial Training	[264]: Adversarial Training ενάντια σε FGSM, PGD, DeepFool, C&W, Universal perturbations στις τιμές τάσης	Πάνω από 80% accuracy για $l = 20$ (με PGD adv training) σε όλα τα σενάρια
VSA	Data Preprocessing	APE-GAN	[264]: Χρήση APE-GAN ενάντια σε FGSM, PGD, DeepFool, C&W, Universal perturbations στις τιμές τάσης	Πάνω από 80% accuracy για $l = 39$ μόνο στο White-box σενάριο σε FGSM, PGD, C&W
ECA	Model Hardening	Adversarial Training	[265]: Adversarial Training για εκπαίδευση robust CNN classifier	5% ASR με $\epsilon = 0.3$
ETD	Privacy	Homomorphic Encryption	[275]: CNN-based detector με χρήση HE για ασφαλή ανίχνευση ρευματοκλοπής	92.67% detection rate και data privacy
FDIAD	Privacy	Federated Learning	[270]: FL-based FDIA detector με χρήση Transformer για ασφαλές FDIA Detection	Καλύτερες επιδόσεις από υπάρχων ανιχνευτές βασισμένους σε CNN, LSTM μοντέλα

Πίνακας 6.10: *Πίνακας με άμυνες ενάντια σε ανταγωνιστικές επιθέσεις σε παραδοσιακά και smart δίκτυα ηλεκτρικής ενέργειας.*

γίζεται αυτό το κόστος όταν πρέπει τα μοντέλα να είναι εύρωστα.

Σε όλες τις περιπτώσεις που εξετάσαμε, δεν ασχοληθήκαμε με την πτώση της ακρίβειας στα κανονικά (clean) δεδομένα, καθώς τα μοντέλα με δεδομένα μετρήσεων έχουν ήδη πολύ καλές ακρίβειες (τάξεως του 90%) και γιατί η πτώση δεν είναι τόσο μεγάλη όσο στις περιπτώσεις του τομέα των εικόνων.

6.3.4.2 Δείκτες αξιολόγησης ευρωστίας

Στην ενότητα 4.4 έχουμε αναφερθεί σε κάποιες γενικές μετρήσεις οι οποίες μπορούν να μας δώσουν μια εικόνα για την ευρωστία ενός μοντέλου. Για τα μοντέλα στα συστήματα ηλεκτρικής ενέργειας, δε χρησιμοποιούνται ιδιαίτερα αυτές οι μετρήσεις, αλλά αξιολογούνται οι επιδόσεις των μοντέλων, κάτω από συνθήκες μεταβλητών πηγών επιθέσεις. Δηλαδή, αξιολογούνται οι τιμές Precision, Recall, ή F1-score για διάφορες τιμές συστημάτων (buses) τα οποία είναι υπό επίθεση και παράγουν ανταγωνιστικές μετρήσεις, οι οποίες θα πρέπει να είναι κοντά στις αποδόσεις του αρχικού μοντέλου [270]. Επίσης, όπως ήδη αναφερθήκαμε στην περίπτωση των VSA ανιχνευτών, οποιοδήποτε μοντέλο παρουσιάζει ακρίβεια ανίχνευσης (detection accuracy) κάτω από 80%, θεωρείται μη αποτελεσματικό [264].

Επίσης, στις περιπτώσεις των μοντέλων Load Forecasting, μπορεί να χρησιμοποιηθεί η Mean Absolute Percentage Error (**MAPE**) μέτρηση, για την αξιολόγηση του σφάλματος πρόβλεψης και της απόκλισης χαρακτηριστικών εισόδου που προκαλείται από την προσθήκη adversarial perturbations ταυτόχρονα [263].

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία, παρουσιάστηκαν αναλυτικά οι κίνδυνοι από ανταγωνιστικές επιθέσεις και επιθέσεις ιδιωτικότητας σε ML μοντέλα, εστιάζοντας σε χρήσεις στους 3 κρίσιμους τομείς των αυτοκινήτων, της υγείας και της ενέργειας. Αντίστοιχα, παρουσιάστηκαν και οι τρόποι αντιμετώπισης και αξιολόγησης της ευρωστίας γενικά και σε κάθε πεδίο ξεχωριστά.

Γενικά, αναδείχθηκε το πρόβλημα των ανταγωνιστικών δειγμάτων, το οποίο είναι εγγενές πρόβλημα των ML μοντέλων, και το πως αυτό επηρεάζει την επίδοση των μοντέλων, μέσω της πτώσης της ακρίβειας και αντίστοιχα της αξιοπιστίας τους. Αντίστοιχα όμως, αναδείχθηκαν και οι τεχνικές με τις οποίες μπορεί να αντιμετωπιστούν ή να μειωθεί η επίπτωση αυτών των επιθέσεων, με διάφορες αμυντικές τεχνικές και κυρίαρχη την ανταγωνιστική μάθηση, όπου χρησιμοποιεί αυτά τα δείγματα κατά την εκπαίδευση του μοντέλου, ώστε να δημιουργήσει εύρωστα μοντέλα και ανθεκτικά απέναντι σε τέτοιες επιθέσεις. Όμως, η ανάπτυξη εύρωστων μοντέλων φέρνει και κάποιες παρενέργειες, όπως μείωση τις κανονικής ακρίβειας του μοντέλου, αύξηση του χρόνου εκπαίδευσης και του υπολογιστικού κόστους, οι οποίες πρέπει να παίρνονται υπόψιν όταν γίνεται εύρωστη εκπαίδευση των μοντέλων.

Επίσης, παρουσιάστηκε το πρόβλημα της ιδιωτικότητας στα ML μοντέλα, δείχνοντας τις ευπάθειες των μοντέλων απέναντι σε επιθέσεις που έχουν σκοπό να παραβιάσουν την ιδιωτικότητα των δεδομένων εκπαίδευσης ή και των ίδιων των παραμέτρων των μοντέλων. Αντίστοιχα, αναδείχθηκαν οι ερευνημένες τεχνικές οι οποίες μπορούν να προσφέρουν προστασία από παραβιάσεις τις ιδιωτικότητας των δεδομένων εκπαίδευσης ή των μοντέλων, αλλά και τεχνικές που προσφέρουν εξ' ορισμού προστασία δεδομένων και με εγγυήσεις όπως η DP, HE, MPC και FL. Όμως, αυτές οι τεχνικές για να προσφέρουν την προστασία ιδιωτικότητας, έρχονται σε αντίθεση με τη χρηστικότητα των μοντέλων και δεδομένων, όπως η DP στην οποία πρέπει να βρεθεί ο κατάλληλος προϋπολογισμός ιδιωτικότητας, ο οποίος θα προσφέρει ασφάλεια και τον ελάχιστο δυνατό θόρυβο στα δεδομένα ή τα αποτελέσματα των μοντέλων ώστε να είναι χρήσιμα.

Όσον αφορά, τις εφαρμογές των επιθέσεων και αμυνών στα πεδία υπό έρευνα, μπορούμε να αναφερθούμε αρχικά στα κοινά χαρακτηριστικά και εφαρμογές. Τα ML μοντέλα αυτόνομης οδήγησης και ιατρικής διάγνωσης εικόνων έχουν και τα δύο εισόδους εικόνας, οπότε πολλές επιθέσεις και αμυντικές τεχνικές είναι κοινές, σε αντίθεση με τα ML μοντέλα των συστημάτων ηλεκτρικής ενέργειας τα οποία έχουν ως εισόδους αριθμητικά δεδομένα. Σύμφωνα με την ταξινόμηση που έγινε, όλες οι περιπτώσεις επιθέσεων μπορεί να είναι white-box ή black-box, evasion ή poisoning και στις εφαρμογές εικόνων, physical ή digital. Αντίστοιχα, οι άμυνες

σε όλες τις περιπτώσεις μπορεί να είναι reactive ή proactive και ανάλογα με την προσέγγιση μπορεί να τροποποιούν τα δεδομένα (data preprocessing), να αυξάνουν την ευρωστία των μοντέλων (model hardening) ή να αξιοποιούν επιπλέον μοντέλα (auxiliary models). Η τεχνική της ανταγωνιστικής εκπαίδευσης είναι η πιο διαδεδομένη μέθοδο άμυνας σε όλα τα πεδία, προσφέροντας από τις καλύτερες επιδόσεις ευρωστίας, όμως αδυνατεί να αντιμετωπίσει δείγματα με τα οποία δεν έχει εκπαιδευτεί και προκαλεί πτώση της κανονικής ακρίβειας. Οι μέθοδοι τροποποίησης των δεδομένων, συνήθως έχουν μικρό κόστος εφαρμογής, δεν τροποποιούν τα μοντέλα, άρα δε χρειάζεται επανεκπαίδευση και μπορεί να λειτουργήσουν ικανοποιητικά για ένα εύρος επιθέσεων, όμως μπορεί να επηρεάσουν και αυτοί την κανονική απόδοση του μοντέλου. Τέλος, οι τεχνικές ανίχνευσης με χρήση άλλων μοντέλων, μπορεί να προσφέρουν πολύ καλές επιδόσεις ανίχνευσης και δεν επηρεάζουν καθόλου τη λειτουργία του αρχικού μοντέλου, αυξάνοντας όμως το υπολογιστικό και χρονικό κόστος κατά την εξαγωγή συμπερασμάτων.

Συγκεκριμένα για τα ML μοντέλα αυτόνομης οδήγησης, η χρήση τεχνικών που τροποποιούν τα δεδομένα φαίνεται να έχει ικανοποιητικά αποτελέσματα στην αντιμετώπιση επιθέσεων με μικρές αρνητικές επιπτώσεις στη λειτουργία του υπόλοιπου συστήματος. Τεχνικές που τροποποιούν τα μοντέλα για βελτίωση της ευρωστίας, ενώ προσφέρουν ικανοποιητική προστασία το πρόβλημα αύξηση του χρόνου και των πόρων εκπαίδευσης, για τόσο μεγάλα μοντέλα βαθιάς μάθησης και πλήθος δεδομένων που χρησιμοποιούνται, μπορεί να είναι καταστήσει αδύνατη την εύρωστη εκπαίδευση. Επίσης, η αύξηση των υπολογιστικών πόρων και του χρόνου απόκρισης του μοντέλου κατά την εξαγωγή συμπερασμάτων με χρήση βοηθητικών μοντέλων, είναι κάτι προβληματικό για αυτά τα μοντέλα, καθώς οι υπολογιστικοί πόροι που διαθέτουν τα οχήματα στα οποία τρέχουν τα μοντέλα είναι περιορισμένοι και η γρήγορη ταχύτητα πρόβλεψης είναι κρίσιμη για την ασφάλεια των οδηγών και του οχήματος. Τα ζητήματα ιδιωτικότητας που εμφανίζονται σε αυτές τις εφαρμογές, μπορούν να λυθούν σε ένα βαθμό με χρήση FL, ώστε να μην κοινοποιούνται εξ΄ αρχής τα δεδομένα, αλλά και για τη δημιουργία πιο αποδοτικών μοντέλων μέσω της κατανεμημένης εκπαίδευσης με περισσότερα και πιο ετερογενή δεδομένα.

Συγκεκριμένα για τα ML μοντέλα ιατρικών διαγνώσεων, η πιο διαδεδομένη τεχνική για δημιουργία εύρωστων μοντέλων είναι η ανταγωνιστική εκπαίδευση, καθώς είναι η πιο αποτελεσματική για την αντιμετώπιση ανταγωνιστικών επιθέσεων και οι επιπτώσεις που προκαλεί στην αύξηση υπολογιστικών πόρων και χρόνου εκπαίδευσης είναι μικρές, καθώς δεν υπάρχουν τόσο μεγάλο πλήθος ιατρικών δυνάμεων για να αποτελεί εμπόδιο, και η ταχύτητα εξαγωγής συμπερασμάτων δεν είναι σε καμία περίπτωση κρίσιμη, όσο η ακρίβεια πρόβλεψης, το οποίο παραμένει πρόβλημα στις περιπτώσεις των καθαρών δειγμάτων. Η ιδιωτικότητα των δεδομένων είναι κρίσιμης σημασίας στον τομέα των ιατρικών δεδομένων, για αυτό και η χρήση DP έχει μελετηθεί εκτενώς, ώστε με τον ορισμό του κατάλληλου privacy budget να δίνονται αυστηρές εγγυήσεις για την ιδιωτικότητα των δεδομένων, ακόμα και αν υπάρξει μείωση της χρηστικότητας των αποτελεσμάτων.

Συγκεκριμένα για τα ML μοντέλα των συστημάτων ηλεκτρικής ενέργειας, επίσης η πιο διαδεδομένη τεχνική για δημιουργία εύρωστων μοντέλων είναι η ανταγωνιστική εκπαίδευση, η οποία είναι από τις πιο αποτελεσματικές μεθόδους, και σε αυτά τα μοντέλα η επίπτωση στην πτώση ακρίβειας δεν είναι τόσο κρίσιμη, καθώς έχουν ήδη πολύ υψηλές ακρίβειες από πριν. Όμως, λόγω του μεγάλου όγκου δεδομένων, όπως και στα ML μοντέλα αυτόνομης οδήγησης, η αύξηση του υπολογιστικού κόστους και χρόνου εκπαίδευσης μπορεί καταστήσουν μην εφικτή

την εκπαίδευση. Αντίστοιχα, για τα ζητήματα ιδιωτικότητας των δεδομένων καταναλωτών, τεχνικές όπως η FL μαζί με HE μπορούν να χρησιμοποιηθούν για τη δημιουργία εύρωστων και ιδιωτικών μοντέλων.

Τέλος, σε ένα πιο γενικό επίπεδο μοντελοποίησης των απειλών για συστήματα στους τομείς που αναλύουμε, μπορεί οι ανταγωνιστικές επιθέσεις ή οι επιθέσεις ιδιωτικότητας να μην είναι η πρώτη απειλή και αντίστοιχα η πρώτη μέριμνα των μηχανικών που τα σχεδιάζουν και των οργανισμών και επιχειρήσεων που τα χρησιμοποιούν, όμως αναδεικνύουν ένα σημαντικό κενό ασφαλείας, το οποίο είναι χαρακτηριστικό των μοντέλων μηχανικής μάθησης, οπότε και ανεξάρτητο υλοποίησης ή σχεδίασης. Όλα τα μοντέλα επηρεάζονται από αυτές τις επιθέσεις, το ζήτημα είναι το κατά πόσον είναι σημαντικό για την κάθε επιχείρηση και οργανισμό να προβλέψει και να επενδύσει στην αντιμετώπιση από αυτές πριν συμβούν και αντίστοιχα να δεχθεί τους συμβιβασμούς που προκύπτουν. Στους κρίσιμους τομείς που εξετάζουμε, τα περιθώρια σφάλματος των μοντέλων είναι πολύ μικρά και υπάρχουν ήδη μέθοδοι που είναι εφαρμόσιμοι για κάθε πεδίο, ώστε να αυξήσουν την ευρωστία και να επηρεάσουν με το λιγότερο δυνατό τρόπο τις υπόλοιπες λειτουργίες των μοντέλων. Αντίστοιχης σημασίας είναι και η προστασία της ιδιωτικότητας, κυρίως των δεδομένων, για τέτοια μοντέλα, και σε κάποιες περιπτώσεις μπορεί να είναι και προϋπόθεσή για να μπορέσει να χρησιμοποιηθεί ή εκπαιδευτεί ένα μοντέλο με πραγματικά δεδομένα.

7.2 Μελλοντικές Επεκτάσεις

Η παρούσα εργασία μπορεί να αποτελέσει έναν αναλυτικό οδηγό για τον τομέα της ευρωστίας και ιδιωτικότητας ML μοντέλων σε κρίσιμα πεδία της βιομηχανίας και των επιχειρήσεων, για την περαιτέρω διερεύνηση ή κατασκευή μεθόδων που μπορούν να εφαρμοστούν σε κάθε πεδίο για τη βελτίωση της ευρωστίας και αντίστοιχα της ιδιωτικότητας αυτών των μοντέλων.

Σε ερευνητικό επίπεδο, υπάρχουν πολλά ανοιχτά προβλήματα, τα οποία έχουμε αναφέρει σταδιακά μέσα στην εργασία, τα οποία θα μπορούσαν να αποτελέσουν την έμπνευση για επιπλέον διερεύνηση και ενασχόληση, όπως:

- Η μη ύπαρξη συστηματικού τρόπου αξιολόγησης της ευρωστίας και σύγκρισης μεταξύ μοντέλων.
- Η ενίσχυση της ευρωστίας για προστασία από πολλαπλά είδη διαταραχών ταυτόχρονα, ειδικά στη γενική του μορφή για όλα τα μοντέλα και διαταραχές.
- Η επίπτωση που έχει στην καθαρή ακρίβεια η χρήση τεχνικών εκπαίδευσης εύρωστων μοντέλων, ειδικότερα σε εφαρμογές που είναι πολύ σημαντική η ακρίβεια.
- Η εφαρμογή DP σε δεδομένα εικόνων για την παροχή αυστηρών εγγυήσεων ιδιωτικότητας και σε τέτοιου είδους δεδομένα με πρακτικό τρόπο.

Σε επίπεδο υλοποίησης, ενδιαφέρον θα παρουσίαζε ένας τυποποιημένος τρόπος δοκιμής και αξιολόγησης της ευρωστίας με αυτοματοποιημένο τρόπο, χρησιμοποιώντας τα υπάρχοντα εργαλεία στα οποία έχουμε αναφερθεί, ώστε να μπορέσει η ευρωστία των ML μοντέλων να ελέγχεται εύκολα και γρήγορα, χωρίς να χρειάζεται από τους μηχανικούς και σχεδιαστές συστημάτων ειδίκευση στον τομέα της εύρωστης και ανταγωνιστικής μηχανικής μάθησης.

Τέλος, η εργασία αυτή αναφέρεται σε δύο σημαντικά χαρακτηριστικά (ευρωστία, ιδιωτικότητα) τα οποία πρέπει να έχουν τα αξιόπιστα και ισχυρά ML μοντέλα, όμως στο γενικότερο πλαίσιο των αξιόπιστων συστημάτων τεχνητής νοημοσύνης (trustworthy AI) υπάρχουν και

άλλα χαρακτηριστικά τα οποία πρέπει να κατέχουν, όπως η ηθική και η ισότητα χωρίς προκαταλήψεις αλλά και η δυνατότητα εξήγησης των αποφάσεων που καταλήγουν. Αυτό θα μπορούσε να είναι ένα επόμενο επίπεδο διερεύνησης για τους τομείς ειδικά που αξιολογούν πιο σημαντικά αυτά τα χαρακτηριστικά.

Βιβλιογραφία

[1] Ian J. Goodfellow, Jonathon Shlens και Christian Szegedy. *Explaining and Harnessing Adversarial Examples*, 2015.

[2] Maria Rigaki και Sebastian Garcia. *A Survey of Privacy Attacks in Machine Learning. ACM Comput. Surv.*, 56(4), 2023.

[3] Hamon R, Junklewitz H και Sanchez Martin JI. *Robustness and Explainability of Artificial Intelligence. JRC Publications Repository*, 1(KJ-NA-30040-EN-N (online)), 2020.

[4] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha και Michael P. Wellman. *SoK: Security and Privacy in Machine Learning. 2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, σελίδες 399–414, 2018.

[5] Gabriel Resende Machado, Eugênio Silva και Ronaldo Ribeiro Goldschmidt. *Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. ACM Comput. Surv.*, 55(1), 2021.

[6] Adnan Qayyum, Junaid Qadir, Muhammad Bilal και Ala Al-Fuqaha. *Secure and Robust Machine Learning for Healthcare: A Survey. IEEE Reviews in Biomedical Engineering*, 14:156–180, 2021.

[7] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou και Miryung Kim. *An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, σελίδες 1–10, 2020.

[8] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[9] Anders Krogh και John Hertz. *A Simple Weight Decay Can Improve Generalization. Advances in Neural Information Processing Systems*J. Moody, S. Hanson και R.P. Lippmann, επιμελητές, τόμος 4. Morgan-Kaufmann, 1991.

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever και Ruslan Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[11] Connor Shorten και Taghi M Khoshgoftaar. *A survey on image data augmentation for deep learning. Journal of big data*, 6(1):1–48, 2019.

[12] Rich Caruana, Steve Lawrence και C. Giles. *Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. Advances in Neural Information*

*Processing Systems*T. Leen, T. Dietterich και V. Tresp, επιμελητές, τόμος 13. MIT Press, 2000.

[13] Yann LeCun, Yoshua Bengio και Geoffrey Hinton. *Deep learning. nature*, 521(7553):436–444, 2015.

[14] Alex Krizhevsky, Ilya Sutskever και Geoffrey E Hinton. *ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems*F. Pereira, C.J. Burges, L. Bottou και K.Q. Weinberger, επιμελητές, τόμος 25. Curran Associates, Inc., 2012.

[15] Karen Simonyan και Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*Yoshua Bengio και Yann LeCun, επιμελητές, 2015.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. *Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Alex Sherstinsky. *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[18] Mark A. Kramer. *Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal*, 37(2):233–243, 1991.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville και Yoshua Bengio. *Generative Adversarial Nets. Advances in Neural Information Processing Systems*Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence και K.Q. Weinberger, επιμελητές, τόμος 27. Curran Associates, Inc., 2014.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser και Illia Polosukhin. *Attention is All you Need. Advances in Neural Information Processing Systems*I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan και R. Garnett, επιμελητές, τόμος 30. Curran Associates, Inc., 2017.

[21] Jonathan Ho, Ajay Jain και Pieter Abbeel. *Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 6840–6851. Curran Associates, Inc., 2020.

[22] Julian Jang-Jaccard και Surya Nepal. *A survey of emerging threats in cybersecurity. Journal of Computer and System Sciences*, 80(5):973–993, 2014. Σπεσιαλ Ισσυε ον Δεπενδαβλε ανδ Σεχυρε δμπυτινγ.

[23] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann και Sharon Xia. *Adversarial Machine Learning-Industry Perspectives. 2020 IEEE Security and Privacy Workshops (SPW)*, σελίδες 69–75, 2020.

[24] Marco Barreno, Blaine Nelson, Anthony D Joseph και J Doug Tygar. *The security of machine learning. Machine Learning*, 81:121–148, 2010.

[25] Anna Jobin, Marcello Ienca και Effy Vayena. *The global landscape of AI ethics guidelines. Nature machine intelligence*, 1(9):389–399, 2019.

[26] N. Carlini και D. Wagner. *Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy (SP)*, σελίδες 39–57, Los Alamitos, CA, USA, 2017. IEEE Computer Society.

[27] Battista Biggio και Fabio Roli. *Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition*, 84:317–331, 2018.

[28] Matt Churilla, Nathan VanHoudnos και Robert Beveridge. *The Challenge of Adversarial Machine Learning.* https://doi.org/10.58012/jrjp-n210, 2023. Ημερομηνία πρόσβασης: 16-12-2023.

[29] Nicholas Carlini. *A Complete List of All (arXiv) Adversarial Example Papers.* https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html, 2019. Ημερομηνία πρόσβασης: 16-12-2023.

[30] Nicholas Carlini. *Lessons Learned from Evaluating the Robustness of Defenses to Adversarial Examples. USENIX Security '19 Open Access*, Santa Clara, CA, 2019. USENIX Association.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow και Rob Fergus. *Intriguing properties of neural networks*, 2014.

[32] Siddhant Garg και Goutham Ramakrishnan. *BAE: BERT-based Adversarial Examples for Text Classification. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*Bonnie Webber, Trevor Cohn, Yulan He και Yang Liu, επιμελητές, σελίδες 6174–6181, Online, 2020. Association for Computational Linguistics.

[33] Nicholas Carlini και David Wagner. *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 2018 IEEE Security and Privacy Workshops (SPW)*, σελίδες 1–7, 2018.

[34] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes και Patrick McDaniel. *Adversarial Examples for Malware Detection. Computer Security – ESORICS 2017*Simon N. Foley, Dieter Gollmann και Einar Snekkenes, επιμελητές, σελίδες 62–79, Cham, 2017. Springer International Publishing.

[35] Xiaoyu Cao και Neil Zhenqiang Gong. *Mitigating Evasion Attacks to Deep Neural Networks via Region-Based Classification. Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC '17, σελίδα 278–287, New York, NY, USA, 2017. Association for Computing Machinery.

[36] Alexey Kurakin, Ian J. Goodfellow και Samy Bengio. *Adversarial Machine Learning at Scale. International Conference on Learning Representations*, 2017.

[37] Abbas Ghaddar, Philippe Langlais, Ahmad Rashid και Mehdi Rezagholizadeh. *Context-aware Adversarial Training for Name Regularity Bias in Named Entity Recognition. Transactions of the Association for Computational Linguistics*, 9:586–604, 2021.

[38] Zico Kolter και Aleksander Madry. *Adversarial Robustness - Theory and Practice*. https://adversarial-ml-tutorial.org/, 2018. Ημερομηνία πρόσβασης: 23-12-2023.

[39] David Evans. *Security and Privacy of Machine Learning - University of Virginia cs6501 Seminar Course*. https://secml.github.io/, 2018. Ημερομηνία πρόσβασης: 23-12-2023.

[40] Weimin Zhao, Sanaa Alwidian και Qusay H. Mahmoud. *Adversarial Training Methods for Deep Learning: A Systematic Review. Algorithms*, 15(8), 2022.

[41] Ian Goodfellow, Nicolas Papernot, Sandy Huang, Pieter Abbeel Yan Duan και Jack Clark. *Attacking machine learning with adversarial examples*. https://openai.com/research/attacking-machine-learning-with-adversarial-examples, 2017. Ημερομηνία πρόσβασης: 26-12-2023.

[42] Maria Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy και Ben Edwards. *Adversarial Robustness Toolbox v1.2.0. CoRR*, 1807.01069, 2018.

[43] Ram Shankar Siva Kumar, David O Brien, Kendra Albert, Salomé Viljöen και Jeffrey Snover. *Failure Modes in Machine Learning Systems*, 2019.

[44] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer και Michael K. Reiter. *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, σελίδα 1528–1540, New York, NY, USA, 2016. Association for Computing Machinery.

[45] Kush R. Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA, 2022.

[46] Kurakin Alexey, Ian J. Goodfellow και Bengio Samy. *Adversarial Examples in the Physical World. Artificial Intelligence Safety and Security*, σελίδες 99–112, 2018.

[47] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras και Adrian Vladu. *Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations*, 2018.

[48] Florian Tramer, Nicholas Carlini, Wieland Brendel και Aleksander Madry. *On Adaptive Attacks to Adversarial Example Defenses. Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 1633–1645. Curran Associates, Inc., 2020.

[49] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik και Ananthram Swami. *The Limitations of Deep Learning in Adversarial Settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, σελίδες 372–387, 2016.

[50] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai και Yongdong Zhang. *APE-GAN: Adversarial Perturbation Elimination with GAN. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 3842–3846, 2019.

[51] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha και Ananthram Swami. *Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. 2016 IEEE Symposium on Security and Privacy (SP)*, σελίδες 582–597, 2016.

[52] Diederik P. Kingma και Jimmy Ba. *Adam: A Method for Stochastic Optimization*, 2017.

[53] Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi και Pascal Frossard. *DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi και Pascal Frossard. *Universal Adversarial Perturbations. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 86–94, 2017.

[55] Yao Deng, Tiehua Zhang, Guannan Lou, Xi Zheng, Jiong Jin και Qing Long Han. *Deep Learning-Based Autonomous Driving Systems: A Survey of Attacks and Defenses. IEEE Transactions on Industrial Informatics*, 17(12):7897–7912, 2021.

[56] Pin Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi και Cho Jui Hsieh. *ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, σελίδα 15–26, New York, NY, USA, 2017. Association for Computing Machinery.

[57] Wieland Brendel *, Jonas Rauber * και Matthias Bethge. *Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. International Conference on Learning Representations*, 2018.

[58] Jianbo Chen, Michael I. Jordan και Martin J. Wainwright. *HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. 2020 IEEE Symposium on Security and Privacy (SP)*, σελίδες 1277–1294, 2020.

[59] Pin Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi και Cho Jui Hsieh. *EAD: elastic-net attacks to deep neural networks via adversarial examples. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[60] Jiawei Su, Danilo Vasconcellos Vargas και Kouichi Sakurai. *One Pixel Attack for Fooling Deep Neural Networks. IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[61] Sayantan Sarkar, Ankan Bansal, Upal Mahbub και Rama Chellappa. *UPSET and ANGRI : Breaking High Performance Image Classifiers*, 2017.

[62] Moustapha Cisse, Yossi Adi, Natalia Neverova και Joseph Keshet. *Houdini: fooling deep structured visual and speech recognition models with adversarial examples. Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, σελίδα 6980–6990, Red Hook, NY, USA, 2017. Curran Associates Inc.

[63] Chaowei Xiao, Bo Li, Jun Yan Zhu, Warren He, Mingyan Liu και Dawn Song. *Generating adversarial examples with adversarial networks. Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, σελίδα 3905–3911. AAAI Press, 2018.

[64] Anish Athalye, Logan Engstrom, Andrew Ilyas και Kevin Kwok. *Synthesizing Robust Adversarial Examples. Proceedings of the 35th International Conference on Machine Learning*Jennifer Dy και Andreas Krause, επιμελητές, τόμος 80 στο *Proceedings of Machine Learning Research*, σελίδες 284–293. PMLR, 2018.

[65] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi και Justin Gilmer. *Adversarial Patch*, 2018.

[66] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes και Patrick McDaniel. *On the (Statistical) Detection of Adversarial Examples*, 2017.

[67] Jan Hendrik Metzen, Tim Genewein, Volker Fischer και Bastian Bischoff. *On Detecting Adversarial Perturbations. International Conference on Learning Representations*, 2017.

[68] Kimin Lee, Kibok Lee, Honglak Lee και Jinwoo Shin. *A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[69] Zhihao Zheng και Pengyu Hong. *Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks. Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[70] Weilin Xu, David Evans και Yanjun Qi. *Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. Proceedings 2018 Network and Distributed System Security Symposium*, NDSS 2018. Internet Society, 2018.

[71] Nicholas Carlini και David Wagner. *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, σελίδα 3–14, New York, NY, USA, 2017. Association for Computing Machinery.

[72] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik και Ananthram Swami. *Practical Black-Box Attacks against Machine Learning. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, σελίδα 506–519, New York, NY, USA, 2017. Association for Computing Machinery.

[73] Andrew Slavin Ros και Finale Doshi-Velez. *Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI-'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[74] Valentina Zantedeschi, Maria Irina Nicolae και Ambrish Rawat. *Efficient Defenses Against Adversarial Attacks. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, σελίδα 39–49, New York, NY, USA, 2017. Association for Computing Machinery.

[75] Chuan Guo, Mayank Rana, Moustapha Cisse και Laurensvan der Maaten. *Countering Adversarial Images using Input Transformations. International Conference on Learning Representations*, 2018.

[76] Dongyu Meng και Hao Chen. *MagNet: A Two-Pronged Defense against Adversarial Examples. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, σελίδα 135–147, New York, NY, USA, 2017. Association for Computing Machinery.

[77] Pouya Samangouei, Maya Kabkab και Rama Chellappa. *Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. International Conference on Learning Representations*, 2018.

[78] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry και Alexey Kurakin. *On Evaluating Adversarial Robustness*, 2019.

[79] Jeremy Cohen, Elan Rosenfeld και Zico Kolter. *Certified Adversarial Robustness via Randomized Smoothing. Proceedings of the 36th International Conference on Machine Learning*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 1310–1320. PMLR, 2019.

[80] Ziang Yan, Yiwen Guo και Changshui Zhang. *Deep Defense: Training DNNs with Improved Adversarial Robustness. Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[81] Shixiang Gu και Luca Rigazio. *Towards Deep Neural Network Architectures Robust to Adversarial Examples*, 2015.

[82] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin και Nicolas Usunier. *Parseval Networks: Improving Robustness to Adversarial Examples. Proceedings of the 34th International Conference on Machine Learning*Doina Precup και Yee Whye Teh, επιμελητές, τόμος 70 στο *Proceedings of Machine Learning Research*, σελίδες 854–863. PMLR, 2017.

[83] Geoffrey Hinton, Oriol Vinyals και Jeffrey Dean. *Distilling the Knowledge in a Neural Network. NIPS Deep Learning and Representation Learning Workshop*, 2015.

[84] Warren He, James Wei, Xinyun Chen, Nicholas Carlini και Dawn Song. *Adversarial Example Defense: Ensembles of Weak Defenses are not Strong. 11th USENIX Workshop on Offensive Technologies (WOOT 17)*, Vancouver, BC, 2017. USENIX Association.

[85] David Warde-Farley και Ian Goodfellow. *11 adversarial perturbations of deep neural networks. Perturbations, Optimization, and Statistics*, 311(5), 2016.

[86] Nilaksh Das, Madhuri Shanbhogue, Shang Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis και Duen Horng Chau. *Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression*, 2017.

[87] Jacob Buckman, Aurko Roy, Colin Raffel και Ian Goodfellow. *Thermometer Encoding: One Hot Way To Resist Adversarial Examples. International Conference on Learning Representations*, 2018.

[88] A. Kloukiniotis, A. Papandreou, A. Lalos, P. Kapsalas, D. V. Nguyen και K. Moustakas. *Countering Adversarial Attacks on Autonomous Vehicles Using Denoising Techniques: A Review. IEEE Open Journal of Intelligent Transportation Systems*, 3:61–80, 2022.

[89] Tianyu Pang, Kun Xu, Chao Du, Ning Chen και Jun Zhu. *Improving Adversarial Robustness via Promoting Ensemble Diversity. Proceedings of the 36th International Conference on Machine Learning*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 4970–4979. PMLR, 2019.

[90] Xuanqing Liu, Minhao Cheng, Huan Zhang και Cho Jui Hsieh. *Towards Robust Neural Networks via Random Self-ensemble. Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[91] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon και Nate Kushman. *PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. International Conference on Learning Representations*, 2018.

[92] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian και Mykel J. Kochenderfer. *Towards Proving the Adversarial Robustness of Deep Neural Networks. Proceedings of the First Workshop on Formal Verification of Autonomous Vehicles (FVAV '17)*Lukas Bulwahn, Maryam Kamali και Sven Linker, επιμελητές, τόμος 257 στο *Electronic Proceedings in Theoretical Computer Science*, σελίδες 19–26, 2017. Turin, Italy.

[93] Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza και Olivier Déforges. *Adversarial example detection for DNN models: a review and experimental comparison. Artif. Intell. Rev.*, 55(6):4403–4462, 2022.

[94] Jiajun Lu, Theerasit Issaranon και David Forsyth. *SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[95] Aditi Raghunathan, Jacob Steinhardt και Percy Liang. *Certified Defenses against Adversarial Examples. International Conference on Learning Representations*, 2018.

[96] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille και Kaiming He. *Feature Denoising for Improving Adversarial Robustness. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[97] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren και Alan Yuille. *Mitigating Adversarial Effects Through Randomization. International Conference on Learning Representations*, 2018.

[98] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal και Matthias Hein. *RobustBench: a standardized adversarial robustness benchmark. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[99] Alex Lamb, Vikas Verma, Juho Kannala και Yoshua Bengio. *Interpolated Adversarial Training: Achieving Robust Neural Networks Without Sacrificing Too Much Accuracy. Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, AISec'19, σελίδα 95–103, New York, NY, USA, 2019. Association for Computing Machinery.

[100] Yao Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov και Kamalika Chaudhuri. *A Closer Look at Accuracy vs. Robustness. Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 8588–8601. Curran Associates, Inc., 2020.

[101] Anonymous. *Towards Bridging the gap between Empirical and Certified Robustness against Adversarial Examples. Submitted to Transactions on Machine Learning Research*, 2022. Ρεθεςτεδ.

[102] Linyi Li, Tao Xie και Bo Li. *SoK: Certified Robustness for Deep Neural Networks. 2023 IEEE Symposium on Security and Privacy (SP)*, σελίδες 1289–1310, 2023.

[103] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu και Suman Jana. *Certified Robustness to Adversarial Examples with Differential Privacy. 2019 IEEE Symposium on Security and Privacy (SP)*, σελίδες 656–672, 2019.

[104] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian και Mykel J. Kochenderfer. *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. Computer Aided Verification*Rupak Majumdar και Viktor Kunčak, επιμελητές, σελίδες 97–117, Cham, 2017. Springer International Publishing.

[105] Jerry Li. *Robustness in Machine Learning (CSE 599-M)*. https://jerryzli.github.io/robust-ml-fall19, 2019. Ημερομηνία πρόσβασης: 29-01-2024.

[106] Eric Wong και Zico Kolter. *Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. Proceedings of the 35th International Conference on Machine Learning*Jennifer Dy και Andreas Krause, επιμελητές, τόμος 80 στο *Proceedings of Machine Learning Research*, σελίδες 5286–5295. PMLR, 2018.

[107] Mark Niklas Mueller, Franziska Eckert, Marc Fischer και Martin Vechev. *Certified Training: Small Boxes are All You Need. The Eleventh International Conference on Learning Representations*, 2023.

[108] Jing Wu, Mingyi Zhou, Ce Zhu, Yipeng Liu, Mehrtash Harandi και Li Li. *Performance Evaluation of Adversarial Attacks: Discrepancies and Solutions*, 2021.

[109] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio και Simon Lacoste-Julien. *A Closer Look at Memorization in Deep Networks. Proceedings of the 34th International Conference on Machine Learning*Doina Precup και Yee Whye Teh, επιμελητές, τόμος 70 στο *Proceedings of Machine Learning Research*, σελίδες 233–242. PMLR, 2017.

[110] Tsui Wei Weng, Huan Zhang, Pin Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho Jui Hsieh και Luca Daniel. *Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. International Conference on Learning Representations*, 2018.

[111] Florian Tramèr, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino και Dan Boneh. *AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, σελίδα 2005–2021, New York, NY, USA, 2019. Association for Computing Machinery.

[112] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, Georgevan den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel και Demis Hassabis. *Mastering the game of Go with deep neural networks and tree search. Nature*, 529:484–503, 2016.

[113] Francesco Croce και Matthias Hein. *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. Proceedings of the 37th International Conference on Machine Learning*Hal Daumé III και Aarti Singh, επιμελητές, τόμος 119 στο *Proceedings of Machine Learning Research*, σελίδες 2206–2216. PMLR, 2020.

[114] Francesco Croce και Matthias Hein. *Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack. Proceedings of the 37th International Conference on Machine Learning*Hal Daumé III και Aarti Singh, επιμελητές, τόμος 119 στο *Proceedings of Machine Learning Research*, σελίδες 2196–2205. PMLR, 2020.

[115] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion και Matthias Hein. *Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. Computer Vision – ECCV 2020*Andrea Vedaldi, Horst Bischof, Thomas Brox και Jan Michael Frahm, επιμελητές, σελίδες 484–501, Cham, 2020. Springer International Publishing.

[116] Dan Hendrycks και Thomas Dietterich. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. International Conference on Learning Representations*, 2019.

[117] Aditi Raghunathan, Jacob Steinhardt και Percy S Liang. *Semidefinite relaxations for certifying robustness to adversarial examples. Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[118] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi και Percy Liang. *Understanding and Mitigating the Tradeoff between Robustness and Accuracy. Proceedings of the 37th International Conference on Machine Learning*Hal Daumé III και Aarti Singh, επιμελητές, τόμος 119 στο *Proceedings of Machine Learning Research*, σελίδες 7909–7919. PMLR, 2020.

[119] Aditi Raghunathan*, Sang Michael Xie*, Fanny Yang, John Duchi και Percy Liang. *Adversarial Training Can Hurt Generalization. ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

[120] Jacob Clarysse, Julia Hörrmann και Fanny Yang. *Why adversarial training can hurt robust accuracy. The Eleventh International Conference on Learning Representations*, 2023.

[121] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner και Aleksander Madry. *Robustness May Be at Odds with Accuracy. International Conference on Learning Representations*, 2019.

[122] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui και Michael Jordan. *Theoretically Principled Trade-off between Robustness and Accuracy. Proceedings of the 36th International Conference on Machine Learning*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 7472–7482. PMLR, 2019.

[123] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar και Aleksander Madry. *Adversarially Robust Generalization Requires More Data. Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi και R. Garnett, επιμελητές, τόμος 31. Curran Associates, Inc., 2018.

[124] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi και Percy S Liang. *Unlabeled Data Improves Adversarial Robustness. Advances in Neural Information Processing Systems*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, τόμος 32. Curran Associates, Inc., 2019.

[125] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin και David Lopez-Paz. *mixup: Beyond Empirical Risk Minimization. International Conference on Learning Representations*, 2018.

[126] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz και Yoshua Bengio. *Manifold Mixup: Better Representations by Interpolating Hidden States. Proceedings of the 36th International Conference on Machine Learning*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 6438–6447. PMLR, 2019.

[127] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P. Mohanty και Matthias Bethge. *Adversarial Vision Challenge*, 2018.

[128] Florian Tramer και Dan Boneh. *Adversarial Training and Robustness for Multiple Perturbations. Advances in Neural Information Processing Systems*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, τόμος 32. Curran Associates, Inc., 2019.

[129] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long και Patrick McDaniel. *Technical Report on the CleverHans v2.1.0 Adversarial Examples Library*, 2018.

[130] Jonas Rauber, Wieland Brendel και Matthias Bethge. *Foolbox: A Python toolbox to benchmark the robustness of machine learning models*, 2018.

[131] Gavin Weiguang Ding, Luyu Wang και Xiaomeng Jin. *advertorch v0.1: An Adversarial Robustness Toolbox based on PyTorch*, 2019.

[132] Zoe Papakipos και Joanna Bitton. *AugLy: Data Augmentations for Robustness*, 2022.

[133] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin και Yanjun Qi. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*, 2020.

[134] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter και Thomas Ristenpart. *Stealing Machine Learning Models via Prediction APIs. 25th USENIX Security Symposium (USENIX Security 16)*, σελίδες 601–618, Austin, TX, 2016. USENIX Association.

[135] Mika Juuti, Sebastian Szyller, Samuel Marchal και N. Asokan. *PRADA: Protecting Against DNN Model Stealing Attacks. 2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, σελίδες 512–527, 2019.

[136] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh και Patrick McDaniel. *The Space of Transferable Adversarial Examples*, 2017.

[137] Daryna Oliynyk, Rudolf Mayer και Andreas Rauber. *I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences. ACM Comput. Surv.*, 55(14ς), 2023.

[138] Javier Galbally, Chris McCool, Julian Fierrez, Sebastien Marcel και Javier Ortega-Garcia. *On the vulnerability of face verification systems to hill-climbing attacks. Pattern Recognition*, 43(3):1027–1038, 2010.

[139] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page και Thomas Ristenpart. *Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. 23rd USENIX Security Symposium (USENIX Security 14)*, σελίδες 17–32, San Diego, CA, 2014. USENIX Association.

[140] Matt Fredrikson, Somesh Jha και Thomas Ristenpart. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, σελίδα 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.

[141] Reza Shokri, Marco Stronati, Congzheng Song και Vitaly Shmatikov. *Membership Inference Attacks Against Machine Learning Models. 2017 IEEE Symposium on Security and Privacy (SP)*, σελίδες 3–18, 2017.

[142] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito και Eric Wallace. *Extracting Training Data*

from Diffusion Models. 32nd USENIX Security Symposium (USENIX Security 23), σελίδες 5253–5270, Anaheim, CA, 2023. USENIX Association.

[143] Mohammad Al-Rubaie και J. Morris Chang. *Privacy-Preserving Machine Learning: Threats and Solutions. IEEE Security & Privacy*, 17(2):49–58, 2019.

[144] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu και Xuyun Zhang. *Membership Inference Attacks on Machine Learning: A Survey. ACM Comput. Surv.*, 54(11ς), 2022.

[145] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz και Michael Backes. *ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*, 2019.

[146] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu και Wenqi Wei. *Demystifying Membership Inference Attacks in Machine Learning as a Service. IEEE Transactions on Services Computing*, 14(6):2073–2089, 2021.

[147] Pratyush Maini, Mohammad Yaghini και Nicolas Papernot. *Dataset Inference: Ownership Resolution in Machine Learning. International Conference on Learning Representations*, 2021.

[148] Nils Lukas, Yuxuan Zhang και Florian Kerschbaum. *Deep Neural Network Fingerprinting by Conferrable Adversarial Examples. International Conference on Learning Representations*, 2021.

[149] Isabell Lederer, Rudolf Mayer και Andreas Rauber. *Identifying Appropriate Intellectual Property Protection Mechanisms for Machine Learning Models: A Systematization of Watermarking, Fingerprinting, Model Access, and Attacks. IEEE Transactions on Neural Networks and Learning Systems*, σελίδες 1–19, 2023.

[150] Honggang Yu, Kaichen Yang, Teng Zhang, Yun Yun Tsai, Tsung Yi Ho και Yier Jin. *CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples. Network and Distributed System Security Symposium*, 2020.

[151] Zhanyuan Zhang, Yizheng Chen και David Wagner. *SEAT: Similarity Encoder by Adversarial Training for Detecting Model Extraction Attack Queries. Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, AISec '21, σελίδα 37–48, New York, NY, USA, 2021. Association for Computing Machinery.

[152] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh και Dhruv Batra. *Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[153] Huadi Zheng, Qingqing Ye, Haibo Hu, Chengfang Fang και Jie Shi. *BDPL: A Boundary Differentially Private Layer Against Machine Learning Model Extraction Attacks. Computer Security – ESORICS 2019*Kazue Sako, Steve Schneider και Peter Y. A. Ryan, επιμελητές, σελίδες 66–83, Cham, 2019. Springer International Publishing.

[154] Thore Graepel, Kristin Lauter και Michael Naehrig. *ML Confidential: Machine Learning on Encrypted Data. Information Security and Cryptology – ICISC 2012*Taekyoung Kwon, Mun Kyu Lee και Daesung Kwon, επιμελητές, σελίδες 1–21, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[155] Liwei Song και Prateek Mittal. *Systematic Evaluation of Privacy Risks of Machine Learning Models. 30th USENIX Security Symposium (USENIX Security 21)*, σελίδες 2615–2632. USENIX Association, 2021.

[156] Jingwen Zhao, Yunfang Chen και Wei Zhang. *Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions. IEEE Access*, 7:48901–48911, 2019.

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

[158] Cynthia Dwork. *Differential Privacy. Automata, Languages and Programming*Michele Bugliesi, Bart Preneel, Vladimiro Sassone και Ingo Wegener, επιμελητές, σελίδες 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[159] Zecheng He, Tianwei Zhang και Ruby B. Lee. *Model inversion attacks against collaborative inference. Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, σελίδα 148–162, New York, NY, USA, 2019. Association for Computing Machinery.

[160] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang και Neil Zhenqiang Gong. *MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. Proceedings of the 2019 ACM SIGSAC Conference on Computer*

and Communications Security, CCS '19, σελίδα 259–274, New York, NY, USA, 2019. Association for Computing Machinery.

[161] Virat Shejwalkar και Amir Houmansadr. *Membership Privacy for Machine Learning Models Through Knowledge Transfer. Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9549–9557, 2021.

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

[163] Florian Tramer και Dan Boneh. *Differentially Private Learning Needs Better Features (or Much More Data). International Conference on Learning Representations*, 2021.

[164] Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith και Borja Balle. *Unlocking High-Accuracy Differentially Private Image Classification through Scale*, 2022.

[165] Wenliang Du και Mikhail J. Atallah. *Secure multi-party computation problems and their applications: a review and open problems. Proceedings of the 2001 Workshop on New Security Paradigms*, NSPW '01, σελίδα 13–22, New York, NY, USA, 2001. Association for Computing Machinery.

[166] Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim και Laurensvan der Maaten. *CrypTen: Secure Multi-Party Computation Meets Machine Learning. Advances in Neural Information Processing Systems*M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang και J. Wortman Vaughan, επιμελητές, τόμος 34, σελίδες 4961–4973. Curran Associates, Inc., 2021.

[167] Olga Ohrimenko, Felix Schuster, Cedric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani και Manuel Costa. *Oblivious Multi-Party Machine Learning on Trusted Processors. 25th USENIX Security Symposium (USENIX Security 16)*, σελίδες 619–636, Austin, TX, 2016. USENIX Association.

[168] Xun Yi, Russell Paulet και Elisa Bertino. *Homomorphic Encryption*, σελίδες 27–46. Springer International Publishing, Cham, 2014.

[169] Joon Woo Lee, Hyungchul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young Sik Kim και Jong Seon No. *Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network. IEEE Access*, 10:30039–30054, 2022.

[170] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson και Blaise Aguera y Arcas. *Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*Aarti Singh και Jerry Zhu, επιμελητές, τόμος 54 στο *Proceedings of Machine Learning Research*, σελίδες 1273–1282. PMLR, 2017.

[171] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust και M. Jorge Cardoso. *The future of digital health with federated learning. npj Digital Medicine*, 3(1), 2020.

[172] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha και Gautam Srivastava. *A survey on security and privacy of federated learning. Future Generation Computer Systems*, 115:619–640, 2021.

[173] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason V. Mancuso, Friederike Jungmann, Marc Matthias Steinborn, Andreas Saleh, Marcus R. Makowski, Daniel Rueckert και Rickmer F. Braren. *End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence*, 3:473 – 484, 2021.

[174] Yijing Li, Xiaofeng Tao, Xuefei Zhang, Junjie Liu και Jin Xu. *Privacy-Preserved Federated Learning for Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems*, 23(7):8423–8434, 2022.

[175] Robin C. Geyer, Tassilo J. Klein και Moin Nabi. *Differentially Private Federated Learning: A Client Level Perspective*, 2019.

[176] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode και Ilya Mironov. *Opacus: User-Friendly Differential Privacy Library in PyTorch. NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[177] Alexander Ziller, Andrew Trask, Antonio Lopardo, Benjamin Szymkow, Bobby Wagner, Emma Bluemke, Jean Mickael Nounahon, Jonathan Passerat-Palmbach, Kritika Prakash, Nick Rose, Théo Ryffel, Zarreen Naowal Reza και Georgios Kaissis. *PySyft: A Library for Easy Federated Learning*, σελίδες 111–139. Springer International Publishing, Cham, 2021.

[178] Adam James Hall, Madhava Jay, Tudor Cebere, Bogdan Cebere, Koen Lennartvan der Veen, George Muraru, Tongye Xu, Patrick Cason, William Abramson, Ayoub Benaissa, Chinmay Shah, Alan Aboudib, Théo Ryffel, Kritika Prakash, Tom Titcombe, Varun Kumar Khare, Maddie Shang, Ionesio Junior, Animesh Gupta, Jason Paumier, Nahua Kang, Vova Manannikov και Andrew Trask. *Syft 0.5: A Platform for Universally Deployable Structured Transparency*, 2021.

[179] Ekim Yurtsever, Jacob Lambert, Alexander Carballo και Kazuya Takeda. *A Survey of Autonomous Driving: Common Practices and Emerging Technologies. IEEE Access*, 8:58443–58469, 2020.

[180] Adnan Qayyum, Muhammad Usama, Junaid Qadir και Ala Al-Fuqaha. *Securing Connected & Autonomous Vehicles: Challenges Posed by Adversarial Machine Lear-*

ning and the Way Forward. IEEE Communications Surveys & Tutorials, 22(2):998–1026, 2020.

[181] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen και Z. Morley Mao. *On Adversarial Robustness of Trajectory Prediction for Autonomous Vehicles. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 15159–15168, 2022.

[182] Hang Bong Kang. *Various Approaches for Driver and Driving Behavior Monitoring: A Review. Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013.

[183] Ruijin Liu, Zejian Yuan, Tie Liu και Zhiliang Xiong. *End-to-End Lane Shape Prediction With Transformers. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, σελίδες 3694–3702, 2021.

[184] Di Feng, Ali Harakeh, Steven L. Waslander και Klaus Dietmayer. *A Review and Comparative Study on Probabilistic Object Detection in Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2022.

[185] Mingxing Tan και Quoc Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 6105–6114. PMLR, 2019.

[186] Li Chen, Tutian Tang, Zhitian Cai, Yang Li, Penghao Wu, Hongyang Li, Jianping Shi, Junchi Yan και Yu Qiao. *Level 2 Autonomous Driving on a Single Device: Diving into the Devils of Openpilot*, 2022.

[187] Jindi Zhang, Yang Lou, Jianping Wang, Kui Wu, Kejie Lu και Xiaohua Jia. *Evaluating Adversarial Attacks on Driving Safety in Vision-Based Autonomous Vehicles. IEEE Internet of Things Journal*, 9(5):3443–3456, 2022.

[188] Shang Tse Chen, Cory Cornelius, Jason Martin και Duen Horng (Polo) Chau. *ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. Machine Learning and Knowledge Discovery in Databases*Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley και Georgiana Ifrim, επιμελητές, σελίδες 52–68, Cham, 2019. Springer International Publishing.

[189] Adith Boloor, Xin He, Christopher Gill, Yevgeniy Vorobeychik και Xuan Zhang. *Simple Physical Adversarial Examples against End-to-End Autonomous Driving Models. 2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, σελίδες 1–7, 2019.

[190] *Tencent Keen Security Lab: Experimental Security Research of Tesla Autopilot.* https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/, 2019. Ημερομηνία πρόσβασης: 13-1-2024.

[191] Ranjie Duan, Xiaofeng Mao, A. Kai Qin, Yun Yang, Yuefeng Chen, Shaokai Ye και Yuan He. *Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink. CoRR*, αβς/2103.06504, 2021.

[192] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash και Tadayoshi Kohno. *Physical Adversarial Examples for Object Detectors. 12th USENIX Workshop on Offensive Technologies (WOOT 18)*, Baltimore, MD, 2018. USENIX Association.

[193] Qiyi He, Xiaolin Meng, Rong Qu και Ruijie Xi. *Machine Learning-Based Detection for Cyber Security Attacks on Connected and Autonomous Vehicles. Mathematics*, 8(8), 2020.

[194] Jiajun Lu, Hussein Sibai, Evan Fabry και David Forsyth. *NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles*, 2017.

[195] Husheng Zhou, Wei Li, Zelun Kong, Junfeng Guo, Yuqun Zhang, Bei Yu, Lingming Zhang και Cong Liu. *DeepBillboard: Systematic Physical-World Testing of Autonomous Driving Systems. Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, σελίδα 347–358, New York, NY, USA, 2020. Association for Computing Machinery.

[196] Zelun Kong, Junfeng Guo, Ang Li και Cong Liu. *PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[197] Adith Boloor, Karthik Garimella, Xin He, Christopher Gill, Yevgeniy Vorobeychik και Xuan Zhang. *Attacking vision-based perception in end-to-end autonomous driving models. Journal of Systems Architecture*, 110:101766, 2020.

[198] Jinghan Yang, Adith Boloor, Ayan Chakrabarti, Xuan Zhang και Yevgeniy Vorobeychik. *Finding Physical Adversarial Examples for Autonomous Driving with Fast and Differentiable Image Compositing*, 2021.

[199] Yang Zhang, Hassan Foroosh, Philip David και Boqing Gong. *CAMOU: Learning Physical Vehicle Camouflages to Adversarially Attack Detectors in the Wild. International Conference on Learning Representations*, 2019.

[200] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Prateek Mittal και Mung Chiang. *Rogue Signs: Deceiving Traffic Sign Recognition with Malicious Ads and Logos*, 2018.

[201] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang και Prateek Mittal. *DARTS: Deceiving Autonomous Cars with Toxic Signs*, 2018.

[202] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno και Dawn Song. *Robust Physical-World Attacks on Deep Learning Visual Classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 1625–1634, 2018.

[203] Huma Rehman, Andreas Ekelhart και Rudolf Mayer. *Backdoor Attacks in Neural Networks – A Systematic Evaluation on Multiple Traffic Sign Datasets. Machine Learning and Knowledge Extraction*Andreas Holzinger, Peter Kieseberg, A Min Tjoa και Edgar Weippl, επιμελητές, σελίδες 285–300, Cham, 2019. Springer International Publishing.

[204] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy και Ling Liu. *Data Poisoning Attacks Against Federated Learning Systems. Computer Security – ESORICS 2020*Liqun Chen, Ninghui Li, Kaitai Liang και Steve Schneider, επιμελητές, σελίδες 480–501, Cham, 2020. Springer International Publishing.

[205] Roland Meier, Thomas Holterbach, Stephan Keck, Matthias Stähli, Vincent Lenders, Ankit Singla και Laurent Vanbever. *(Self) Driving Under the Influence: Intoxicating Adversarial Network Inputs. Proceedings of the 18th ACM Workshop on Hot Topics in Networks*, HotNets '19, σελίδα 34–42, New York, NY, USA, 2019. Association for Computing Machinery.

[206] Omid Poursaeed, Isay Katsman, Bicheng Gao και Serge Belongie. *Generative Adversarial Perturbations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[207] Shaohua Ding, Yulong Tian, Fengyuan Xu, Qun Li και Sheng Zhong. *Trojan Attack on Deep Generative Models in Autonomous Driving. Security and Privacy in Communication Networks*Songqing Chen, Kim Kwang Raymond Choo, Xinwen Fu, Wenjing Lou και Aziz Mohaisen, επιμελητές, σελίδες 299–318, Cham, 2019. Springer International Publishing.

[208] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh και Patrick McDaniel. *Ensemble Adversarial Training: Attacks and Defenses. International Conference on Learning Representations*, 2018.

[209] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu και Jun Zhu. *Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[210] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias και Gigel Macesanu. *A survey of deep learning techniques for autonomous driving. Journal of Field Robotics*, 37(3):362–386, 2020.

[211] Anh Nguyen, Tuong Do, Minh Tran, Binh X. Nguyen, Chien Duong, Tu Phan, Erman Tjiputra και Quang D. Tran. *Deep Federated Learning for Autonomous Driving. 2022 IEEE Intelligent Vehicles Symposium (IV)*, σελίδες 1824–1830, 2022.

[212] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam και Isaac S. Kohane. *Adversarial attacks on medical machine learning. Science*, 363(6433):1287–1289, 2019.

[213] Justin Ker, Lipo Wang, Jai Rao και Tchoyoson Lim. *Deep Learning Applications in Medical Image Analysis. IEEE Access*, 6:9375–9389, 2018.

[214] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman και Shanay Rab. *Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks*, 3:58–73, 2022.

[215] Kyriakos D. Apostolidis και George A. Papakostas. *A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis. Electronics*, 10(17), 2021.

[216] Gaël Varoquaux και Veronika Cheplygina. *Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine*, 5(1):48, 2022.

[217] Meghavi Rana και Megha Bhushan. *Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications*, 82(17):26731–26769, 2023.

[218] Joseph Redmon, Santosh Divvala, Ross Girshick και Ali Farhadi. *You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 779–788, 2016.

[219] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville και Yoshua Bengio. *Generative Adversarial Networks. Commun. ACM*, 63(11):139–144, 2020.

[220] Angela Zhang, Lei Xing, James Zou και Joseph Wu. *Shifting machine learning for healthcare from development to deployment and from models to data. Nature Biomedical Engineering*, 6:1–16, 2022.

[221] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane και Andrew L. Beam. *Adversarial Attacks Against Medical Deep Learning Systems. arXiv e-prints*, σελίδα αρΞι:1804.05296, 2018.

[222] Georgios Kaissis, Marcus R. Makowski, Daniel Rückert και Rickmer F. Braren. *Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence*, 2:305–311, 2020.

[223] Binyu Tian, Qing Guo, Felix Juefei-Xu, Wen Le Chan, Yupeng Cheng, Xiaohong Li, Xiaofei Xie και Shengchao Qin. *Bias Field Poses a Threat to DNN-Based X-Ray Recognition. 2021 IEEE International Conference on Multimedia and Expo (ICME)*, σελίδες 1–6, 2021.

[224] David Kügler, Andreas Bucher, Johannes Kleemann, Alexander Distergoft, Ali Jabhe, Marc Uecker, Salome Kazeminia, Johannes Fauser, Daniel Alte, Angeelina Rajkarnikar, Arjan Kuijper, Tobias Weberschock, Markus Meissner, Thomas Vogl και Anirban Mukhopadhyay. *Physical Attacks in Dermoscopy: An Evaluation of Robustness for clinical Deep-Learning*, 2019.

[225] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro και Nassir Navab. *Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López και Gabor Fichtinger, επιμελητές, σελίδες 493–501, Cham, 2018. Springer International Publishing.

[226] Aminul Huq και Mst. Tasnim Pervin. *Analysis of Adversarial Attacks on Skin Cancer Recognition. 2020 International Conference on Data Science and Its Applications (ICoDSA)*, σελίδες 1–4, 2020.

[227] Biprodip Pal, Debashis Gupta, Md. Rashed-Al-Mahfuz, Salem A. Alyami και Mohammad Ali Moni. *Vulnerability in Deep Transfer Learning Models to Adversarial Fast Gradient Sign Attack for COVID-19 Prediction from Chest Radiography Images. Applied Sciences*, 11(9), 2021.

[228] Guohua Cheng και Hongli Ji. *Adversarial Perturbation on MRI Modalities in Brain Tumor Segmentation. IEEE Access*, 8:206009–206015, 2020.

[229] Jai Kotia, Adit Kotwal και Rishika Bharti. *Risk Susceptibility of Brain Tumor Classification to Adversarial Attacks. Man-Machine Interactions 6*Aleksandra Gruca, Tadeusz Czachórski, Sebastian Deorowicz, Katarzyna Harężlak και Agnieszka Piotrowska, επιμελητές, σελίδες 181–187, Cham, 2020. Springer International Publishing.

[230] Dawen Wu, Shishi Liu και Jian Ban. *Classification of Diabetic Retinopathy Using Adversarial Training. IOP Conference Series: Materials Science and Engineering*, 806(1):012050, 2020.

[231] Siqi Liu, Arnaud Arindra Adiyoso Setio, Florin C. Ghesu, Eli Gibson, Sasa Grbic, Bogdan Georgescu και Dorin Comaniciu. *No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting With Adversarial Attacks. IEEE Transactions on Medical Imaging*, 40(1):335–345, 2021.

[232] Aleksandra Vatian, Natalia Gusarova, Natalia Dobrenko, Sergey Dudorov, Niyaz Nigmatullin, Anatoly Shalyto και Artem Lobantsev. *Impact of Adversarial Examples on the Efficiency of Interpretation and Use of Information from High-Tech Medical Images. 2019 24th Conference of Open Innovations Association (FRUCT)*, σελίδες 472–478, 2019.

[233] Chen Chen, Chen Qin, Huaqi Qiu, Cheng Ouyang, Shuo Wang, Liang Chen, Giacomo Tarroni, Wenjia Bai και Daniel Rueckert. *Realistic Adversarial Data Augmentation for MR Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu και Leo Joskowicz, επιμελητές, σελίδες 667–677, Cham, 2020. Springer International Publishing.

[234] Francesco Calivá, Kaiyang Cheng, Rutwik Shah και Valentina Pedoia. *Adversarial Robust Training of Deep Learning MRI Reconstruction Models. Machine Learning for Biomedical Imaging*, 1:1–32, 2021.

[235] Kaiyang Cheng, Francesco Calivá, Rutwik Shah, Misung Han, Sharmila Majumdar και Valentina Pedoia. *Addressing The False Negative Problem of Deep Learning MRI Reconstruction Models by Adversarial Attacks and Robust Training. Proceedings of the Third Conference on Medical Imaging with Deep Learning*Tal Arbel, Ismail Ben Ayed, Marleende Bruijne, Maxime Descoteaux, Herve Lombaert και Christopher Pal, επιμελητές, τόμος 121 στο *Proceedings of Machine Learning Research*, σελίδες 121–135. PMLR, 2020.

[236] Fei Fei Xue, Jin Peng, Ruixuan Wang, Qiong Zhang και Wei Shi Zheng. *Improving Robustness of Medical Image Diagnosis with Denoising Convolutional Neural Networks. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew Thian Yap και Ali Khan, επιμελητές, σελίδες 846–854, Cham, 2019. Springer International Publishing.

[237] Hanwool Park, Amirhossein Bayat, Mohammad Sabokrou, Jan S. Kirschke και Bjoern H. Menze. *Robustification of Segmentation Models Against Adversarial Perturbations in Medical Imaging. Predictive Intelligence in Medicine*Islem Rekik, Ehsan Adeli, Sang Hyun Park και Maria del C. Valdés Hernández, επιμελητές, σελίδες 46–57, Cham, 2020. Springer International Publishing.

[238] Achyut Mani Tripathi και Ashish Mishra. *Fuzzy Unique Image Transformation: Defense Against Adversarial Attacks On Deep COVID-19 Models*, 2020.

[239] Saeid Asgari Taghanaki, Kumar Abhishek, Shekoofeh Azizi και Ghassan Hamarneh. *A Kernelized Manifold Mapping to Diminish the Effect of Adversarial Perturbations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[240] Anisie Uwimana και Ransalu Senanayake. *Out of Distribution Detection and Adversarial Attacks on Deep Neural Networks for Robust Medical Image Analysis. ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[241] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar και Ulfar Erlingsson. *Scalable Private Learning with PATE. International Conference on Learning Representations*, 2018.

[242] Zongkun Sun, Yinglong Wang, Minglei Shu, Ruixia Liu και Huiqi Zhao. *Differential Privacy for Data and Model Publishing of Medical Data. IEEE Access*, 7:152103–152114, 2019.

[243] Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert και Georgios Kaissis. *Medical imaging deep learning with differential privacy. Scientific Reports*, 11(1):13524, 2021.

[244] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig και John Wernsing. *CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, σελίδα 201–210. JMLR.org, 2016.

[245] Joppe W. Bos, Kristin Lauter και Michael Naehrig. *Private predictive analysis on encrypted medical data. Journal of Biomedical Informatics*, 50:234–243, 2014. Special Issue on Informatics Methods in Medical Privacy.

[246] Alexander Wood, Kayvan Najarian και Delaram Kahrobaei. *Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. ACM Comput. Surv.*, 53(4), 2020.

[247] Dong Li, Xiaofeng Liao, Tao Xiang, Jiahui Wu και Junqing Le. *Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. Computers & Security*, 90:101701, 2020.

[248] Raylin Tso, Abdulhameed Alelaiwi, Sk Md Mizanur Rahman, Mu En Wu και M Shamim Hossain. *Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. Journal of Signal Processing Systems*, 89:51–59, 2017.

[249] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang και Hairong Qi. *Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, σελίδες 2512–2520, 2019.

[250] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis και Wei Shi. *Federated learning of predictive models from federated Electronic Health Records. International Journal of Medical Informatics*, 112:59–67, 2018.

[251] Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano, Laura Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin και Jayashree Kalpathy-Cramer. *Federated Learning for Breast Density Classification: A Real-World Implementation. Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu και Ziyue Xu, επιμελητές, σελίδες 181–191, Cham, 2020. Springer International Publishing.

[252] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor και Hamid R Tizhoosh. *Federated learning and differential privacy for medical image analysis. Scientific reports*, 12(1):1953, 2022.

[253] Xiang Bai, Hanchen Wang, Liya Ma, Yongchao Xu, Jiefeng Gan, Ziwei Fan, Fan Yang, Ke Ma, Jiehua Yang, Song Bai και others. *Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. Nature Machine Intelligence*, 3(12):1081–1089, 2021.

[254] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu και Zhiming Ma. *Improved OOD Generalization via Adversarial Training and Pretraing. Proceedings of the 38th International Conference on Machine Learning*Marina Meila και Tong Zhang, επιμελητές, τόμος 139 στο *Proceedings of Machine Learning Research*, σελίδες 11987–11997. PMLR, 2021.

[255] Xi Fang, Satyajayant Misra, Guoliang Xue και Dejun Yang. *Smart Grid — The New and Improved Power Grid: A Survey. IEEE Communications Surveys & Tutorials*, 14(4):944–980, 2012.

[256] Ibrahim Yilmaz και Ambareen Siraj. *Avoiding Occupancy Detection From Smart Meter Using Adversarial Machine Learning. IEEE Access*, 9:35411–35430, 2021.

[257] Lei Cui, Youyang Qu, Longxiang Gao, Gang Xie και Shui Yu. *Detecting false data attacks using machine learning techniques in smart grid: A survey. Journal of Network and Computer Applications*, 170:102808, 2020.

[258] Ali Sayghe, Olugbenga Moses Anubi και Charalambos Konstantinou. *Adversarial Examples on Power Systems State Estimation. 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, σελίδες 1–5, 2020.

[259] Eklas Hossain, Imtiaj Khan, Fuad Un-Noor, Sarder Shazali Sikander και Md. Samiul Haque Sunny. *Application of Big Data and Machine Learning in Smart Grid, and Associated Security Concerns: A Review. IEEE Access*, 7:13960–13988, 2019.

[260] Olufemi A. Omitaomu και Haoran Niu. *Artificial Intelligence Techniques in Smart Grid: A Survey. Smart Cities*, 4(2):548–568, 2021.

[261] Muhammad Qamar Raza και Abbas Khosravi. *A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renewable and Sustainable Energy Reviews*, 50:1352–1372, 2015.

[262] Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Saad Mekhilef, Moh Yamani Idna Idris, Willem Van Deventer, Bend Horan και Alex Stojcevski. *Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews*, 81:912–928, 2018.

[263] Yize Chen, Yushi Tan και Deepjyoti Deka. *Is Machine Learning in Power Systems Vulnerable? 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, σελίδες 1–6, 2018.

[264] Qun Song, Rui Tan, Chao Ren και Yan Xu. *Understanding Credibility of Adversarial Examples against Smart Grid: A Case Study for Voltage Stability Assessment. Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, e-Energy '21, σελίδα 95–106, New York, NY, USA, 2021. Association for Computing Machinery.

[265] Iman Niazazari και Hanif Livani. *Attack on Grid Event Cause Analysis: An Adversarial Machine Learning Approach. 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, σελίδες 1–5, 2020.

[266] Ali Sayghe, Junbo Zhao και Charalambos Konstantinou. *Evasion Attacks with Adversarial Deep Learning Against Power System State Estimation. 2020 IEEE Power & Energy Society General Meeting (PESGM)*, σελίδες 1–5, 2020.

[267] Jiangnan Li, Yingyuan Yang και Jinyuan Stella Sun. *SearchFromFree: Adversarial Measurements for Machine Learning-based Energy Theft Detection. 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, σελίδες 1–6, 2020.

[268] Yize Chen, Yushi Tan και Baosen Zhang. *Exploiting Vulnerabilities of Load Forecasting Through Adversarial Attacks. Proceedings of the Tenth ACM International Conference on Future Energy Systems*, e-Energy '19, σελίδα 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[269] Jingbo Hao και Yang Tao. *Adversarial attacks on deep learning models in smart grids. Energy Reports*, 8, 2022. 2021 6th International Conference on Clean Energy and Power Generation Technology.

[270] Yang Li, Xinhao Wei, Yuanzheng Li, Zhaoyang Dong και Mohammad Shahidehpour. *Detection of False Data Injection Attacks in Smart Grid: A Secure Federated Deep Learning Approach. IEEE Transactions on Smart Grid*, 13(6):4862–4872, 2022.

[271] Jiangnan Li, Yingyuan Yang, Jinyuan Stella Sun, Kevin Tomsovic και Hairong Qi. *Towards Adversarial-Resilient Deep Neural Networks for False Data Injection Attack Detection in Power Grids. 2023 32nd International Conference on Computer Communications and Networks (ICCCN)*, σελίδες 1–10, 2023.

[272] Junfei Wang και Pirathayini Srikantha. *Stealthy Black-Box Attacks on Deep Learning Non-Intrusive Load Monitoring Models. IEEE Transactions on Smart Grid*, 12(4):3479–3492, 2021.

[273] Abdulrahman Takiddin, Muhammad Ismail, Usman Zafar και Erchin Serpedin. *Robust Electricity Theft Detection Against Data Poisoning Attacks in Smart Grids. IEEE Transactions on Smart Grid*, 12(3):2675–2684, 2021.

[274] Jiwei Tian, Tengyao Li, Fute Shang, Kunrui Cao, Jing Li και Mete Ozay. *Adaptive Normalized Attacks for Learning Adversarial Attacks and Defenses in Power Systems.*

2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), σελίδες 1–6, 2019.

[275] Donghuan Yao, Mi Wen, Xiaohui Liang, Zipeng Fu, Kai Zhang και Baojia Yang. *Energy Theft Detection With Energy Privacy Preservation in the Smart Grid. IEEE Internet of Things Journal*, 6(5):7659–7669, 2019.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλ.	βλέπε
κ.α.	και άλλα
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
π.χ.	παραδείγματος χάριν
ΕΕ	Ευρωπαϊκή Ένωση
CPU	Central Processing Unit
GPU	Graphics Processing Unit
TPU	Tensor Processing Unit
CUDA	Compute Unified Device Architecture
API	Application Programming Interface
B	Byte
GB	Gigabyte
MB	Megabyte
AI	Artificial Intelligence
ML	Machine Learning
CV	Computer Vision
NN	Neural Networks
MLP	Multi Layer Perceptron
DNN	Deep Neural Networks
CNN	Convolutional Neural Networks
AE	Autoencoders
DAE	Denoising Autoencoders
RNN	Recurrent Neural Networks
VGG	Visual Geometry Group
GAN	Generative Adversarial Network
LLM	Large Language Model
NLP	Natural Language Processing
YOLO	You Only Look Once
ASR	Attack Success Rate
CLEVER	Cross Lipschitz Extreme Value for nEtwork Robustness
LLS	Local Loss Sensitivity
RT	Robust Training
RST	Robust Self-Training
ADM	Auxiliary Detection Mode
ART	Adversarial Robustness Toolbox
FGSM	Fast Gradient Sign Method

BIM	Basic Iterative Method
PGD	Projected Gradient Descen
JSMA	Jacobian Saliency Map Attack
C&W	Carlini & Wagner
ZOO	Zeroth Order Optimization Attack
BA	Boundary Attack
HSJA	HopSkipJump Attack
DoS	Denial of Service
DDoS	Distributed Denial of Service
CVE	Common Vulnerabilities and Exposures
PPML	Privacy Preserving Machine Learning
DP	Differential Privacy
SGD	Stochastic Gradient Descent
DPSGD	Differentially Private Stochastic Gradient Descent
BDPL	Boundary Differential Privacy Layer
FL	Federated Learning
MPC	Multi Party Computation
HE	Homomorphic Encryption
AV	Autonomous Vehicles
ADS	Autonomous Driving Systems
ADAS	Advanced Driver Assistance Systems
CAV	Connected and Autonomous Vehicles
LIDAR	Light Detection and Ranging
CAD	Computer Aided Diagnosis
MRI	Magnetic Resonance Imaging
CT	Computer Tomography
PET	Positron Emission Tomography
DR	Diabetic Retinopathy
EHR	Electronic Health Record
PV	Photovoltaics
SG	Smart Grid
EV	Electronic Vehicle
IoT	Internet of Things
EMS	Energy Management System
SCADA	Supervisory Control and Data Acquisition system
RTU	Remote Terminal Units
SE	State Estimation
FDIA	False Data Injection Attacks
NILM	Nonintrusive Load Monitoring
LF	Load Forecasting

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
μηχανική μάθηση	machine learning
βαθιά μάθηση	deep learning
εκπαίδευση	training
εξαγωγή συμπερασμάτων	inference
χαρακτηριστικά	features
διάνυσμα	vector
γενίκευση	generalization
κλίση	gradient
συνάρτηση κόστους	loss function
σφάλμα δοκιμής	test error
σφάλμα πρόβλεψης	prediction error
σύνολο δεδομένων	dataset
επαύξηση	augmentation
συνάρτηση στόχου	objective function
συνάρτηση ενεργοποίησης	activation function
συνάρτηση πρόβλεψης	prediction function
ακρίβεια	accuracy
επιβλεπόμενη	supervised
μη επιβλεπόμενη	unsupervised
ενισχυτική	reinforcement
παλινδρόμηση	regression
ταξινόμηση	classification
ομαδοποίηση	clustering
νευρώνας	neuron
νευρωνικά δίκτυα	neural networks
στρώμα	layer
κρυφό	hidden
εμπρόσθια τροφοδότηση	feed forward
οπισθοδιάδοση	backpropagation
χωρητικότητα	capacity
υπερπροσαρμογή	overfitting
υποπροσαρμογή	underfitting
κατάβαση κλίσης	gradient descent
απόρριψη	dropout
κανονικοποίηση	regularization
συνέλιξη	convolution

όριο απόφασης	decision boundary
ευριστική	heuristic
βελτιστοποιητής	optimizer
χάρτης προεξοχής	saliency map
κυβερνοεπίθεση	cyberattack
κυβερνοασφάλεια	cybersecurity
μοντέλο απειλής	threat model
πεδίο επιθέσεων	attack surface
κακόβουλο λογισμικό	malware
ευάλωτος	vulnerable
επίθεση	attack
άμυνα	defense
αντίπαλος	adversarial
ανταγωνιστικό παράδειγμα	adversarial example
ανταγωνιστική μάθηση	adversarial training
διαταραχή	perturbation
εύρωστος	robust
ευρωστία	robustness
ιδιωτικός	private
ιδιωτικότητα	privacy
ανωνυμοποίηση	anonymization
αποδεδειγμένος	certified
εμπειρικός	empirical
ανίχνευση	detection
στοχευμένος	targeted
μη στοχευμένος	untargeted
σκόπιμος	intentional
μη σκόπιμος	unintentional
δηλητηρίαση	poisoning
εισβολή	evasion
προσαρμοστικός	adaptive
επαναληπτικός	iterative
μεταφορά	transfer
αναπληρωτής	surrogate
υποκατάστατο	substitute
αντιδραστικός	reactive
προληπτικός	proactive
πολλαπλότητα	manifold
ακραίες τιμές	outliers
μετρήσεις	metrics
εγγυήσεις	guarantees
επαλήθευση	verification
συμβιβασμός	trade-off

συγκριτική αξιολόγηση	benchmark
ανοιχτού κώδικα	open-source
ερώτημα	query
εξαγωγή	extraction
αντιστροφή	inversion
μέλος	member
προϋπολογισμό ιδιωτικότητας	privacy budget
αυτόνομη οδήγηση	autonomous driving
υποβοηθητική οδήγηση	driving asstistance
πορεία	trajectory
φροντίδα υγείας	healthcare
ιατρική απεικόνιση	medical imaging
κατάτμηση	segmentation
δερματοσκόπηση	dermoscopy
ακτινολογία	radiology
οφθαλμολογία	ophthalmology
βυθοσκοπικό	fundoscopic
έξυπνα δίκτυα	smart grids
έξυπνος μετρητής	smart meter