Süni İntellekt Əsaslı Adversial Hücumlar və Onlara Qarşı Kibertəhlükəsizlik Mexanizmləri

İsgəndər Məmmədov
Bakı Dövlət Universiteti, Bakı, Azərbaycan
mammadov.isgandar.nazim@bsu.edu.az

***Xülasə*— Müasir kibertəhlükəsizlik sistemləri getdikcə daha çox süni intellekt və maşın öyrənməsi modellərinə əsaslanır. Lakin bu modellərin özləri də yeni tip hücumlara — adversarial hücumlara qarşı həssasdır. Bu hücumlar modelin giriş verilənlərini minimal dəyişikliklərlə manipulyasiya etməklə yanlış nəticələr çıxarmasına səbəb olur. Məqalədə adversial hücumların mahiyyəti, onların növləri və bu hücumlara qarşı süni intellekt əsaslı müdafiə mexanizmləri təhlil olunur. Eyni zamanda real tətbiq ssenariləri və gələcək istiqamətlər müzakirə edilir.**

Açar sözlər*— *adversial hücumlar; süni intellekt; təhlükəsizlik modelləri; kibertəhlükəsizlik.

I. GİRİŞ

Rəqəmsal transformasiya dövründə kibertəhlükəsizlik məsələləri daha da aktuallaşmışdır. Ənənəvi təhlükəsizlik sistemləri artıq mürəkkəb və adaptiv hücumlara qarşı kifayət qədər effektiv deyil. Bu səbəbdən süni intellekt və maşın öyrənməsi texnologiyaları kibertəhlükəsizlik sistemlərinə inteqrasiya olunur. Süni intellekt əsaslı sistemlər böyük həcmli məlumatları analiz edərək anomaliyaları aşkar edə bilir və real-vaxt rejimində qərarlar qəbul edir [1]. Lakin bu sistemlərin özləri də yeni nəsil hücumlara məruz qalır. Bu hücumlar adversial hücumlar adlanır və süni intellekt modellərinin zəif tərəflərindən biridir. Adversial hücum, modelin təlim prosesi zamanı görmədiyi və ya qəsdən manipulyasiya edilmiş giriş məlumatları ilə sistemi aldatmaq cəhdidir. Dərin neyron şəbəkələri təsvirlərin tanınması və təbii dilin emalı (Natural Language Processing, NLP) kimi sahələrdə insan səviyyəsində məhsuldarlıq göstərsə də, onların qərar vermə prosesi qara qutu xarakteri daşıyır [2]. Bu qeyri-şəffaflıq, təcavüzkarlara giriş məlumatlarına kiçik, lakin strateji dəyişikliklər edərək sistemi tamamilə yanlış nəticələrə sövq etməyə imkan verir.

II. TƏDQİQAT

Adversial hücumlar kibertəhlükəsizlik sistemləri üçün ciddi risklər yaradır [3]. Bu hücumların əsas xüsusiyyəti ondan ibarətdir ki, giriş məlumatlarında çox kiçik dəyişikliklər etməklə modelin nəticəsi tamamilə dəyişdirilə bilər. Süni intellekt modellərinin manipulyasiyaya açıq olması, hücumların aşkarlanmasının çətinliyi, təhlükəsizlik sistemlərinin aldadılması adversial hücumlar üçün əsas problemlərdir.

Adversial hücumların aşağıdakı növləri vardır:

1. *Qaçış hücumları (Evasion Attacks)*

Model işlədiyi anda giriş məlumatı dəyişdirilir və sistem aldadılır. Bu hücumlar model artıq öyrədildikdən sonra – yəni "test mərhələsində" həyata keçirilir. Məqsəd modeli aldatmaq və səhv nəticə çıxarmağa məcbur etməkdir. Bu modeldə giriş məlumatına (məsələn, şəkil, mətn) çox kiçik dəyişikliklər edir. Amma bu dəyişikliklər elə edilir ki, insan demək olar ki, onu görünməz olaraq qəbul edir. İnsan bunu qəbul etsə də model qəbul etmir, və onun üçün ciddi fərq yaradır.

Nümunə üçün şəkil tanıma sistemində bir panda şəkli üzərində azca "səs-küy" əlavə edilir. Model isə bunu artıq panda olaraq yox, məsələn gibbon kimi tanıya bilər. Bu da azacıq dəyişimin nəticəsidir. İnsan üçün nəzərə çarpmayan bu dəyişiklik model tərəfindən tamamilə fərqli şəkildə interpretasiya oluna bilər. Bu problemin yaratdığı risklərə nəzər salanda görə bilərik ki, həqiqətən də ciddi fəsadlar yarana bilər. Məsələn, avtonom avtomobillərdə yol nişanlarının səhv tanınması, biometrik sistemlərin aldadılması böyük xətalara yol açа bilər.

2. *Zəhərlənmə hücumları (Poisoning Attacks)*

Modelin öyrədildiyi təlim verilənlər toplusuna (dataset)zərərli məlumatlar daxil edilir və model yanlış öyrədilir. Bu hücumlar modelin təlim mərhələsində həyata keçirilir. Hücumun əsas məqsədi modelin öyrənmə prosesini pozmaqdır. Təlim verilənlər toplusuna qəsdən səhv və ya manipulyasiya olunmuş məlumat göndərilir və model bu yanlış məlumatı "doğru" kimi öyrənir.

Bu hücumların iki növü vardır: Etiket zəhərlənməsi (Label poisoning) və Arxaqapı (Backdoor) hücumları vardır. Etiket zəhərlənməsi nəticəsində düzgün məlumat yanlış etiketlənir və səhv göndərilir. Arxaqapı hücumlarında isə modelə gizli "trigger" əlavə olunur. Nümunə üçün spam filtr sistemini göstərə bilərik. Burada spam mesajlar normal kimi işarələnir və nəticədə model spamı düzgün ayırd edə bilmir. Təbii ki, sistemli şəkildə qərar verən Süni intellekt modellərində, kritik tətbiqlərdə məsələn, tibbi AI sistemlərində ciddi risklər və nəticələr yaranır.

3. *Modelin çıxarılması (Model Extraction Attacks)*

Hücum edən şəxs modelin davranışını analiz edərək onun strukturunu təxmin edir. Bu hücumda hücumçu modelin özünü "oğurlamağa" çalışır. Hücumun məqsədi qara qutu (black-box) modelin funksiyasını təkrar qurmaqdır. Hücum zamanı hücumçu modelə çoxlu sorğular göndərir və nəticədə giriş-çıxış cütlərini analiz edərək oxşar modellər qurmağa çalışır və orijinal modelin davranışını təqlid edən yeni model əldə edərək hücum edir. Bu kommersiya modellərinin oğurlanması və intellektual mülkiyyətin pozulması kimi risklər yaradır.

DOI: 10.25045/forensic.gov.2026.52

4. Üzvlük inferensiyası hücumları (Membership Inference Attacks)

Bu hücum məlumat məxfiliyinə yönəlib. Modelin öyrədildiyi məlumatların məxfiliyi pozulur və istifadəçi məlumatları ifşa olunur. Məqsəd isə konkret bir məlumatın modelin təlim verilənlər toplusunda olub-olmadığını müəyyən etməkdir. Modelin müəyyən girişlərə verdiyi reaksiyalar analiz olunur və model təlimində gördüyü məlumatlara fərqli etimad ilə cavab verir və bu fərqlərdən istifadə edilərək nəticə çıxarılır. Məsələn, tibbi modeldə konkret bir şəxsin məlumatının istifadə olunub-olunmadığını müəyyən etmək üçün istifadə edilə bilər. Hər hansı şəxsin şəxsi məlumatlarından istifadə edilərək ifşa edilməsi, GDPR kimi məxfilik qanunlarının pozulması riskləri isə bu hücumlara xas risklərdir.

Süni intellekt sistemlərini adversial hücumlardan qorumaq üçün müxtəlif müdafiə yanaşmaları mövcuddur. Bu yanaşmalar əsasən modelin davamlılığını, məxfiliyini və şəffaflığını artırmağa yönəlib. Süni intellekt əsaslı müdafiə sistemlərini aşağıdakı kimi xarakterizə etmək olar:

Rəqib təlim. Model xüsusi olaraq "saxta" hücum nümunələri ilə öyrədilir. Bu, modelin davamlılığını artırır və hücumlara qarşı daha dayanıqlı edir. Rəqib təlim, yəni adversial Training, süni intellekt modellərinin daha davamlı və hücumlara qarşı dayanıqlı olmasını təmin edən əsas metodlardan biridir. Bu yanaşmada model yalnız adi, təmiz verilənlər üzərində deyil, həm də xüsusi olaraq hazırlanmış və modelin zəif cəhətlərini üzə çıxaran adversial nümunələr üzərində öyrədilir. Beləliklə, model təkcə normal halları deyil, həm də manipulyasiya olunmuş girişləri düzgün tanımağı öyrənir. Bu metod xüsusilə qaçış hücumlarına qarşı effektiv hesab olunur, çünki model əvvəlcədən bu tip aldadıcı dəyişikliklərlə qarşılaşmış olur. Bununla belə, rəqib təlim yüksək hesablama resursları tələb edir və bütün mümkün hücum ssenarilərini əhatə etmək praktik olaraq çətindir.

Müdafiə distillasiyası (Defensive Distillation). Modelin qərarvermə prosesini daha sabit etmək üçün istifadə olunur. Bu metod modelin kiçik dəyişikliklərə qarşı həssaslığını azaldır. Bu üsulda ilkin olaraq bir model öyrədilir və onun çıxışları – yəni ehtimal paylanmaları – daha sonra ikinci modelin təlimində istifadə olunur. Nəticədə ikinci model daha "yumşaq" qərar sərhədlərinə malik olur və kiçik dəyişikliklərə qarşı daha az həssaslıq göstərir. Bu yanaşma modelin ümumiləşdirmə qabiliyyətini artırsa da, son illərdə inkişaf edən daha mürəkkəb hücum metodlarına qarşı hər zaman yetərli müdafiə təmin etmir. Buna görə də müdafiə distillasiyası adətən digər təhlükəsizlik mexanizmləri ilə birlikdə tətbiq olunur [4].

Anomaliyaların aşkarlanması. Süni intellekt vasitəsilə sistemdə qeyri-adi davranışlar müəyyən edilir. Anomaliyaların aşkarlanması isə Anomaly Detection çərçivəsində inkişaf etdirilən və sistemə daxil olan məlumatların normal davranışdan kənara çıxmasını müəyyən etməyə yönəlmiş yanaşmadır. Bu metodda model əvvəlcə normal verilənlərin statistik və ya struktur xüsusiyyətlərini öyrənir, daha sonra isə yeni daxil olan məlumatları bu öyrənilmiş paylanma ilə müqayisə edir. Əgər ciddi uyğunsuzluq aşkar olunarsa, bu hal potensial hücum və ya anormal davranış kimi qiymətləndirilir. Bu yanaşmanın əsas üstünlüyü ondan ibarətdir ki, o, yalnız əvvəlcədən məlum olan hücumları deyil, həm də yeni və gözlənilməz təhlükələri aşkar edə bilir. Xüsusilə real-vaxt rejimində işləyən sistemlərdə və kibertəhlükəsizlik tətbiqlərində bu metod geniş istifadə olunur [5, 6].

İzah edilə bilən süni intellekt. Bu yanaşma modelin qərarlarının izah olunmasını təmin edir. İzah edilə bilən süni intellekt, yəni Explainable Artificial Intelligence, müdafiə mexanizmləri arasında bir qədər fərqli, lakin çox vacib rol oynayır. Bu yanaşma modelin qərarlarının insan tərəfindən başa düşülə bilən şəkildə təqdim edilməsini təmin edir. Ənənəvi "qara qutu" modellərdə qərarın necə verildiyini anlamaq çətin olduğu halda, izah edilə bilən süni intellekt metodları modelin hansı xüsusiyyətlərə əsaslandığını və hansı səbəbdən konkret nəticəyə gəldiyini göstərir. Bu isə yalnız etibarlılığı artırmır, həm də potensial hücumların və manipulyasiyaların aşkarlanmasını asanlaşdırır. Məsələn, əgər model qərar verərkən qeyri-adi və ya əlaqəsiz xüsusiyyətlərə əsaslanırsa, bu, adversial müdaxilənin göstəricisi ola bilər. Ümumilikdə, bu müdafiə yanaşmaları süni intellekt sistemlərinin təhlükəsizliyini təmin etmək üçün bir-birini tamamlayan mexanizmlər kimi çıxış edir. Müasir praktikada adətən bu metodların kombinasiyası tətbiq olunur: rəqib təlim modelin davamlılığını artırır, anomaliyaların aşkarlanması hücumları müəyyən edir, müdafiə distillasiyası sabitliyi gücləndirir, izah edilə bilən süni intellekt isə bütün prosesin şəffaf və nəzarət edilə bilən olmasını təmin edir. Belə kompleks yanaşma olmadan, xüsusilə kritik sahələrdə istifadə olunan süni intellekt sistemlərinin etibarlı və təhlükəsiz işləməsi mümkün deyil.

Praktikada tətbiq sahələrinə isə şəbəkə təhlükəsizliyi, maliyyə sistemləri, biometrik sistemlər, bulud texnologiyalarını göstərmək olar. Real-vaxtd rejimində trafik analizi və hücumların aşkarlanması şəbəkə təhlükəsizliyinin qorunmasını yaradır. Fırıldaqçılıq əməliyyatlarının aşkarlanması maliyyə sistemlərinə olunmuş hücumlarının qarşısını almaq üçün istifadə edilir. Üz tanıma və identifikasiya sistemlərinin qorunması biometrik sistemlər üçün əsasdır. Bulud texnologiyaları üçün məlumatların qorunması və giriş nəzarətinin olması mütləqdir, çünki olunacaq hər hansı hücum məlumatların itməsinə və zərər görməsinə səbəb ola bilər.

Adversial hücumların riyazi modeli

Adversial hücumun məqsədi verilmiş x girişinə elə bir β (perturbasiya (kiçik dəyişiklik)) əlavə etməkdir ki, modelin çıxışı $f(x+ \beta) \neq f(x)$ olsun, lakin $\| \beta \| \leq ɛ$ şərti daxilində bu dəyişiklik insan tərəfindən hiss edilməsin [7].

Təlim Mərhələsində Hücumlar

Təcavüzkar təlim verilənlər bazasına zərərli məlumatlar daxil edərək modelin öyrənmə prosesini korlayır. Bu, gələcəkdə modelin müəyyən arxa qapısitəsilə idarə olunmasına şərait yaradır.

Çıxış Mərhələsində Hücumlar

Ən geniş yayılmış hücum növüdür. Burada model artıq təlim keçib, lakin real zamanlı istifadə zamanı aldadılır.

- Carlini & Wagner (C&W) Hücumu: Ən güclü hücumlardan biri hesab olunur və optimallaşdırma funksiyasına əsaslanır.
- DeepFool: Modelin qərar sərhədlərini (decision boundaries) keçmək üçün tələb olunan minimum perturbasiyanı hesablayır [8].

Kibertəhlükəsizlik kontekstində süni intellekt sistemlərinin qorunması çoxsəviyyəli yanaşma tələb edir və bu yanaşmanın əsas istiqamətlərindən biri modelin möhkəmləndirilməsi, digəri isə giriş məlumatlarının əvvəlcədən emal olunmasıdır. Bu iki istiqamət bir-birini tamamlayaraq Süni intellekt və Kibertəhlükəsizlik sahələrində adversial hücumlara qarşı daha dayanıqlı sistemlərin qurulmasına imkan yaradır. Modelin möhkəmləndirilməsi çərçivəsində ən geniş tətbiq olunan metodlardan biri rəqib təlim, yəni Adversarial Training hesab olunur. Bu yanaşmada model yalnız adi təlim nümunələri ilə deyil, həm də məqsədli şəkildə dəyişdirilmiş, yəni adversialnümunələrlə birlikdə öyrədilir. Riyazi baxımdan bu proses min–max optimallaşdırma problemi kimi ifadə olunur: model bir tərəfdən səhv ehtimalını minimuma endirməyə çalışır, digər tərəfdən isə ən pis halda belə (yəni ən güclü hücum ssenarisində) düzgün nəticə verməyə optimallaşdırılır. Bu, əslində model ilə hücumçu arasında bir növ "oyun" qurur. Nəticədə model yalnız normal verilənlərə deyil, həm də potensial manipulyasiyalara qarşı daha davamlı olur. Bununla belə, bu metodun əsas çətinliyi ondan ibarətdir ki, bütün mümkün hücum variantlarını əvvəlcədən nəzərə almaq mümkün deyil və bu proses yüksək hesablama resursları tələb edir.

Modelin möhkəmləndirilməsində istifadə olunan digər üsul isə qradient maskalanması (Gradient Masking) kimi tanınan qradient maskalanmasıdır. Bu yanaşmanın məqsədi hücumçunun modelin daxili strukturunu analiz etməsini çətinləşdirməkdir. Bildiyimiz kimi, bir çox adversial hücumlar modelin qradient məlumatlarına əsaslanır və bu qradientlər vasitəsilə giriş verilənlərində hansı dəyişikliklərin modeli aldada biləcəyi hesablanır. Qradient maskalanması zamanı modelin çıxış funksiyası və ya aktivasiya mexanizmləri elə dəyişdirilir ki, qradientlər ya çox kiçik, ya da qeyri-stabil olur. Nəticədə hücumçu üçün effektiv istiqamət tapmaq çətinləşir. Lakin bu metod tam təhlükəsizlik təmin etmir, çünki bəzi hallarda hücumçular alternativ yanaşmalarla bu məhdudiyyəti aşa bilir. Buna görə də qradient maskalanması daha çox əlavə müdafiə qatlarından biri kimi istifadə olunur.

Digər əsas istiqamət isə giriş məlumatlarının emalıdır ki, burada məqsəd hücumun təsirini modelə çatmadan əvvəl azaltmaqdır. Bu çərçivədə tətbiq olunan metodlardan biri küyün təmizlənməsidir və bu prosesdə tez-tez avtoenkoderlərdən istifadə olunur. Autoenkoder tipli neyron şəbəkələr giriş verilənlərini sıxılmış formada kodlayıb yenidən bərpa etməklə onun əsas struktur xüsusiyyətlərini saxlayır, lakin əlavə edilmiş təsadüfi və ya məqsədli səs-küyləri aradan qaldıra bilir. Bu, xüsusilə şəkil və audio məlumatlarda effektivdir. Məsələn, adversial olaraq dəyişdirilmiş bir şəkil əvvəlcə avtoenkoderdən keçirilir və nəticədə daha "təmiz" versiyası modelə daxil olur. Bununla da hücumun təsiri əhəmiyyətli dərəcədə azalır.

Giriş məlumatlarının emalı sahəsində geniş istifadə olunan digər metod isə JPEG sıxılmasıdır. Bu üsul xüsusilə təsvir məlumatlarında tətbiq olunur və məqsədi yüksək tezlikli komponentləri azaltmaqdır. Adversial hücumlar adətən insan gözü ilə görünməyən, lakin model üçün kritik olan yüksək tezlikli dəyişikliklər əlavə edir. JPEG sıxılması zamanı bu yüksək tezlikli komponentlər sıxılma prosesi nəticəsində zəiflədilir və ya tamamilə aradan qaldırılır. Beləliklə, modelə daxil olan təsvir daha sadə və daha az manipulyasiya olunmuş formada olur. Bununla belə, bu metodun da çatışmazlıqları var, çünki həddindən artıq kompressiya faydalı məlumatın da itirilməsinə səbəb ola bilər və modelin ümumi performansına mənfi təsir göstərə bilər [8].

Ümumilikdə, modelin möhkəmləndirilməsi və giriş məlumatlarının emalı adversial hücumlara qarşı müdafiənin iki mühüm və bir-birini tamamlayan istiqamətidir. Birinci yanaşma modelin daxilində davamlılığı artırmağa yönəldiyi halda, ikinci yanaşma hücumun təsirini modelə çatmamış azaltmağa çalışır. Müasir sistemlərdə bu metodlar adətən birlikdə tətbiq olunur və bu da süni intellekt əsaslı sistemlərin daha təhlükəsiz və etibarlı işləməsini təmin edir [9].

Hücumun Aşkarlanması

Hücumun aşkarlanması sistemləri adətən əvvəlcə "normal" verilənlərin statistik xüsusiyyətlərini öyrənir. Bu xüsusiyyətlərə verilənlərin paylanması, xüsusiyyətlərarası əlaqələr, tezlik komponentləri və digər struktur parametrlər daxildir. Daha sonra sistemə daxil olan yeni sorğular bu öyrənilmiş model ilə müqayisə edilir. Əgər yeni verilən əvvəlki paylanmadan əhəmiyyətli dərəcədə kənara çıxırsa, bu, potensial olaraq manipulyasiya olunmuş və ya adversial hücum nəticəsində yaradılmış nümunə kimi qiymətləndirilir. Bu proses statistik kənarlaşmaların aşkarlanmasına əsaslanır və çox vaxt ehtimal modelləri, klasterləşdirmə və ya neyron şəbəkə əsaslı yanaşmalar vasitəsilə həyata keçirilir. Bu cür sistemlər xüsusilə ona görə effektivdir ki, onlar yalnız əvvəlcədən məlum olan hücumları deyil, həm də yeni və daha əvvəl rast gəlinməyən hücum növlərini aşkar edə bilir. Məsələn, adversial olaraq dəyişdirilmiş bir şəkil və ya mətn modeli aldatmaq məqsədi daşısa da, onun statistik xüsusiyyətləri adi verilənlərdən fərqlənə bilər. Sistem bu fərqi aşkar edərək ya həmin sorğunu bloklayır, ya da əlavə yoxlama mərhələsinə yönləndirir. Bəzi hallarda isə sistem istifadəçiyə və ya administratora xəbərdarlıq siqnalı göndərir. Hücumun aşkarlanması mexanizmlərində müxtəlif texniki yanaşmalar tətbiq olunur. Bunlara giriş verilənlərinin xüsusiyyətlərinin dərin analizi, daxili model aktivasiyalarının monitorinqi və ya xüsusi detektor modellərin qurulması daxildir. Məsələn, bəzi sistemlər əsas modeldən əlavə olaraq ikinci bir "detektor model" istifadə edir. Bu model yalnız verilənin təbii və ya süni olduğunu müəyyən etməyə fokuslanır. Digər yanaşmalarda isə əsas modelin gizli qatlarında yaranan aktivasiya nümunələri analiz edilir və anormal davranışlar aşkar edilir. Bununla belə, bu yanaşmanın da müəyyən çətinlikləri mövcuddur. Ən əsas problemlərdən biri yanlış pozitiv və yanlış neqativ nəticələrdir. Bəzən tamamilə normal bir verilən səhvən hücum kimi qiymətləndirilə bilər və ya əksinə, yaxşı hazırlanmış bir adversial nümunə aşkarlanmadan keçə bilər. Bundan əlavə, hücumçular da öz metodlarını təkmilləşdirərək bu deteksiya sistemlərini keçməyə çalışırlar ki, bu da "silahlanma yarışı" kimi xarakterizə olunan bir vəziyyət yaradır.

III. TƏDQİQAT VƏ MÜQAYİSƏLİ TƏHLİL

Rəqib (adversarial) hücumlara qarşı müdafiə üsullarının seçilməsi yalnız onların təhlükəsizlik baxımından effektivliyi ilə deyil, eyni zamanda hesablama mürəkkəbliyi, tətbiq sahəsi və modelin ümumi məhsuldarlığına təsiri ilə də qiymətləndirilməlidir. Müasir tədqiqatlar göstərir ki, universal və bütün hücum növlərinə qarşı eyni dərəcədə effektiv olan müdafiə metodu mövcud deyildir. Buna görə də müdafiə

strategiyasının seçilməsi tətbiqin xüsusiyyətləri, istifadə olunan modelin arxitekturası və gözlənilən hücum ssenariləri nəzərə alınmaqla həyata keçirilməlidir.

Bu tədqiqat çərçivəsində geniş istifadə olunan dörd əsas müdafiə yanaşması – rəqib təlim (Adversarial Training), müdafiə distillasiyası (Defensive Distillation), giriş verilənlərinin transformasiyası (Input Transformation) və dayanıqlı optimallaşdırma (Robust Optimization) – effektivlik, hesablama xərci və potensial zəifliklər baxımından müqayisə edilmişdir. Müqayisənin nəticələri Cədvəl 1-də təqdim olunur.

CƏDVƏL I. RƏQİB HÜCUMLARINA QARŞI ƏSAS MÜDAFİƏ METODLARININ MÜQAYİSƏLİ TƏHLİLİ

Müdafiə Metodu	Effektivlik	Hesablama xərci	Zəif tərəfi
Rəqib Təlim	Yüksək	Çox yüksək	Yalnız tanınan hücumlara qarşıdır
Müdafiə Distillasiyası	Orta	Orta	C&W hücumuna qarşı zəifdir
Giriş Transformasiyası	Orta	Aşağı	Təsvir keyfiyyətini azalda bilər
Möhkəm Optimallaşdırma	Yüksək	Yüksək	Modelin ümumi dəqiqliyini azalda bilər

Cədvəl I-də təqdim olunan nəticələr göstərir ki, rəqib təlim təhlükəsizlik baxımından ən etibarlı metodlardan biri hesab olunur. Bu yanaşmada model təlim mərhələsində həm orijinal, həm də qəsdən dəyişdirilmiş nümunələr üzərində öyrədildiyindən, hücumlara qarşı daha davamlı olur. Bununla belə, bu metod əlavə rəqib nümunələrin yaradılmasını tələb etdiyi üçün hesablama xərci əhəmiyyətli dərəcədə artır və böyük verilənlər bazaları üzərində təlim prosesini uzadır. Müdafiə distilləsi neyron şəbəkəsinin çıxış ehtimallarını yumşaldaraq modelin həssaslığını azaltmağa çalışır. Bu metod hesablama baxımından daha sərfəli olsa da, son illərdə hazırlanmış optimallaşdırma əsaslı hücumlar, xüsusilə Carlini–Wagner (C&W) hücumu qarşısında kifayət qədər dayanıqlı olmadığı müəyyən edilmişdir. Buna görə də bu yanaşma müstəqil müdafiə vasitəsi kimi deyil, əlavə təhlükəsizlik mexanizmi kimi daha məqsədəuyğun hesab edilir. Giriş transformasiyası üsulu təsvirlərin sıxılması, filtrlənməsi, yenidən ölçüləndirilməsi və ya təsadüfi çevrilmələri vasitəsilə rəqib perturbasiyalarını zəiflətməyə çalışır. Bu metodun əsas üstünlüyü aşağı hesablama xərci və mövcud sistemlərə asan inteqrasiya edilməsidir. Lakin transformasiya zamanı faydalı xüsusiyyətlərin də qismən itirilməsi modelin normal verilənlər üzərində dəqiqliyinin azalmasına səbəb ola bilər. Möhkəm optimallaşdırma isə model parametrlərinin ən pis ssenarilərdə belə sabit işləməsini təmin etməyə yönəlmiş riyazi optimallaşdırma metodlarına əsaslanır. Bu yanaşma yüksək müdafiə səviyyəsi təmin etsə də, təlim prosesinin mürəkkəbliyini artırır, daha çox hesablama resursu tələb edir və bəzi hallarda modelin təmiz verilənlər üzərində təsnifat dəqiqliyini aşağı sala bilir. Aparılmış müqayisəli təhlil göstərir ki, müdafiə metodları arasında təhlükəsizlik səviyyəsi ilə hesablama xərci arasında birbaşa asılılıq mövcuddur. Yüksək müdafiə təmin edən metodlar adətən daha çox hesablama resursu və uzun təlim müddəti tələb edir. Əksinə, hesablama baxımından qənaətcil yanaşmalar mürəkkəb və adaptiv rəqib hücumlarına qarşı kifayət qədər etibarlı olmaya bilər. Buna görə də son illərdə aparılan elmi tədqiqatlarda hibrid müdafiə strategiyalarına üstünlük verilir. Belə yanaşmalarda rəqib təlim giriş transformasiyası, anomaliyaların aşkarlanması və ya möhkəm optimallaşdırma üsulları ilə birlikdə tətbiq edilir. Bu kombinasiyalar həm modelin dayanıqlığını artırır, həm də təhlükəsizlik və hesablama səmərəliliyi arasında daha optimal tarazlığın əldə edilməsinə imkan yaradır. Beləliklə, real tətbiqlərdə müdafiə metodunun seçimi konkret sistemin təhlükəsizlik tələbləri, hesablama imkanları və istifadə ssenarilərinə uyğun şəkildə müəyyən edilməlidir.

IV. NƏTİCƏ VƏ GƏLƏCƏK PERSPEKTİVLƏR

Nəticə etibarilə demək olar ki, süni intellekt sistemlərinin effektiv müdafiəsi yalnız ayrı-ayrı texnologiyaların tətbiqi ilə deyil, inteqrasiya olunmuş və çoxsəviyyəli yanaşmalar vasitəsilə mümkündür. Məsələn, rəqib təlim modelin davamlılığını artırsa da, bu metod bütün mümkün hücum ssenarilərini əhatə edə bilmir. Eyni zamanda anomaliyaların aşkarlanması yeni və naməlum hücumların müəyyən edilməsində effektiv olsa da, bəzən yanlış pozitiv nəticələr verə bilər. Müdafiə distilləsi modelin stabilliyini artırsa da, daha mürəkkəb hücumlar qarşısında zəif qala bilir. İzah edilə bilən süni intellekt isə təhlükəsizlik mexanizmi olmaqla yanaşı, həm də bu sistemlərin qərarlarının şəffaflığını təmin edərək istifadəçi etibarını artırır və hücumların analizini asanlaşdırır. Bu səbəbdən müasir yanaşmalar bu metodların sinerjisini əsas götürür və kompleks müdafiə arxitekturalarının qurulmasını prioritet hesab edir. Gələcək perspektivlər baxımından bu sahədə bir neçə mühüm istiqamət xüsusi diqqət çəkir. İlk növbədə, adaptiv və özünü öyrədən müdafiə sistemlərinin yaradılması əsas hədəflərdən biridir. Belə sistemlər real-vaxt rejimində yeni hücum strategiyalarını aşkar edib onlara uyğunlaşa biləcək. Bununla yanaşı, İzah edilə bilən süni intellekt (Explainable Artificial Intelligence, XAI) metodlarının daha da inkişaf etdirilməsi gözlənilir, çünki gələcəkdə yalnız dəqiq deyil, həm də izah oluna bilən modellərə ehtiyac artacaq. Xüsusilə tibb, maliyyə və hüquq kimi kritik sahələrdə qərarların şəffaflığı təhlükəsizlik qədər vacib olacaqdır. Digər mühüm istiqamət isə məxfilik yönümlü öyrənmə metodlarının inkişafıdır. Bu baxımdan diferensial məxfilik, federativ öyrənmə və təhlükəsiz çoxtərəfli hesablama kimi yanaşmaların daha geniş tətbiqi gözlənilir. Bu metodlar modelin öyrənmə prosesində istifadə olunan məlumatların qorunmasını təmin etməklə yanaşı, üzvlük nəticəsi kimi hücumların qarşısını almağa kömək edəcək. Eyni zamanda, süni intellekt sistemlərinin hüquqi tənzimlənməsi və standartlaşdırılması da gələcəkdə bu sahənin inkişafına ciddi təsir göstərəcək. Beynəlxalq səviyyədə qəbul olunacaq təhlükəsizlik standartları və etik çərçivələr həm texnologiyanın inkişafını istiqamətləndirəcək, həm də istifadəçi hüquqlarının qorunmasını təmin edəcək [5]. Bundan əlavə, gələcək tədqiqatlarda hibrid modellərin — yəni həm statistik, həm də simvolik yanaşmaları birləşdirən sistemlərin – inkişafı mühüm rol oynaya bilər. Bu cür modellər daha izah edilə bilən və daha davamlı qərar mexanizmlərinə malik ola bilər. Eyni zamanda kvant hesablama texnologiyalarının inkişafı ilə süni intellekt təhlükəsizliyində yeni imkanlar və eyni zamanda yeni risklər ortaya çıxacaq. Bu isə mövcud müdafiə mexanizmlərinin yenidən nəzərdən keçirilməsini zəruri edəcək. Yekun olaraq qeyd etmək olar ki, süni intellekt əsaslı sistemlərin təhlükəsizliyi dinamik və daim inkişaf edən bir sahədir. Mövcud müdafiə

yanaşmaları mühüm nailiyyətlər təqdim etsə də, gələcəkdə daha ağıllı, adaptiv və inteqrasiya olunmuş təhlükəsizlik mexanizmlərinə ehtiyac olacaqdır. Bu istiqamətdə aparılan elmi tədqiqatlar və praktiki tətbiqlər yalnız texnoloji inkişafı deyil, həm də cəmiyyətin bu texnologiyalara olan etibarını müəyyən edəcək əsas amillərdən biri olacaqdır.

ƏDƏBİYYAT

[1] C. Szegedy et al., "Intriguing properties of neural networks," in Proc. Int. Conf. Learn. Represent. (ICLR), Banff, AB, Canada, 2014, pp. 1–10.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015.

[3] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in IEEE Symp. Secur. Privacy (SP), San Jose, CA, USA, 2016, pp. 582–597.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models robust to adversarial attacks," in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, 2018.

[5] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," IEEE Access, vol. 6, pp. 14410–14430, Mar. 2018.

[6] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," ACM Comput. Surv., vol. 54, no. 5, pp. 1–36, Jun. 2021.

[7] H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," Int. J. Autom. Comput., vol. 17, no. 2, pp. 151–178, Apr. 2020.

[8] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," Expert Syst. Appl., vol. 238, Art. no. 121742, Mar. 2024.

[9] K. He, D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," IEEE Commun. Surveys Tuts., vol. 25, no. 1, pp. 538–566, Firstquarter 2023.

Artificial Intelligence-Based Adversarial Attacks and Cybersecurity Mechanisms Against Them

Isgandar Mammadov
Baku State University, Baku, Azerbaijan
mammadov.isgandar.nazim@bsu.edu.az

Abstract- Modern cybersecurity systems are increasingly based on artificial intelligence and machine learning models. However, these models themselves are vulnerable to a new type of attack adversarial attacks. These attacks cause the model to produce false results by manipulating the input data with minimal changes. The article analyzes the essence of adversarial attacks, their types, and artificial intelligence-based defense mechanisms against these attacks. At the same time, real application scenarios and future directions are discussed.

Keywords—adversarial attacks; artificial intelligence; security models; cybersecurity.