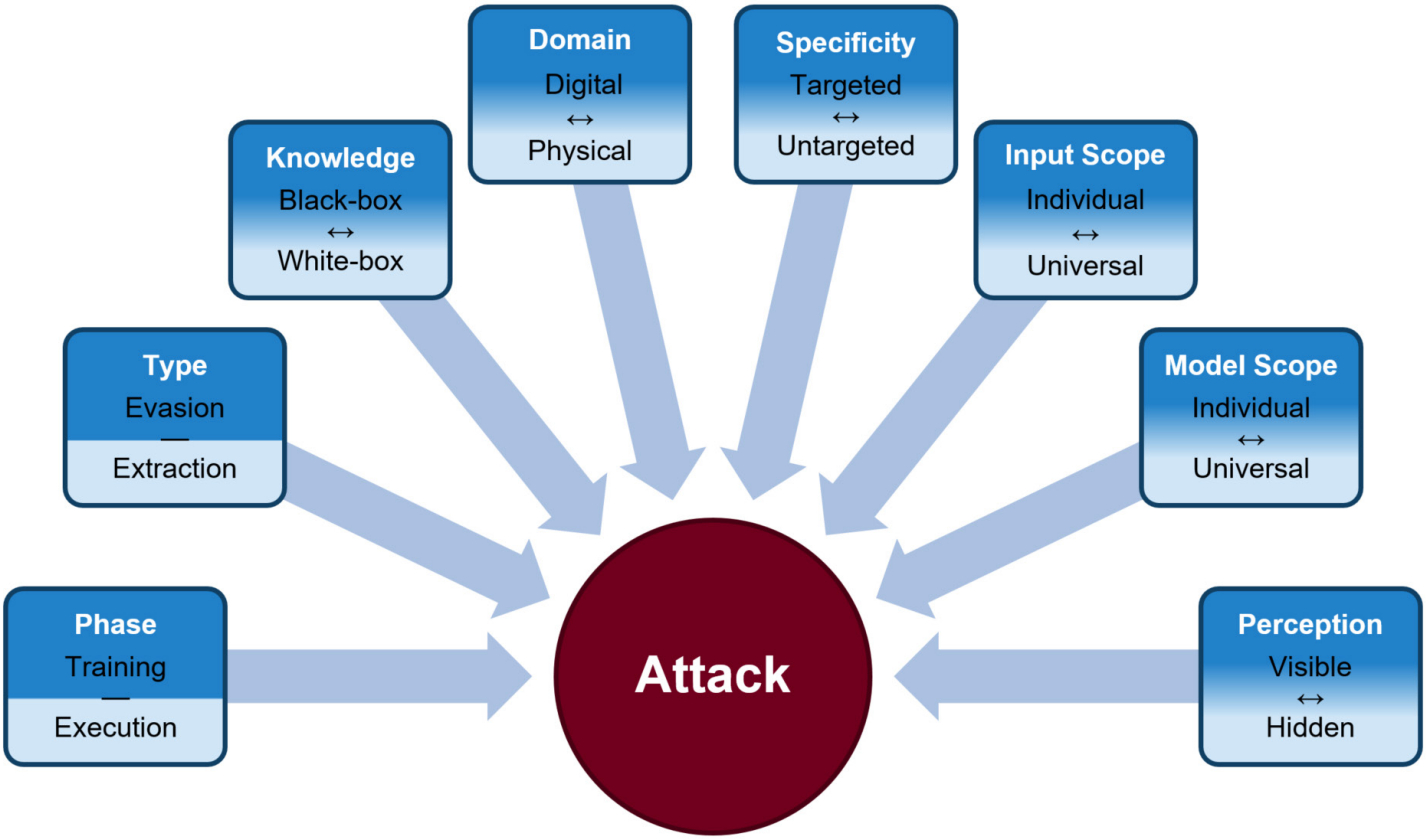
Hot- och sårbarhetsanalys av attacker mot AI i trådlösa kommunikationssystem

ERIK AXELL, ARWID KOMULAINEN, MARCUS KARLSSON,
PATRIK ELIARDSSON OCH ANDREAS ANDERSSON



FOI-R--5646--SE
ISSN 1650-1942 Oktober 2024

Erik Axell, Arwid Komulainen, Marcus Karlsson,
Patrik Eliardsson och Andreas Andersson

Hot- och sårbarhetsanalys av attacker mot AI i trådlösa kommunikationssystem

Titel	Hot- och sårbarhetsanalys av attacker mot AI i trådlösa kommunikationssystem
Title	Threat and Vulnerability Analysis of Attacks Against AI in Wireless Communication Systems
Rapportnr / Report No.	FOI-R--5646--SE
Månad / Month	Oktober / October
Utgivningsår / Year	2024
Antal sidor / Pages	33
ISSN	1650-1942
Kund / Customer	Försvarsmakten
Forskningsområde	Ledningsteknologi
FoT område	Ledning och MSI
Projektnr / Project No.	E51546
Godkänd av / Approved by	Christian Jönsson
Ansvarig avdelning	Telekrig

Sammanfattning

Införandet av AI, med tekniker baserade på maskininlärning (ML), i trådlösa kommunikationssystem ger nya möjligheter att effektivisera utnyttjandet av tillgängliga radioresurser, men riskerar också att införa nya typer av sårbarheter. För trådlösa kommunikationssystem är attacker och sårbarheter inte lika välstuderade som inom andra teknikområden, exempelvis bild- och språkbehandling. Området som studerar attacker mot ML-modeller benämns *adversarial machine learning* (AML). Begreppet AML brukar vanligtvis även inkludera försvar mot sådana attacker, men i denna rapport fokuserar vi på attackerna och hoten de utgör mot trådlösa kommunikationssystem.

Syftet med denna rapport är att analysera vilka typer av AML-attacker mot trådlösa kommunikationssystem som utgör realistiska hot samt vilka konsekvenser dessa kan få i kommunikationssystemen. Rapporten presenterar en översiktlig beskrivning av olika egenskaper som karakteriserar AML-attacker mot trådlös kommunikation, ger exempel på sådana attacker samt diskuterar deras realiserbarhet och påverkan. Målet är att bedöma vilka AML-attacker som utgör allvarliga hot och som därför kräver utveckling av motåtgärder.

En skillnad för trådlös kommunikation, jämfört med andra teknikområden, är att potentiella attacker sker fysiskt via radiogränssnittet. Detta gör att AML-attacker mot trådlösa kommunikationssystem måste ta hänsyn till de effekter som radiokanalen innebär. Radiokanalen ökar komplexiteten för attacken vilket inte är fallet för andra tillämpningar och flera studier försummar radiokanalens inverkan.

AML-attacker kan även utnyttjas av kommunikationssystem för att påverka motståndare, exempelvis genom medveten falsksignalering för att försvåra ML-baserad signalspaning.

Utifrån de publikationer som analyserats i denna studie går det inte att avfärda AML-attacker som ett hot inom trådlös kommunikation; det finns exempel inom flera av de undersökta områdena som vi bedömer är realiserbara. Dessvärre saknar majoriteten av arbetena realistiska kanalaspekter, många utgår från förenklade kommunikationsscenarier och saknar ofta kvalitativa jämförelser med traditionella brusstörningar. Det är därmed svårt att avgöra om AML-attacker inom trådlös kommunikation är ett allvarligt hot. Kunskapen om vilka egenskaper som krävs för att realisera AML-attacker inom trådlös kommunikation utgör dock en bra grund för framtida sårbarhetsanalyser.

Nyckelord: AML, AI, maskininlärning, radiokommunikation, störning

Abstract

The introduction of AI, with techniques based on machine learning (ML), in wireless communication systems provides new opportunities to make more efficient use of available radio resources, but also risks introducing new types of vulnerabilities. For wireless communication systems, attacks and vulnerabilities are not as well studied as in other technology areas, such as image and language processing.

The field that studies attacks against ML models is called adversarial machine learning (AML). The term AML usually also includes defense methods against such attacks, but in this report we focus on the attacks and the threats they constitute to wireless communication systems.

The purpose of this report is to analyze the types of AML attacks against wireless communication systems that pose realistic threats and the consequences they can have on communication systems. The report presents an overview of the different features that characterize AML attacks against wireless communications, provides examples of such attacks and discusses their realizability and impact. The aim is to assess which AML attacks constitute serious threats and therefore require the development of countermeasures.

One difference for wireless communication, compared to other technologies, is that potential attacks occur physically via the radio interface. This means that AML attacks against wireless communication systems must take into account the effects of the radio channel. The radio channel increases the complexity of the attack which is not the case for other applications and several studies neglect the impact of the radio channel.

AML attacks can also be used by communication systems to impact on adversaries, for example through deliberate false signaling to make ML based signals intelligence more difficult.

Based on the publications analyzed in this study, it is not possible to dismiss AML attacks as a threat in wireless communications; there are examples in several of the areas examined that we believe are feasible. On the other hand the majority of works lack realistic channel aspects, many are based on simplified communication scenarios and often lack qualitative comparisons with traditional jamming. It is thus difficult to determine whether AML attacks in wireless communications are a serious threat. Knowledge of the characteristics required to realize AML attacks in wireless communications, however, provides a good basis for future vulnerability assessments.

Keywords: AML, AI, machine learning, radio communication, jamming

Innehållsförteckning

1 Inledning

Införandet av AI, med tekniker baserade på maskininlärning (ML), i trådlösa kommunikationssystem ger nya möjligheter att effektivisera utnyttjandet av tillgängliga radioresurser, men riskerar också att medföra nya sårbarheter. En översikt över användning av ML i trådlösa kommunikationssystem ges i [1]. Attacker mot ML-baserad teknik har studerats inom andra teknikområden, exempelvis bild- och språkbehandling [2, 3]. För trådlösa kommunikationssystem är dessa attacker och sårbarheter inte lika välstuderade, även om det finns ett fåtal studier inom området [4, 5].

I denna rapport används begreppet ML-modell, i enlighet med vetenskaplig litteratur inom AI-området, för att beteckna en tränad ML-arkitektur (och dess parametrar) som kan utföra någon form av inferens. Detta ska inte förväxlas med en modell av ett kommunikationssystem som används för exempelvis utveckling och utvärdering av algoritmer. I rapporten används även begreppen ML-attack eller bara attack, vilka avser attacker som är riktade mot en (eller flera) ML-modell(er). Sådana attacker brukar benämnas *adversarial machine learning* attack eller förkortat AML attack. Begreppet AML inkluderar även försvar mot ML-attacker, men i denna rapport fokuserar vi på attackerna och hoten de utgör mot trådlösa kommunikationssystem.

I rapporten ges exempel på AML-attacker som beskrivs i vetenskaplig litteratur inom trådlös kommunikation. Attackerna baseras på att offret använder ML-modeller för den aktuella tillämpningen. I detta arbete beskrivs attackerna under förutsättning att ML används, men det görs ingen ytterligare analys av huruvida ML är den bästa tekniken för att lösa varje given uppgift.

Syftet med denna rapport är att analysera vilka typer av ML-attacker mot trådlösa kommunikationssystem som utgör realistiska hot samt vilka konsekvenser dessa kan få. Rapporten presenterar en översiktlig beskrivning av olika egenskaper som karakteriserar ML-attacker, ger exempel på ML-attacker mot kommunikationssystem samt diskuterar deras realiserbarhet och påverkan. Målet är att bedöma vilka attacker som utgör allvarliga hot och som därför kräver utveckling av motåtgärder.

2 Översikt av attackegenskaper

I detta kapitel beskrivs egenskaper som kan tillskrivas AML-attacker samt deras relevans för attacker mot ML-modeller i trådlösa kommunikationssystem. En kategorisering av AML-attacker görs baserat på dessa egenskaper. Syftet med kategoriseringen är att underlätta hot- och sårbarhetsanalys genom att att identifiera hur attackers olika egenskaper påverkar exempelvis realiserbarhet och sårbarhet.

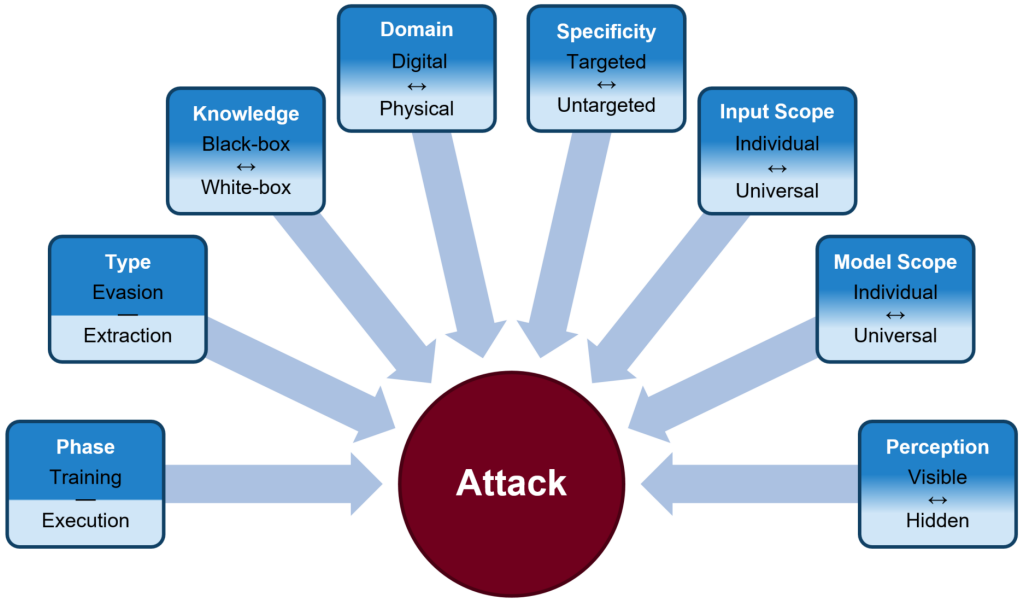
Med vår kategorisering av attackegenskaper kan AML-attacker beskrivas av samtliga egenskaper oberoende av varandra, dvs. att alla kombinationer av egenskaper är möjliga utfall och att egenskapernas definitioner inte överlappar. Några egenskaper, som i denna rapport beskrivs separat, grupperas ihop till en gemensam attackegenskap i [4]. Kategoriseringen i [4] gör därför att det finns beroenden mellan de beskrivna egenskaperna. För att göra en tydligare beskrivning av attacker baserat på egenskaper vilka kan beskrivas oberoende av varandra, görs därför en förfinad kategorisering i denna rapport, med exempel på vad de innebär i kontexten trådlös kommunikation.

Kategoriseringen av attackers egenskaper i denna rapport utgår från [3] och [4], men med vissa modifieringar för att bättre beskriva AML-attacker inom trådlös kommunikation. I [3] presenterades en översikt av egenskaper för ML-attacker inom andra teknikområden än trådlös kommunikation. Ett sätt att kategorisera attacker mot ML-modeller i trådlösa kommunikationssystem finns presenterat i [4]. Denna kategorisering skiljer sig något från [3]. I [4] kategoriserar författarna ML-attacker mot trådlösa system enligt tre egenskaper: typ av attack, mål med attack och kunskap för att utföra attacken. I artikeln diskuteras även vad som särskiljer ML-attacker mot trådlösa kommunikationssystem från attacker inom andra områden. De särskiljande faktorer som nämns är radiokanalens egenskaper, indirekt åtkomst till tränings- och testdata samt en stor spridning i vilka egenskaper hos radiokommunikationssystemen som de neurala näten behöver lära sig [4]. I [5] ges en sammanställning av attacker mot ML-modeller i kommunikationssystem, men med fokus på specifika tekniker för att utforma attacker snarare än att kategorisera egenskaper hos attacker. Dessutom ges i [5] en översikt av specifika försvarsalgoritmer mot sådana attacker.

Flertalet av de begrepp och definitioner som vi väljer att använda i denna rapport återfinns även i [3, 4]. Eftersom de attackegenskaper som diskuteras i [3] kommer från andra teknikområden har kategoriseringen av egenskaper i denna rapport modifierats något för att bättre passa för trådlös kommunikation. Även kategoriseringen i [4] har sitt ursprung i andra teknikområden, men med förklarande exempel på vad olika egenskaper kan innebära för trådlös kommunikation.

Vi använder några av de begrepp som är etablerade på engelska för att benämningar på egenskaper hos attacker ska vara samstämmiga med existerande vetenskaplig litteratur. Vi bedömer att svenska översättningar av vissa av egenskaperna kan skapa mer förvirring av begreppen än det underlättar förståelsen.

Figur 2.1 illustrerar den kategorisering av attackegenskaper som används i denna rapport, nämligen *Phase*, *Type*, *Knowledge*, *Domain*, *Specificity*, *Input Scope*, *Model Scope* och *Perception*. Två av egenskaperna kan anta diskreta värden, det vill säga en attacks *Phase* kan beskrivas som *Training* eller *Execution* och egenskapen *Type* kan vara *Evasion* eller *Extraction*. Övriga kategorier är egenskaper som kan beskrivas på en kontinuerlig skala. Alla dessa egenskaper definieras och förklaras mer utförligt i kommande avsnitt. I kapitel 3 beskrivs sedan exempel på AML-attacker i trådlös kommunikation och hur de kan kategoriseras av dessa egenskaper.



Figur 2.1: Kategorisering av olika typer av attackegenskaper.

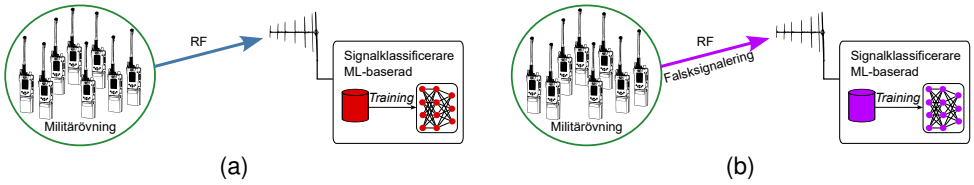
2.1 Phase

En ML-attack kan genomföras i två olika faser: *träningsfasen* (eng. training phase) och *exekveringsfasen* (eng. execution phase, ibland även kallat test phase eller inference phase). I [4] används även termen *trojan* om attacker som sker i både träningsfasen och exekveringsfasen. Vid en sådan attack införs bakdörrar i ML-modellen under träningsfasen som sedan utnyttjas under exekveringsfasen.

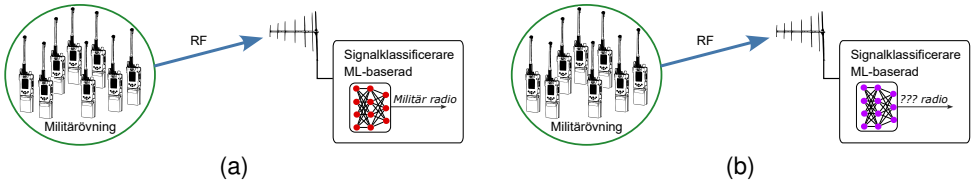
Attacker under exekveringsfasen bygger generellt på att indata till ML-modellen modifieras genom att en störning adderas. Störningen är designad för att påverka ML-modellens prediktion och kan bygga på sårbarheter som införts under träningsfasen.

Attacker under träningsfasen bygger på att träningsdata manipuleras för att införa modellfel, vilka orsakar försämrad prestanda eller sårbarheter i ML-modellen som kan attackeras senare. Sådana attacker benämns *förgiftning* av data (eng. poisoning eller causative attack). Genom en attack under träningsfasen är det möjligt att påverka en ML-modell så att den lär sig felaktiga mönster. Manipulering av träningsdata kräver att angriparen har möjlighet att påverka träningsdata. Inom trådlös kommunikation finns flertalet studier om ML-modeller för modulationsklassificering. Det finns publikt tillgängliga dataset för att träna sådana ML-modeller. Ett sätt att undvika korrekt klassificering av den egna modulationstypen är att införa förgiftade träningsdata i publika dataset.

Ett alternativt scenario för genomförande av en attack under träningsfasen kan vara att signaler i luften från ett radiokommunikationssystem spelas in av motståndaren, exempelvis vid militära övningar, i syfte att skapa träningsdata för klassificeringsalgoritmer. I figur 2.2a spelar en motståndare in radiosignaler vid en militärövning med syfte att träna sin signalklassificerare. Ett system eller en operatör som är medvetna om att data samlas in av en motståndare kan då manipulera sina signaler under övningen i syfte att förgifta träningsdata hos motståndaren, se figur 2.2b. För att veta hur signalerna ska modifieras för att effektivt skapa förgiftade träningsdata hos motståndaren krävs förmodligen ingående kunskap om den ML-modell som ska tränas. Under exekveringsfasen används den tränade modellen för att klassificera radiosignalen, se figur 2.3, där den förgiftade modellen kommer göra felaktiga klassificeringar. I exemplet ovan har såle-



Figur 2.2: Inspelning av träningsdata under militär övning. Inspelade data används i träningsfasen för att träna en signalklassificerare. I (a) samlas korrekta data in medan i (b) görs medveten falsksignalering med syfte att förgifta träningsdata.



Figur 2.3: Vid exekveringsfasen klassificerar den tränade modellen radiosignalen. I (a) görs en korrekt klassificering medan i (b) misslyckas klassificering på grund av att modellen tränats på felaktiga (förgiftade) data.

des kommunikatören attackerat motståndaren under träningsfasen för att motståndarens klassificering ska misslyckas i exekveringsfasen.

I fallet att Försvarsmakten själv utvecklar ML-modeller finns möjligheten att kontrollera träningsdata, men då produkter köps in av tillverkare som i sin tur nyttjar ML-modeller är det svårare att kontrollera att träningsdata inte är manipulerade. Om träningsdata kontrolleras helt av den som tränar sin modell, exempelvis om data kan genereras syntetiskt, är risken för förgiftade träningsdata liten. Principer för att skapa träningsdata med god kvalitet bör dock fortfarande följas för att garantera robust kommunikation [6, Kap. 2].

Sammantaget bedömer vi att attacker under både träningsfasen och exekveringsfasen är realiserbara i kontexten trådlös kommunikation. Genom att själv skapa eller samla in träningsdata istället för att använda publika dataset minskar risken för förgiftning av data. Attacker under exekveringsfasen är troligt den vanligaste attackytan, dock försvåras dessa attacker av radiokanalens egenskaper jämfört med motsvarande attacker inom andra teknikområden.

2.2 Type

I denna rapport används begreppet *typ* (eng. type) för att skilja på två olika typer av attacker: *evasion* som syftar till att en ML-modell ska fatta felaktiga beslut och *extraction* som syftar till att extrahera data från en ML-modell. Inom befintlig litteratur ges egenskapen *typ* något olika betydelse och det varierar mellan publikationer vad som inkluderas i begreppet. I [3] nämns följande attacktyper: poisoning, evasion och extraction; poisoning betecknas därmed som en egen attack-typ i [3] medan vi använder begreppet för att beskriva en attack som utförs under träningsfasen, se avsnitt 2.1.

Begreppet evasion ges olika betydelse i olika publikationer: i vissa innefattar det endast förmågan att få en ML-modell att fatta ett felaktigt beslut [4], i andra publikationer är det även inkluderat i begreppet att attacken ska ske på ett sätt som är svårt att upptäcka [3]. Vi väljer här att i definitionen av en *typ* av attack enbart inkludera påverkan av en ML-modells förmåga till att fatta korrekta beslut. Huruvida en attack är svår att upptäcka eller ej placeras istället under kategorin *perception*, se avsnitt 2.8.

ML-attacker som är av typen evasion angriper ML-modeller exempelvis genom att

manipulera data under exekveringsfasen. Det kan exempelvis innebära AML-baserad störning mot trådlösa kommunikationssystem som nyttjar ML-modeller för mottagning. Syftet med attacken är att manipulera data på ett sätt som lurar en ML-modell att fatta ett felaktigt beslut. Manipulering sker typiskt genom att en störning adderas till signalen som agerar indata till ML-modellen. En kommunikationstillämpning av ML som studerats mycket är modulationsklassificering: att utgående från en mottagen samplad signal avgöra vilken modulationsform som använts för att generera signalen. För exemplet med en modulationsklassificerare kan en evasion-attack bestå i att modifiera sin utsända signal för att undvika klassificering av en redan tränad ML-modell. För att designa en effektiv sådan attack krävs relativt stor kunskap om den ML-modell som skall attackeras. Av det som publicerats kring AML mot trådlösa system kan en stor del klassas som attacker av typen evasion, exempelvis tidigare nämnda attacker mot modulationsklassificering [4].

Stora datamängder krävs för att träna ML-modeller och för en angripare är åtkomst till dessa data eftertraktansvärt av flera anledningar. I ett realistiskt scenario är det troligt att angriparen inte har tillgång till offrets ML-modell utan behöver träna en egen så kallad skuggmodell som imiterar offrets ML-modell. Attacker som syftar till att utvinna information om en ML-modell kallas vanligtvis *extraction attack* eller *exploratory attack*. Genom att extrahera data från ML-modellen som ska attackeras kan en bättre skuggmodell konstrueras. Skuggmodellen kan senare användas för att utveckla attacker mot offrets ML-modell. Exempel på hur detta kan genomföras mot språkmodeller ges i [3]. Hur effektiva attacker framtagna med hjälp av en skuggmodell är mot det tänkta offret benämns *överförbarhet* (eng. transferability).

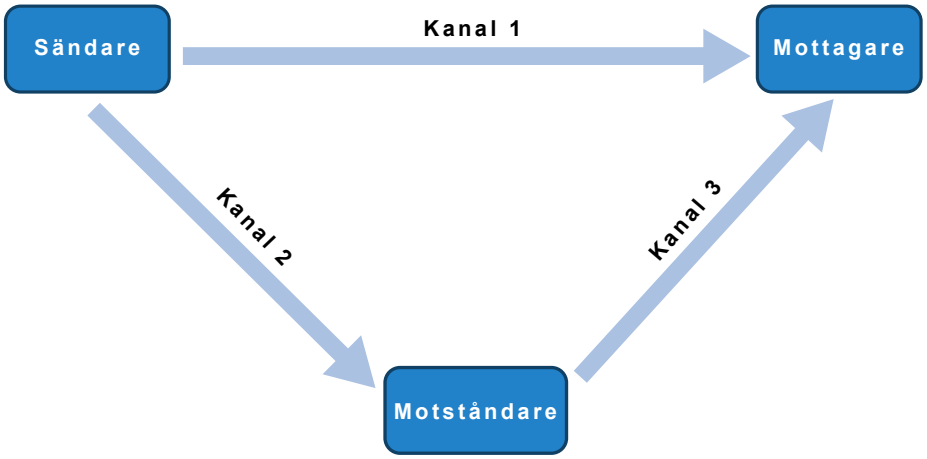
Det är även troligt att ML-modeller som tas fram för att ersätta funktioner i trådlösa kommunikationssystem, exempelvis modulation och kodning, störskydd och kanalaccess, behöver tränas på ej öppna data och parametersättas med parametrar som omfattas av sekretess. Risk för informationsläckage vid användning av ML-modeller för trådlösa kommunikationssystem tränade på hemliga eller känsliga data är något som bör beaktas.

Både evasion och extraction är realiserbara och i ett realistiskt scenario behöver i många fall en angripare utföra båda attackerna: en extraction-attack i syfte att bygga upp en skuggmodell som sedan kan användas för att konstruera en evasion-attack anpassad mot den ML-modell som ska attackeras. Den trådlösa radiokanalens egenskaper gör båda typerna av attacker svårare att genomföra. Indata till skuggmodellen påverkas av radiokanalen på ett svårförutsägbart vis vilket leder till modellfel kontra offrets ML-modell. På samma sätt är det svårt att förutse hur en evasion-attack påverkas av radiokanalen till mottagaren.

2.3 Knowledge

Inom denna kontext syftar *kunskap* (eng. knowledge) på hur mycket angriparen känner till om det attackerade systemet, i synnerhet sådan kunskap som kan främja allvarlighetsgraden av attacken och svårigheten för en angripare att designa attacken.

Kunskap delas i detta sammanhang upp i tre nivåer: *black-box*, *white-box* och *gray-box*. En black-box-attack innebär att angriparen inte har någon insikt i det anfallna systemet i sig, bortsett från möjligheten att observera indata och undersöka vad effekten av dessa är. I andra änden av skalan existerar white-box-attacker, i vilka angriparen har fullständig kunskap om det attackerade systemet. Då detta handlar om kunskap i systemets digitala domän så innebär white-box att angriparen har komplett kunskap om offrets ML-modell såsom modellarkitektur, vikter och hyperparametrar. White-box-attacker används ofta i den vetenskapliga litteraturen för att betrakta det värsta fallet av hur robust



Figur 2.4: Illustration av hur radiokanalen påverkar möjligheten att utföra AML mot trådlös kommunikation. En motståndare känner inte till kanalen mellan sändare och mottagare och de två andra kanalerna behöver estimeras. I ett mobilt scenario varierar därtill kanalerna dynamiskt över tid.

ett system är mot attacker. Gray-box-attacker befinner sig någonstans mellan white- och black-box-attacker och innebär att angriparen har delvis kunskap om systemet. Det kan exempelvis vara vetskap om ett neuralt nätverks arkitektur men inte värdet på vikterna. En gray-box-attack kan utspela sig så att angriparen försöker emulera hur den förväntar sig att offrets faktiska system ser ut via en skuggmodell, för att sedan använda denna emulering som mål för att skapa en white-box-attack och nyttja denna attack mot det systemet som ursprungligen var målet för attacken.

En ren white-box-attack bedöms inte vara realiserbar i praktiken och behandlas därför inte vidare i denna rapport. Att observera in- och utdata för att genomföra en black-box-attack mot ett trådlöst kommunikationssystem bedöms vara realiserbart i många fall. Att ha delvis information baserat på t.ex. kända principer, information via andra källor eller genom att extrahera information via observation av in- och utdata, bedöms också vara realiserbart i vissa fall.

2.4 Domain

Egenskapen *domain* beskriver i vilket gränssnitt en attack sker mot ett trådlöst kommunikationssystem. En attack via radiogränssnittet, dvs. via radiovågor, benämns som en *fysisk* (eng. physical) attack. Den andra typen av attack som är möjlig är en *digital* attack.

En fysisk attack för trådlös kommunikation sker via radiogränssnittet. En angripare sänder således en attacksignal på samma bärvågsfrekvens som radiomottagaren är konfigurerad till. Attacksignalen påverkas därmed av radiovågens utbredning mellan angriparen och radiomottagaren. Dämpningen från vågutbredningen av attack-signalen gör att angriparen måste befinna sig inom ett givet avstånd från radiomottagaren för att ha möjlighet till att påverka radiomottagaren. Avståndet beror bland annat på attacksignalens uteffekt, vågutbredningen och antennegenskaper hos angriparen och radiomottagaren. Figur 2.4 visar en illustration av en radiosändare, radiomottagare och en motståndare samt de individuella kanalerna mellan dem. Oftast ses motståndaren som angriparen då den försöker påverka radiomottagaren, men det kan också förekomma fall där radiosändaren angriper motståndaren. Detta kan ske om motståndaren lyssnar på radiotrafiken. Då kan sändaren medvetet angripa motståndarens ML-modell och då är motståndaren att betrakta som ett offer för ML-attacken.

En digital attack sker genom att direkt påverka indata till en ML-modell, vilket in-

nebär att angripa kommunikationssystemet inifrån. Det kan vara genom att påverka träningen av ML-modellen eller påverka signalbehandlingen genom skadlig kod. En digital attack för ett trådlöst kommunikationssystem är att betrakta som en cyberattack. För trådlösa kommunikationssystem bedöms en digital attack som osannolik medan en fysisk attack som mer sannolik.

2.5 Specificity

Specificity avser hur riktat resultatet av en attack är, dvs. om attacken har som mål att ge ett specifikt resultat eller om målet är att resultatet bara ska avvika från det korrekta. En *specifik* (eng. targeted) attack mot exempelvis modulationsklassificering är avsedd att påverka klassificeringsresultatet att ange ett specifikt modulationsformat. En attack som är *ospecifik* (eng. untargeted) har som mål att klassificeringen ska ange ett felaktigt modulationsformat, men utan att ta hänsyn till vilket.

Attacker mot trådlösa kommunikationssystem kan vara både specifik eller ospecifik, beroende på vad syftet är med attacken och vilken kunskap den som utför attacken har om kommunikationssystemet. I många fall krävs mer kunskap om offret för att utföra en targeted attack. I princip kan traditionell störning med exempelvis vitt Gaussiskt brus betraktas som en untargeted attack, vilken därmed kan utföras utan någon kunskap om offret.

Med känd information eller antaganden om offrets ML-modell kan attacker utformas inom hela intervallet av egenskapen specificity, dvs. från specifik till ospecifik. Möjligheten att realisera attacker, med avseende på specificity, faller ofta tillbaka på vilken kunskap som krävs om offrets ML-modell för att utföra attacken. Realiserbarheten har därför ett beroende till egenskapen *knowledge*.

2.6 Input Scope

Input scope för en attack eller störning beskriver hur riktad en attack är med hänsyn till insignalen. Om attacken är unikt framtagen för varje insignal kallas den *individuell* (eng. individual). Om en attack konstrueras baserat på en mängd insignaler och används mot flera insignaler kallas den *universell* (eng. universal). Målet med attacken, oavsett omfattning, är att med en liten förändring av insignalerna öka sannolikheten att offret gör fel.

För individuella attacker skapas attacksignalerna genom att analysera insignalerna separat och varje insignal kan därför få en unik attacksignal. Det gör individuella attacker potentiellt mer kraftfulla eftersom varje attacksignal är anpassad till en specifik insignal. Ett exempel på individuella attacker från bildanalys ges i [7], där attacksignalen skapas genom att hitta den minsta förskjutningen av bilden som gör att bilden felklassificeras.

Vid en universell attack är varje attacksignal baserat på en mängd olika insignaler och varje attacksignal kan appliceras på flera insignaler. I extremfallet kan det handla om en enda attacksignal som appliceras på samtliga insignaler [8, 9]. Attacksignalen kan konstrueras på olika sätt, beroende på kunskap (se avsnitt 2.3). I [8] skapas attacksignalen genom att ackumulera attacksignaler för individuella sampel ur en delmängd av träningsdata. En fördel med universella attacker är att angriparen inte behöver någon vetskap om nuvarande insignal eftersom attacksignalen är skapad för en godtycklig insignal. Universella attacker kan även konstrueras på ett sådant sätt som gör attacksignalen invariant mot fasskift vilket gör att angriparen och offret inte behöver vara faskoherenta för att attacken ska vara effektiv[9].

Både indivuduella och universella attacker är generaliserbara mellan träningsmäng-

der. I det universella fallet innebär detta att den framtagna attacksignalen lyckas lura offret med hög sannolikhet även för sampel utanför den delmängd som användes för att skapa attacksignalen [8]. För individuella attacker visar [7] att attacksignaler skapade för en modell tränad med en viss datamängd även fungerar om modellen tränas på en disjunkt datamängd.

Både individuella och universella attacker anses vara realiserbara men individuella attacker ställer högre krav på angriparen. För en individuell attack kan insignalen estimeras eller predikteras för att därefter användas till att skapa en attack. Att kunna estimera utan fördröjning eller prediktera felfritt kan dock anses vara orimligt, men detta faller inom egenskapen *domain* i avsnitt 2.4. Universella attacker är mer intuitivt realiserbara: attacken konstrueras baserat på en förbestämd mängd sampel och används sedan på godtycklig insignal.

2.7 Model Scope

Model scope för en attack beskriver hur riktad en attack är med hänsyn till ML-modellen. En attack som är unikt framtagen för att fungera väl på en ML-modell kallas *individuell* och en attack som tar hänsyn till flera ML-modeller kallas *universell*. Notera att egenskapen model scope inte säger något om egenskapen input scope. Vi kan ha en attack med individuell input scope men universell model scope och vice versa.

Författarna i [10] demonstrerar hur välkonstruerade attacker har hög *transferability*, dvs. att de har stor påverkan även då de appliceras på en ML-modell med annan arkitektur, givet att målet för ML-modellerna (i detta fall igenkänning av handskrivna siffror) är detsamma. Då attacker tenderar att generalisera väl mellan olika ML-modeller och olika träningsmängder kan attacksignaler konstrueras utan att känna till detaljer om offret, exempelvis genom att använda en skuggmodell, och fortfarande vara effektiva [10].

Model scope påverkar inte realiserbarheten i sig, då model scope handlar mer om vilka ML-modeller som tas i beaktning när attacken skapas och utvärderas. Kännedom av ML-modellerna när attacken konstrueras är mer en fråga om kunskap som diskuteras i avsnitt 2.3.

2.8 Perception

I [3] definieras *perception* som egenskapen att en attack kan upptäckas av en människa eller ej. I detta arbete utökar vi detta begrepp till att innefatta hur lätt en attack är att upptäcka, oavsett om det görs av en människa eller av tekniska lösningar. Perception kan betraktas som en kontinuerlig skala där ändpunkterna delas in *synlig* (eng. visible) eller *dold* (eng. hidden).

Attacker mot trådlös kommunikation kan tänkas befinna sig i stora delar av denna skala. Exempelvis kan syftet med en störattack mot en neural nätverks-baserad mottagare enbart vara riktad mot att förstöra möjligheten för offret att kommunicera, utan att ta hänsyn till om attacken är lätt att upptäcka. I andra fall, exempelvis för att vilseleda en modulationsklassificerare, kan det finnas incitament att en pågående attack ska vara svår att upptäcka.

Var gränserna går för när en attacks bedöms vara synlig eller dold är komplext att avgöra. Flertalet publikationer fokuserar på attackens förmåga att åstadkomma fel och perceptions-aspekten studeras inte alls. Ett vanligt angreppssätt för att försvåra upptäckt är att begränsa sändareffekt för attacken [11, 12]. I några studier utvärderas prestanda för att upptäcka attacken och anges t.ex. i form av kvoten mellan attackens effekt och bruseffekt [9]. För att fullständigt karakterisera perception krävs dock ytterligare stor-

heter, exempelvis detektionssannolikhet och falsklarmssannolikhet. Även om en sådan fullständig karaktärisering kan göras så är det nästintill omöjligt att avgöra när en attack kan betraktas som ändlägena visible respektive hidden.

Eftersom karakteriseringen av perception är komplex så är det också svårt att bedöma realiserbarheten av denna egenskap i generella fall. I vissa fall saknar perception helt relevans, medan i andra fall kan det vara helt avgörande att attacken inte kan upptäckas för att förhindra motåtgärder. Realiserbarheten bör därför analyseras beroende på tillämpning och i vilka scenarier attacker kan förväntas. För trådlös kommunikation har t.ex. offrets och angriparens positioner i förhållande till varandra stor betydelse för om en attack kan upptäckas eller ej.

2.9 Attacker över tid

De egenskaper för attacker som har beskrivits hittills utgår från en enskild, momentan attack. För att kontinuerligt påverka ett kommunikationssystem behöver attacker utföras upprepade gånger, eventuellt genom anpassning, över tid. De egenskaper som har beskrivits tidigare gäller alltså för varje enskild attack under ett givet tidsintervall.

Vissa attacker som pågår över tid kan beskrivas som spelteoretiska problem där två motståndare duellerar med olika mål. Sådana attacker tenderar att omfatta båda typerna *evasion* och *extraction* samt utföras under både träningsfasen och exekveringsfasen.

En teknik som ofta används för att kontinuerligt uppdatera och anpassa ML-modeller är så kallad *förstärkningsinlärning* (eng. reinforcement learning, RL). För RL-baserade dueller görs en skillnad mellan om det går att påverka målets observationer direkt eller inte. I [13, 14] bedöms det orelistiskt att påverka målets observationer direkt i praktiska system, så en alternativ taktik är att utveckla en policy som gör att den tränade agenten inte vet vad den ska göra.

Vid dessa duellsituationer, där ML-modellen tränas med återkoppling beroende på sitt agerande, påverkas både den egna och motståndarens ML-modeller av varandras beslut och agerande. Sårbarheten och konsekvensen av en sådan duell är svår att förutsäga, men att vinna duellen handlar snarast om att överlista motståndaren.

Hur återkopplingen ska ske är en utmaning för realisering av attacker i en duellsituation. För att få återkoppling måste offrets beslut eller agerande mätas via radiogränssnittet, vilket dessutom tar radioresurser. I vissa tillämpningar är det svårt att få återkoppling och därmed att realisera denna typ av adaptiv attack, exempelvis vid ML-attacker mot en passiv signalspanare som inte inte sänder.

3 Attacker och påverkan på kommunikationssystem

I detta kapitel ges en sammanställning av exempel på attacker mot ML i kommunikationssystem. Attackerna kategoriseras även enligt de egenskaper som presenterades i kapitel 2. Sammanställningen har grupperats inom tre delområden; spektrumavkänning, modulationsklassificering och radioresursallokering.

Användningen av ML inom trådlös kommunikation var tidigt ute inom just spektrumavkänning och modulationsklassificering. Detta är områden som baseras på detektion och klassificering, vilka är naturliga tillämpningar av ML. Därför finns en relativt stor mängd forskningslitteratur, även med studier av attacker, inom just dessa områden. Området radioresursallokering innefattar egentligen en betydligt större bredd av tillämpningar för effektivt utnyttjande av radiokommunikation. Eftersom ML för trådlös kommunikation är ett relativt ungt forskningsområde, så finns relativt lite forskningslitteratur om attacker mot radioresursallokering.

I sammanställningen av ML-attacker görs en något förenklad bedömning av attackerna uttryckt i ytterligheterna av respektive kategori. Sammanställningen innehåller enbart attacker som genomförs via radiogränssnittet och som vi bedömer är rimliga att genomföra i praktisk tillämpning för trådlös kommunikation. Kategorierna *domain* och *perception* redovisas därför inte av följande skäl:

- En digital attack bedöms inte vara praktiskt realiserbar inom trådlös kommunikation. I den mån en antagonist kan direkt påverka indata till en ML-modell så är det snarare att betrakta som en cyberattack, vilken i grunden genomförs via andra sårbarheter i systemet. Alla attacker i sammanställningen inkluderar en radiokanal och är, i varierande grad, fysiska attacker.

- Exakt var gränserna går för synlig respektive dold i kategorin *perception* är komplext att bedöma, vilket diskuterades i 2.8. Även om en fullständig karaktärisering kan göras så är det nästintill omöjligt att avgöra när en attack kan betraktas som synlig respektive dold. Vi bedömer därför att en förenklad bedömning av attacker i termer av ändlägena inte är meningsfull och därför utelämnas denna kategori i den översiktliga sammanställningen.

3.1 Attacker mot spektrumavkänning

Spektrumavkänning är en funktion för att estimera om spektrum eller en specifik kanal används för tillfället eller ej. Detta är en viktig funktion i radiosystem som har förmågan att dynamiskt välja vilken kanal och bandbredd som för tillfället är mest effektivt att utnyttja. Maskininlärning har under de senaste åren använts för att avgöra om en kanal är ledig eller ej [15]. Detta kommer inte utan risker för att dessa algoritmer kan attackeras. Nedan ges några exempel på sådana attacker. Tabell 3.1 visar en sammanställning av dessa exempel och hur de kan karakteriseras enligt egenskaperna i avsnitt 2.

Ett scenario som studeras i [16] är en primär sändare, en sekundär användare som vill använda kanalen när den inte används av den primära användaren samt en angripare som vill lura den sekundära användaren att sända när kanalen är upptagen. Den sekundära användaren antas använda ett neuralt nätverk för att estimera om kanalen är ledig och angriparen tränar upp en skuggmodell med motsvarande syfte. Attacken utförs med så kallad *maximum received perturbation power* (MRPP) med målet att den sekundära användaren ska felaktigt klassificera en pågående sändning som en outnyttjad kanal. Förutsättningarna är att olika kanaler antas mellan de tre involverade parterna (Figur 2.4).

På grund av att kanalerna är olika tränas motståndarens skuggmodell på andra sampel än den sekundära användarens. För att designa den störning som ska sändas utgår angriparen från den mottagna signalen, dock antas kanalen mellan angripare och sekundär användare vara känd. Resultaten visar på vilka effekter radiokanalen har på prestandan, framförallt i form av minskad framgång med attackerna då exakta indata inte är kända.

I [17, 18] beskrivs tre attacker. Den första attacken bygger en skuggmodell av kommunikationssändarens prediktion av när spektrum är ledigt. Skuggmodellen byggs av motståndaren genom att detektera sändningar och lyssna på om *ack* eller *nack* skickas från offret [19]. Den andra attacken är att störsändning sker enbart då kommunikationssändaren är i sin spektrumavkänningsfas, för att förhindra att den sedan ska fatta beslut om att sända. Det krävs mindre energi för att lura sändaren i sensingfasen än att störa ut meddelandet som sedan skickas. Den sista attacken sker då kommunikationssändaren utför omträning av sin ML-modell för spektrumavkänning. Genom att påverka omträningen behövs ingen störsändning utan sändaren fattar själv felaktiga beslut. En liknande attack som beskrivs i [18] genomförs även i [20]. I [20] utvärderas också en försvarsstrategi som innebär att kommunikationssändaren medvetet fattar felaktiga beslut för att på så viss påverka träningen av motståndarens skuggmodell. Med den beskrivna försvarsstrategin, som är en attack mot motståndarens skuggmodell under dess träningsfas, ökar motståndarens sannolikhet för felaktig detektion från 4% till 21% och felsannolikheten ökar från 18% till 25% [20].

Med förstärkningsinlärning kan RL-agenter tränas för att utföra dynamisk spektrumaccess. I [21] används en RL-agent för att välja ut en bra kanal bland flertalet kanaler. En attack skapas genom att låta en attack RL-agent träna sig på att detektera bra kanaler. Attackagenten belönas när den väljer en bra kanal och samma som offret och bestraffas om den väljer en annan kanal än offret.

En metod för att generera strukturerat brus för att lura en spektrumavkännare beskrivs [22]. Metoden kallas *embedded communication method* (ECM). Metoden baseras på en singelvärdesuppdelning av kommunikationssignalen för att nyttja den som strukturerat brus. Bruset placeras i stopbandet vilket gör det svårupptäckt. I utvärderingen av ECM får den större påverkan i fallet black-box än metoderna *basic iterative method* (BIM), *momentum iterative method* (MIM) and FGSM. I white-box fallet är resultatet det omvända. Att bruset döljs i stopbandet gör att det kan filtreras bort av en smalbandig mottagare men får en påverkan hos en bredbandig mottagare.

Tabell 3.1: Sammanfattning av attacker mot spektrumavkänning.

Attack	Type		Phase		Knowledge		Specificity		Input Scope		Model Scope		Ref.
	ev	ex	train	exec	black	gray	target	untarget	ind	univ	ind	univ	
MRPP, kanal-access	✓	✓	✓	✓	✓		✓		✓			✓	[16]
Lyssnar på *ack* och *nack*		✓		✓	✓			✓		✓		✓	[17, 18, 19, 20]
Prediktion av sändningsmönster	✓			✓		✓		✓		✓	✓		[17, 18]
Påverkar omträningen	✓		✓			✓		✓		✓		✓	[17, 18]
RL-baserad störning	✓	✓		✓	✓			✓		✓	✓		[21]
Störsignal överlagrad på kommunikationssignal	✓			✓	✓			✓	✓			✓	[22]

3.2 Attacker mot modulationsklassificering

Modulationsklassificering ämnar identifiera vilken typ av modulation en på förhand okänd signal har. Modulationsklassificering underlättar bland annat spektrumavkänning, identifiering av interferens och identifiering av fientliga signaler. De senaste åren har fokus skiftat från klassiska metoder som särdragsextrahering och likelihood-baserade test till ML-baserade metoder [23]. Nedan följer exempel på attacker som utnyttjar svagheter i dessa ML-baserade metoder. Tabell 3.2 visar en sammanställning av dessa exempel och hur de kan karakteriseras enligt egenskaperna i avsnitt 2.

I [24] definierar författarna en klass av problem de kallar *Generalized Wireless AML* som är olika typer av attacker mot en modulationsklassificerare. Både att modifiera en legitim användares signal för att den ska felklassificeras och att modifiera egen utsänd signal behandlas i samma ramverk. Attacken som presenteras bygger på att använda FIR-filter för att modifiera I/Q-sampel och kanaleffekter modelleras.

I [9] konstrueras en universell black-box-attack baserad på en delmängd av träningsdata med hjälp av *principal component analysis* (PCA). En attacksignal skapad på detta vis kan skada ett offer mer än en klassisk störsignal som endast består av Gaussiskt brus. Attacken i sig är digital vilket gör den icke realiserbar enligt diskussionen i början av avsnitt 3. Denna attack är, trots att den är digital, en byggsten för många andra attacker och används ofta som referens.

Attacken i [9] utökas i [25] från digital till fysisk genom att introducera en fädande kanal mellan angripare och offer. En legitim mottagare (offret) ska klassificera modulationstypen i signalen som skickas från en legitim sändare samtidigt som angriparen skickar signaler för förhindra korrekt klassificering. Angriparen har statistisk kännedom om kanalen till offret och använder detta och en skuggmodell för att konstruera individuella attacker för en mängd insignaler. Givet dessa attacker används PCA för att få fram en universell attack, kallad *universal adversarial perturbation* (UAP), som fungerar för godtycklig kanalrealisation och insignal.

En metod för att generera olika universella attacker presenteras i [11] där en *pertubation generator model* (PGM) producerar universella attacker som skickas över en AWGN[1]-kanal. Attackerna optimeras för att likna vitt Gaussiskt brus för att vara svåra att upptäcka och det euklidiska avståndet mellan olika attacker maximeras för att försvåra eventuellt försvar. Artikeln visar att det är förhållandevis lätt att bekämpa en enskild universell attack, men attacker från deras PGM är mer motståndskraftig.

I [12] presenteras en PGM som utvärderas i ett scenario med en verklig, fädande kanal och multimodal data (bild, video, tal, text). Angriparen antas kunna få statistisk kännedom om kanalen till offret genom utnyttja utskickade pilotsignaler. Attacken är konstruerad för att fungera oavsett vilken typ av data som skickas och utan att angriparen är synkroniserad med sändaren eller offret. Även här diskuteras försvarsmekanismer och experimenten visar att attacken är förhållandevis svår att försvara sig mot.

Attackerna FGSM (fast gradient sign method), PGD (projected gradient descent) och DeepFool utvärderas för en angripare som har olika mycket kunskap om offrets ML-modell [26]. Med full kunskap om offrets ML-modell, *white-box*, är attacken mest effektiv. Desto mindre kunskap som angripare har om offrets ML-modellen desto sämre verkan har attacken. Även attackens signalstyrka i förhållande till signalen och bruset utvärderas. Jämfört med att störa med AWGN behöver de studerade attackerna mindre effekt för att få samma verkan.

Störning av modulationsklassificering behandlas ofta som ett isolerat problem med en mottagare som ämnar klassificera en signal utskickad av en fientlig sändare. Sända-

[1] additive white Gaussian noise

rens primära mål är i detta fall att lura mottagaren, inte att överföra information som är det normala i ett kommunikationssystem. I ett mer realistiskt scenario är informationsöverföring det primära och att undvika att bli klassificerad av en fientlig mottagare är sekundärt. Denna utvidgning av problemet ger sändaren mindre utrymme att permutera signalen eftersom en alltför stor förändring påverkar informationsflödet negativt.

I [27] studeras hur en legitim sändare kan förhindra modulationsklassificering av en informationsbärande signal med hjälp av AML. Den legitima sändaren modifierar den informationsbärande signalen i syfte att förhindra en avlyssnande enhet att identifiera signalens modulation. Målet med att lura den avlyssnande enheten balanseras med att bibehålla förmågan att kommunicera över den legitima länken. I [28] undersöks hur felrättande koder påverkar prestanda och i [29] tar författarna hänsyn till signalens spektrala egenskaper för att försvåra för avlyssnaren att upptäcka att den informationsbärande signalen har modifierats.

Det ramverk som presenteras i [27, 28, 29] gör det möjligt att ta hänsyn till informationsöverföring, förmåga att vilseleda och förmåga att undvika upptäckt samtidigt. Arbetet är fortfarande i sin linda och attackerna baseras än så länge på information om offret som normalt sätt inte är tillgänglig för en angripare, vilket gör att samtliga attacker räknas som white-box.

Tabell 3.2: Sammanfattning av attacker mot modulationsklassificering.

Attack	Type		Phase		Knowledge		Specificity		Input Scope		Model Scope		Ref.
	ev	ex	train	exec	black	gray	target	untarget	ind	univ	ind	univ	
FIR-filter-baserad	✓			✓	✓		✓	✓		✓		✓	[24]
PCA, UAP	✓			✓	✓			✓		✓	✓		[25]
PGM	✓			✓	✓			✓		✓		✓	[11]
PGM	✓			✓	✓			✓		✓		✓	[12]
FGSM, PGD, DeepFool	✓			✓		✓		✓		✓		✓	[26]

3.3 Attacker mot radioresursallokering

I ett radionät med flera användare behöver kommunikationsresurser, bland annat bandbredd, tidluckor och effekt fördelas mellan användare. Fördelningen förändras dynamiskt över tid och behöver ta hänsyn till parametrar såsom användares mobilitet, frekvenstillgång, kanalegenskaper och typ av tjänster. Matematiskt optimala fördelningsalgoritmer går oftast inte att använda på grund av för hög beräkningskomplexitet och svårighet att modellera problemet. Traditionellt har därför heuristiska, suboptimala metoder använts för att lösa problemet. På senare tid har ML-baserade algoritmer för radioresursallokering i radionät i högre grad studerats. Nedan ges exempel på attacker utformade mot olika former av radioresursallokering baserat på ML-modeller. Tabell 3.3 visar en sammanställning av dessa exempel och hur de kan karakteriseras enligt egenskaperna i avsnitt 2.

I [30] utvärderas attacker mot effektfördelning mellan underbärvågor i multi-user MIMO[2]. Scenariot är att basstationen använder ett neuralt nätverk för att fördela effekten mellan underbärvågor baserat på uppmätta kanalskattningar. En angripare skapar en störning i syfte att minska datatakten för enstaka eller alla användare. En skuggmodell används för att skapa attacker, dock är det vagt kring skuggmodellen.

Samma problem som [30] studeras i [31] med syfte att jämföra olika attacker både för ett white-box- och ett black-box-scenario. Vidare undersöks hur *adversarial training* kan användas för att göra effektfördelningen mer robust mot attacker. En sidoeffekt av detta är ökad robusthet även mot slumpmässiga störningar. Scenariot som studeras i pappret är relativt teoretiskt, ML-modellernas indata är mobilanvändarnas positioner och attackerna genererar störningar i form av positionsförskjutningar. Resultaten är teoretiskt sett intressanta men kravet på direkt tillgång till input för ML-modellen renderar attacken svår att realisera i praktiken.

Kanalanvändning studeras i [32, 33] i vilka offrets användning av kanaler predikteras och attacken utförs genom att välja kanal att sända störeffekt på. Det krävs ingen explicit kunskap om offrets modell, men däremot antas att relationen mellan in- och ut-data är känd helt utan fel och utan fördröjning. Scenariot och kanalmodellen som används är väldigt simpelt, i praktiken lär sig nätet övergånssannolikheten för en Markov-modell med två tillstånd. Det är därför svårt utgående från dessa papper att uttala sig kring realiserbarhet och påverkan hos dessa typer av attacker.

Författarna till [32, 33] utökar problemställningen i [34] till ett scenario i vilket DRL används för kombinerat kanalval och effektfördelning. Målet med attacken är att minimera den summerade datatakten (eng. sum rate) för alla användare. Attackens prestanda jämförs med en optimal lösning som alltid väljer rätt kanaler att störa och en dum lösning som slumpmässigt väljer kanaler och resultaten visar på prestanda nära den optimala lösningen. Metoder för att motverka attacken analyseras också. I förhållande till tidigare papper av samma författare används här en mer realistisk kanalmodell och attackerna förefaller för det undersökta scenariot realiserbara.

I [35] används en liknade RL-agent som offrets för att minska offrets summerade datatakt. Attackagenten antas kunna avlyssna offrens återkoppling om summerad datatakt för att kunna lära sig hur gynnsam attacken var.

End-to-end-baserade ML-system är ett område som syftar till att ersätta flera (eller alla) block i en kommunikationskedja med en ML-modell. I [36] studeras AML-attacker mot ett end-to-end-baserat ML-system, som tränas att utföra motsvarande modulation och felrättande kodning. Kanalen är förenklad och modelleras endast med vitt Gaussiskt brus och synkroniseringsfel. AML-attacker skapas mot en skuggmodell av det neurala

[2] multiple-input and multiple-output

nätverket med annan arkitektur än offrets neurala nät. Slutsatsen är att attacker kan skapas mer effektivt under dessa förutsättningar än störning med vitt Gaussiskt brus. Kanalmodellen i [36] är dock kraftigt förenklad vilket innebär att en verklig fysisk attack är svårare att utföra.

Tabell 3.3: Sammanfattning av attacker mot radioresursallokering

Attack	Type		Phase		Knowledge		Specificity		Input Scope		Model Scope		Ref.
	ev	ex	train	exec	black	gray	target	untarget	ind	univ	ind	univ	
Analytisk och FGM med skuggmodell	✓			✓	✓		✓		✓			✓	[30]
FGSM, PGDM	✓	✓	✓	✓	✓			✓		✓	✓		[31]
RL (duell)	✓	✓	✓	✓	✓			✓		✓	✓		[32, 33]
RL (duell)	✓	✓	✓	✓	✓			✓		✓	✓		[34]
RL (duell)	✓	✓		✓	✓			✓		✓	✓		[35]
Varianter av FGM	✓			✓	✓			✓	✓	✓		✓	[36]

4 Diskussion

En aspekt av attacker mot ML inom trådlös kommunikation som skiljer området från sådana attacker inom andra tillämpningsområden, exempelvis bild- och textanalys, är den trådlösa radiokanalen. Attacker mot trådlös kommunikation behöver i de flesta realistiska fall vara fysiska attacker och påverkas därmed av kanalens egenskaper såsom dämpning, flervägsutbredning och fördröjning. Kanalens effekter måste tas i beaktning vid design av attacker för att attackerna ska ge önskad effekt i ett praktiskt scenario. För vissa typer av attacker är det flera kanalvägar, se figur 2.4, som påverkar hur attacken kan genomföras och vilken påverkan den får. Att kanalen, eller kanalerna, aldrig är helt känd i praktiken försvårar genomförandet och effektiviteten av en attack. En stor del av det som finns publicerat kring attacker mot ML inom trådlös kommunikation bortser helt ifrån eller gör grova förenklingar av radiokanalen, vilket gör det svårt att avgöra hur stort hot en viss attack utgör i verkligheten. Vidare är kommunikationsscenarierna som studeras för ML-attacker ofta förenklade: ett fåtal noder, simpla nätverkslösningar och brist på dynamik i kommunikationsscenarierna. Även dessa förenklade utgångsscenarier begränsar möjligheten att utvärdera attackers realiserbarhet och effekt i mer verklighetsnära scenarier. Detta kan dels bero på att området är relativt nytt och dels på att de som studerar området huvudsakligen fokuserar på tillämpningar av ML-modeller snarare än radiokommunikation.

Många av ML-attackerna baseras på att en så kallad skuggmodell skapas av offrets ML-modell i radiosystemet som ska attackeras. Målet med skuggmodellen är att fungera som en kopia av den ML-modell som angrips för att konstruera en effektiv attack. Med en bra skuggmodell har motståndaren tillräcklig kännedom om hur offrets ML-modell fungerar och kan på så vis påverka dess beslut och därmed försämra kommunikationens prestanda. Utan en skuggmodell behöver angriparen först ta reda på vilket besluts som fattats av offret för att sedan utföra attacken. Med en skuggmodell vet (med ett visst fel) motståndaren vilket beslut offret kommer att fatta. Då kan angriparen agera snabbare och påverka offret på ett annat sätt som tidigare inte var möjligt. En form av motattack är också att påverka motståndarens skapande av en skuggmodell. Detta kan ske genom att medvetet fatta felaktiga beslut och då baseras skuggmodellen på felaktig information. Detta är att betrakta som en form av poisoning-attack, som motmedel, över en radiokanal.

Universella attacker kan vara effektiva mot en ML-modell jämfört med klassisk brusstörning trots att modellens arkitektur och träningsdata är okända. Eftersom angriparen ofta inte behöver ha perfekt kunskap om kanalen fungerar universella attacker väl även i fall där kanalen mellan angripare och offer är helt eller delvis okänd [12, 11]. I fall där modellarkitekturen eller träningsdata är kända kan de universella attackerna göras än mer effektiva, genom att göra dem bättre anpassade till offret och därmed mer individuella. Det är därför viktigt att begränsa spridningen av information om de ML-modeller som används.

Träningsdata av tillräcklig mängd och kvalitet är avgörande för hur väl en ML-algoritm presterar. Viktiga principer för hur träningsdata bör skapas och kontrolleras behandlas i [6]. Risken för att egna träningsdatamängder infekteras av en motståndare har vi bedömt som en digital attack, vilket snarare är att betrakta som ett cyberhot. Detta ligger utanför studiens omfattning och har därför inte studerats vidare. Att nyttja existerande publika dataset kan innebära risk för förgiftning som är svårt att ha kontroll över. Tillverkare av radiokommunikationssystem som utnyttjar AI-teknik bör också beakta kvalitet och kontroll av träningsdata.

Sammantaget är bedömningen att det är svårare att utföra effektiva attacker mot ML i trådlösa kommunikationssystem, jämfört med andra teknikområden, framförallt på

grund av den trådlösa kanalen. Det är dock möjligt att utföra attacker trots stor osäkerhet om kanalen och offrets ML-modell. Det är svårt att bedöma allvarligheten i realistiska attacker, eftersom majoriteten av de publikationer som analyserats saknar realistiska kanalmodeller, utgår från förenklade kommunikationsscenarier eller saknar jämförelser med traditionell brusstörning. Det går inte att utesluta att realistiska AML-attacker kan utföras mot trådlösa kommunikationssystem med allvarliga konsekvenser. Därför bör robusthetshöjande åtgärder beaktas vid konstruktion och användning av ML-modeller i trådlösa kommunikationssystem.

5 Slutsatser

Utifrån de publikationer som analyserats i denna studie går det inte avfärda attacker mot ML-modeller som ett hot inom trådlös kommunikation. Det finns exempel inom flera av de undersökta områdena som vi bedömer är realiserbara. Majoriteten av studierna saknar dock realistiska kanalaspekter. Många av dessa utgår från förenklade kommunikationsscenarier och saknar ofta kvalitativa jämförelser med traditionella brusstörningar. Det är därmed svårt att bedöma om attacker mot ML inom trådlös kommunikation är ett allvarligt hot. Kunskapen om vilka egenskaper som krävs för att realisera AML-attacker inom trådlös kommunikation utgör dock en grund för framtida sårbarhetsanalyser.

Attacker mot maskininlärning för trådlösa kommunikationssystem är inte lika brett studerat som för andra tillämpningar som exempelvis bild- och språkbehandling. En väsentlig skillnad på attacker mot trådlös kommunikation, jämfört med andra teknikområden, är att potentiella attacker kommer ske fysiskt via radiogränssnittet. Detta gör att attackerna mot ML-modeller för trådlösa kommunikationssystem måste ta hänsyn till de effekter som radiokanalen innebär. Radiokanalen ökar komplexiteten för attacken vilket inte är fallet för andra tillämpningar och flera studier försummar radiokanalens inverkan.

Vissa typer av attacker förutsätter att ML-modellen är helt känd, s.k. white-box-attack. Denna typ av attack bedöms som osannolik. Mer sannolika attacker är så kallade black-box- eller gray-box-attacker där ingen eller viss information om ML-modellen finns eller kan observeras. Ett tillvägagångssätt som utnyttjas av flera attacker är att skapa en kopia, så kallad skuggmodell, av den modellen som ska angripas.

En observation från studerade arbeten är att få av dem jämför ML-attacker med klassisk störning, än färre i realistiska scenarier. Det är därför svårt att bedöma i vilken utsträckning AML-attacker är mer effektiva än traditionell brusstörning. Dessutom är komplexiteten hos ML-attacker ofta större i jämförelse med klassisk störsändning med brus. I detta arbete har inte lämpligheten att använda ML bedömts för de olika tillämpningarna. AML-attackerna baseras på att ML faktiskt används. I många tillämpningar inom trådlös kommunikation är det sannolikt att traditionella tekniker är mer lämpade och då är AML-attacker inte ett hot.

ML-attacker kan även utnyttjas av kommunikationssystem för att påverka motståndare. Det kan exempelvis ske genom medveten falsksignalering för att försvåra signalspaning.

En viktig aspekt som inte får glömmas bort är kontroll över träningsdata för ML-modellerna. Detta har inte berörts nämnvärt i denna rapport. Viktiga principer för hantering av träningsdata har behandlats i en tidigare rapport [6]. Vår bedömning är att förgiftning av träningsdata är en form av digital attack, vilken kan undvikas genom god hantering av träningsdata och förtroende för leverantörer.

Referenser

[1] E. Axell, P. Eliardsson, K. Hägglund, P. Brännström och C. Svensson, "Overview of machine learning in communication systems," Totalförsvarets forskningsinstitut, FOI-R--5275--SE, 2021.

[2] X. Yuan, P. He, Q. Zhu och X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, nr. 9, s. 2805–2824, 2019.

[3] F. Kamrani, L. Kanestad, L. Luotsinen, B. Pelzer, J. Sabel, V. Sandström och A. Tegen, "Attacking and Deceiving Military AI Systems," Totalförsvarets forskningsinstitut, FOI-R--5396--SE, mars 2023.

[4] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu och L. Qian, "Adversarial Machine Learning in Wireless Communications Using RF Data: A Review," *IEEE Commun. Surveys Tuts.*, vol. 25, nr. 1, s. 77–100, 2023.

[5] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain och H. Vincent Poor, "Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 25, nr. 4, s. 2245–2298, 2023.

[6] E. Axell, K. Hägglund, P. Eliardsson, A. Andersson, A. Komulainen och M. Karlsson, "Intelligent robust radiokommunikation," Totalförsvarets forskningsinstitut, FOI-R--5540--SE, 2023.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow och R. Fergus, "Intriguing properties of neural networks," 2014, arXiv:1312.6199.

[8] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi och P. Frossard, "Universal adversarial perturbations," 2017, arXiv:1610.08401.

[9] M. Sadeghi och E. G. Larsson, "Adversarial Attacks on Deep-Learning Based Radio Signal Classification," *IEEE Commun. Lett.*, vol. 8, nr. 1, s. 213–216, 2019.

[10] N. Papernot, P. McDaniel och I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," 2016, arXiv:1605.07277.

[11] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel och D. Towsley, "Robust adversarial attacks against DNN-based wireless communication systems," 2021, arXiv:2102.00918.

[12] J.-W. Chang, K. Sun, N. Heydaribeni, S. Hidano, X. Zhang och F. Koushanfar, "Magmaw: Modality-Agnostic Adversarial Attacks on Machine Learning-Based Wireless Communication Systems," okt. 2023, arXiv:2311.00207.

[13] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu och M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," 2019, arXiv:1703.06748.

[14] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine och S. Russell, "Adversarial policies: Attacking deep reinforcement learning," 2021, arXiv:1905.10615.

[15] S. N. Syed, P. I. Lazaridis, F. A. Khan, Q. Z. Ahmed, M. Hafeez, A. Ivanov, V. Poulkov och Z. D. Zaharis, "Deep Neural Networks for Spectrum Sensing: A Review," *IEEE Access*, vol. 11, s. 89 591–89 615, 2023.

[16] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu och S. Ulukus, "Channel effects on surrogate models of adversarial attacks against wireless signal classifiers," *Proc. IEEE Int. Conf. on Commun. (ICC)*, 2021, s. 1–6.

[17] Y. E. Sagduyu, Y. Shi och T. Erpek, "Adversarial deep learning for over-the-air spectrum poisoning attacks," *IEEE Trans. Mobile Comput.*, vol. 20, nr. 2, s. 306–319, 2021.

[18] ——, "IoT network security from the perspective of adversarial deep learning," *Proc. IEEE Int. Conf. on Sens., Commun., and Netw. (SECON)*, 2019, s. 1–9.

[19] Y. Shi, T. Erpek, Y. E. Sagduyu och J. H. Li, "Spectrum data poisoning with adversarial deep learning," *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2018, s. 407–412.

[20] Y. Shi, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu och J. H. Li, "Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies," *Proc. IEEE Int. Conf. on Commun. Work. (ICC Workshops)*, 2018, s. 1–6.

[21] F. Wang, C. Zhong, M. C. Gursoy och S. Velipasalar, "Resilient dynamic channel access via robust deep reinforcement learning," *IEEE Access*, vol. 9, s. 163 188–163 203, 2021.

[22] M. Liu, H. Zhang, Z. Liu och N. Zhao, "Attacking Spectrum Sensing With Adversarial Deep Learning in Cognitive Radio-Enabled Internet of Things," *IEEE Trans. Rel.*, vol. 72, nr. 2, s. 431–444, 2023.

[23] W. Xiao, Z. Luo och Q. Hu, "A review of research on signal modulation recognition based on deep learning," *Electronics*, vol. 11, nr. 17, 2022. [Online]. URL: https://www.mdpi.com/2079-9292/11/17/2764

[24] F. Restuccia, S. D'Oro, A. Al-Shawabka, B. Costa Rendon, K. Chowdhury, S. Ioannidis och T. Melodia, "Generalized wireless adversarial deep learning," *Computer Networks*, vol. 216, s. 109264, 2022.

[25] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek och S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," *Proc. Annual Conf. on Information Sciences and Systems (CISS)*, 2020, s. 1–6.

[26] W. Zhang, M. Krunz och G. Ditzler, "Stealthy adversarial attacks on machine learning-based classifiers of wireless signals," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, s. 261–279, 2024.

[27] B. Flowers, R. M. Buehrer och W. C. Headley, "Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications," *IEEE Trans. Inf. Forensics Security*, vol. 15, s. 1102–1113, 2020.

[28] M. DelVecchio, B. Flowers och W. C. Headley, "Effects of Forward Error Correction on Communications Aware Evasion Attacks," *Proc. IEEE Int. Symp. on Personal, Indoor, Mobile Radio Commun. (PIMRC)*, aug. 2020, s. 1–7.

[29] M. DelVecchio, V. Arndorfer och W. C. Headley, “Investigating a Spectral Deception Loss Metric for Training Machine Learning-based Evasion Attacks,” *Proc. 2nd ACM Workshop on Wireless Security and Machine Learning*, juli 2020, s. 43–48.

[30] B. Kim, Y. Shi, Y. E. Sagduyu, T. Erpek och S. Ulukus, “Adversarial attacks against deep learning based power control in wireless communications,” *IEEE Globecom Workshops (GC Wkshps)*, 2021, s. 1–6.

[31] B. R. Manoj, M. Sadeghi och E. G. Larsson, “Downlink power allocation in massive MIMO via deep learning: Adversarial attacks and training,” *IEEE Trans. on Cogn. Commun. Netw.*, vol. 8, nr. 2, s. 707–719, 2022.

[32] C. Zhong, F. Wang, M. C. Gursoy och S. Velipasalar, “Adversarial jamming attacks on deep reinforcement learning based dynamic multichannel access,” *Proc. IEEE Wireless Commun. Network. Conf. (WCNC)*, 2020, s. 1–6.

[33] F. Wang, C. Zhong, M. C. Gursoy och S. Velipasalar, “Defense strategies against adversarial jamming attacks via deep reinforcement learning,” *Proc. Annual Conf. on Information Sciences and Systems (CISS)*, 2020, s. 1–6.

[34] F. Wang, M. C. Gursoy och S. Velipasalar, “Adversarial reinforcement learning in dynamic channel access and power control,” *Proc. IEEE Wireless Commun. Network. Conf. (WCNC)*, 2021, s. 1–6.

[35] ——, “Adversarial reinforcement learning in dynamic channel access and power control,” *Proc. IEEE Wireless Commun. Network. Conf. (WCNC)*, 2021, s. 1–6.

[36] M. Sadeghi och E. G. Larsson, “Physical adversarial attacks against end-to-end autoencoder communication systems,” *IEEE Commun. Lett.*, vol. 23, nr. 5, s. 847–850, 2019.

FOI
Totalförsvarets forskningsinstitut
164 90 Stockholm

Tel: 08-55 50 30 00
Fax: 08-55 50 31 00

www.foi.se