Topi Luukkanen

KONEOPPIMINEN KYBERHYÖKKÄYKSISSÄ

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2018

TIIVISTELMÄ

Luukkanen, Topi
Koneoppiminen kyberhyökkäyksissä
Jyväskylä: Jyväskylän yliopisto, 2018, 26 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja(t): Seppänen, Ville

Tässä kirjallisuuskatsauksessa tarkastellaan koneoppimisen hyödyntämistä kyberturvallisuuden kontekstissa kyberhyökkäysten näkökulmasta. Koneoppimisella on lukuisia ja paljon tutkittuja sovelluksia kyberturvallisuuden edistäjänä, joten tutkielma keskittyy vähemmän tiedostettuun varjopuoleen kyberuhkien mahdollistajana, eli kuinka koneoppimista voidaan käyttää kyberhyökkäyksen apuvälineenä ja millaisia hyödynnettäviä haavoittuvuuksia koneoppimiseen tukeutuvissa järjestelmissä on. Yksi tärkeimmistä löydöistä tutkimuksessa oli taksonominen luokittelu koneoppimiseen kohdennettuihin kyberhyökkäyksiin. Luokittelu erottelee kolme vaikuttavaa tekijää hyökkäykseen: hyökkääjän mahdollisuuden vaikuttaa mallin harjoitusdataan, tietoturvaloukkauksen tyyppi ja hyökkäyksen tarkkuus. Edistyneitä koneoppimismenetelmiä voidaan käyttää kohteena olevan koneoppimismallin varastamiseen sekä haavoittuvuuksien löytämiseen jo hyvin pienilläkin määrillä harjoitusdataa.

Asiasanat: koneoppiminen, kyberuhka, kyberhyökkäys, vihamielinen koneoppiminen

ABSTRACT

Luukkanen, Topi
Machine Learning in Cyberattacks
Jyväskylä: University of Jyväskylä, 2018, 26 pp.
Information Systems, Bachelor's Thesis
Supervisor(s): Seppänen, Ville

This bachelor's thesis inspects the usage of machine learning in cybersecurity domain from an attack perspective. Machine learning has established its position as a vital part of every cybersecurity system and its applications in cyber defense are well-researched. Therefore, this literature review focuses on machine learning's less acknowledged side as an enabler of certain cyberthreats. Ergo, how machine learning can be utilized as a tool for cyberattacks and what types of vulnerabilities machine learning systems are prone to. One of the main findings of the literature review was a taxonomy for attacks against machine learning systems. The three axes defined for classifying cyberattacks include the attacker's ability to affect the model's training data, quality of information security violation and intended scope of the attack. The attacker can benefit from using advanced machine learning methods to discover vulnerabilities and infer vital aspects of the model with minimal amounts of model's training data.

Keywords: machine learning, cyberthreat, cyberattack, adversarial machine learning

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Tänä päivänä tieto on ubiikkia – ihmisten on mahdollista päästä käsiksi haluamaansa informaatioon missä vain ja milloin vain. Tietokoneet ja tietojärjestelmät eivät ole enää teknologisesta ympäristöstään eristäytyneitä yksiköitä, vaan internetin yhdistämiä solmuja kybertoimintaympäristössä. Tämä on mahdollistanut geologiset rajat rikkovan reaaliaikaisen tiedottamisen ja kansainvälisten yhteisöjen muodostumisen, mutta myös uudenlaisten ja ennalta tuntemattomien uhkakuvien realisoitumisen. Hyökkäykset tietoverkkoja, käyttäjien laitteita ja jopa kansainvälisiä tietojärjestelmiä kohtaan ovat nykyään arkipäivää. (Alpaydin, 2016) Kyberhyökkäyksillä voidaan lamauttaa kriittisen infrastruktuurin ohjausjärjestelmiä ja aiheuttaa laajoja toimintahäiriöitä niin verkossa kuin fyysisessä maailmassa, sekä vaikuttaa kansainväliseen politiikkaan ja liikemaailmaan (Puolustusministeriö, 2013). Tämän vuoksi tiedon, ICT-infrastruktuurin ja kybertoimintaympäristön suojaamisen merkitys on jatkuvassa kasvussa.

Symantecin (2018) raportin mukaan vuonna 2017 analysoiduista yhteyspyynnöistä kahdeksan prosenttia voitiin johtaa haittaohjelmiin ja yksittäisen bottiverkon tutkittiin lähettäneen yli kymmenen miljoonaa haitallista sähköpostia vain puolen vuoden aikana. Myös kyberhyökkäykset seuraavat globaaleja markkinatrendejä: vuoteen 2016 verrattuna hyökkäykset IoT-laitteisiin kuusinkertaistuivat ja webpalveluihin sulautetut kryptovaluuttalouhijat muodostivat lähes neljänneksen kaikista estetyistä selainpohjaisista kyberuhkista. Kyberrikollisuuden kustannuksiksi arvioidaan pahimmillaan kuusisataa miljardia dollaria vuodessa (Lewis, 2018).

Koneoppimisessa rakennetaan matemaattisia malleja, joiden avulla datasta voidaan tehdä ennustuksia tai tuottaa uutta merkityksellistä informaatiota. Se on mahdollistanut kompleksisten tilastollisten mallien rakentamisen ilman yksityiskohtaista ohjelmointia (Alpaydin, 2016). Laajasti käytössä oleviin koneoppimisen sovelluksiin kuuluvat muun muassa hakukoneet, hahmontunnistus, suosittelujärjestelmät, itseohjaavat kulkuneuvot ja bottiverkkojen havaitsemisjärjestelmät (Shi, Sagduyu & Grushin, 2017). Yksi koneoppimisen suurimmista eduista on sen kyky pystyä käsittelemään ja analysoimaan esimerkiksi sosiaalis-

ten medioiden ja verkkoliikenteen tuottamaa valtavaa datamäärää (Portugal, Alencar & Cowan 2017).

Koneoppimismenetelmiä hyödyntäviä kyberpuolustusjärjestelmiä on tuotettu ja tutkittu varsin paljon (mm. Livadas, Walsh, Lapsley & Strayer, 2006; Zanero & Serazzi, 2008; Buczak & Guven, 2015; Iglesias & Zseby, 2015), mutta huomattavasti vähemmän tunnetaan koneoppimista hyödyntäviä kyberhyökkäyksiä. Kattavan huomionsa vuoksi, tässä tutkielmassa kyberpuolustus ja koneoppimisen osuus siinä jätetään hyvin suppeaksi. Tutkielma vastaa seuraaviin tutkimuskysymyksiin:

- Millä tavoin koneoppiminen voi olla kyberhyökkäyksen kohteena?
- Voiko koneoppimismenetelmiä hyödyntää kyberhyökkäyksissä?

Tutkielma suoritettiin kirjallisuuskatsauksena. Hakukanavina lähteille toimivat IEEE Explorer, JYKDOK, Google Scholar, Scopus ja Elsevier. Lähteitä etsittiin pääasiallisesti hakusanoilla “machine learning“, “cyber security”, “cyber threat” sekä niiden eri yhdistelmillä. Hieman myöhemmin löytynyt hakutermi “adversarial machine learning” osoittautui tutkielman kannalta tärkeäksi, tuottaen suuren osan kyberturvallisuutta ja koneoppimista yhdistävistä tutkielman lähdeartikkeleista. Vakiintuneita suomenkielisiä käännöksiä ei useille termeille löytynyt, joten tässä tutkielmassa käytetään pitkälti Suomen kyberturvallisuusstrategian (Puolustusministeriö, 2013) sekä Limnéllin, Majewskin ja Salmisen (2014) käyttämiä kyberturvallisuuden termejä.

Tutkielman sisältö on rakennettu seuraavasti: johdannon jälkeinen toinen luku käsittelee koneoppimista, sen lähestymistapoja ja keskeisiä haasteita. Kolmas luku keskittyy kyberturvallisuuteen, sen erottamiseen tieto- ja ICT-turvallisuudesta sekä kyberhyökkäysten eri luokittelutapoihin. Neljäs luku tarkastelee koneoppimista kyberhyökkäysten kohteena sekä konkreettisena työkaluna hyökkäyksille. Viidennessä luvussa esitellään tutkielman johtopäätökset ja näkemykset jatkotutkimuksen suunnista.

2 KONEOPPIMINEN

Tässä luvussa käsitellään koneoppimisen taustoja, vakiintuneita lähestymistapoja sekä keskeisiä oppimisongelmia. Luku 2.1 esittelee tiivistetyn koneoppimisen määritelmän, luku 2.2 koneoppimisen yleiset menetelmät ja niiden käyttömahdollisuuksiin vaikuttavat lähtöasetelmat. Luku 2.3 käsittelee koneoppimiseen liittyviä yleisesti tunnistettuja haasteita.

2.1 Koneoppimisen määritelmä

Koneoppimisessa rakennetaan tilastoihin pohjautuvia matemaattisia malleja, joiden pohjalta voidaan tehdä päätelmiä ja ennustuksia (Alpaydin, 2014, 1). Yksi ensimmäisistä menestyksekkäistä ja siksi tunnetuimmista koneoppimisen taidonnäytteistä oli Arthur Samuelin 1950-luvulla kehittämä tammea pelaava ohjelma (Michalski, Carbonell & Mitchell, 2013). Samuelin (1959) mukaan koneoppimisen tavoitteena on ohjelmoida tietokone oppimaan itsenäisesti kokemuksen pohjalta, jolloin ohjelman sisältöä ei tarvitse määritellä pikkutarkasti. Jotta tuotetun mallin voidaan sanoa olevan oppiva, sen suorituskyvyn, kuten luokittelu- tai ennustustarkkuuden, kehityksen tulee olla kasvujohteista harjoittamisen aikana. Tätä voidaan arvioida erilaisten numeerisen suorituskykymittarien avulla (Alpaydin, 2014, 2).

2.2 Koneoppimismenetelmät

Koneoppimisen menetelmät ovat eriteltävissä neljään kategoriaan: ohjattuun (supervised), ohjaamattomaan (unsupervised) ja puoliohjattuun (semi-supervised) oppimiseen sekä vahvistusoppimiseen (reinforcement learning). Lähestymistavan valintaan vaikuttavat ratkaisevasti saatavilla olevan havaintoaineiston määrä sekä metadata käytettävästä aineistosta. Koko havaintoaineisto

tai osa siitä voi olla merkattu selitteillä (label), jotka osoittavat kullekin havainnolle oikeellisen luokan (Portugal, Alencar & Cowan, 2017.)

Mikäli havaintoaineiston jokaiselle ilmentymälle voidaan osoittaa selite, on mahdollista hyödyntää ohjattua oppimista. Kerätty havaintoaineisto erotellaan harjoitus- ja testidataksi: harjoitusdatalla koulutetaan malli ja testidatalla arvioidaan sen tarkkuutta. Tämän lähestymistavan tuottama malli, luokitin, yhdistää piirreavaruuden (feature space) arvot oikeisiin luokkiin (Er, Ave & Wang, 2016). Piirreavaruus on yksinkertaisimmillaan kaksiulotteinen alue (ks. kuvio 1), jossa piirteiden saamia arvoja kuvataan. Piirreavaruuden ulotteisuus eli kompleksisuus määräytyy piirteiden lukumäärän mukaan (Kotsiantis, 2007.) Erin ym. (2016) mukaan ohjatun oppimisen haasteena – ja monesti sen hyödyntämisen esteenä – on menetelmän vaatima suuri datamäärä luotettavan tuloksen saavuttamiseksi.

Ohjaamattomassa oppimisessa selitteillä merkattua aineistoa mallin harjoittamiseksi ei ole saatavilla, jolloin havainnoille ei voi osoittaa ennalta määrättyjä luokkia. Sen sijaan mallin tavoitteena on löytää datassa piileviä säännönmukaisuuksia ja havaintojen ryhmittymiä. Esimerkiksi, sosiaalisista medioista kerättyä massadataa analysoimalla ohjaamattoman oppimisalgoritmin avulla on mahdollista jaotella käyttäjät erilaisiin persoonallisuuskategorioihin ja käyttäjäprofiileihin, joiden avulla käyttäjille voidaan kohdistaa relevantteja mainoksia ja sisältöä. (Portugal ym. 2017)

Puoliohjatussa oppimisessa selitteellistä havaintoaineistoa on hyvin vähän, eikä datan määrä yksinään ole riittävä ohjattujen oppimismenetelmien hyödyntämiseksi, joten se yhdistetään selitteettömän havaintoaineston kanssa. Oppiminen ja havaintojen luokkien päättely perustuvat olettamukseen, että lähekkäin sijaitsevat havainnot kuuluvat suurella todennäköisyydellä samaan luokkaan. (Er ym., 2016)

Vahvistusoppimista on Barton ja Dietterichin (2004) mukaan mahdollista hyödyntää siinä tapauksessa, jos algoritmille syötettävien harjoitusesimerkkien selitteitä ei ole saatavilla, mutta on mahdollista arvioida, oliko muutos mallin suoriutumisessa toivotunlainen. Algoritmi seuraa ja pyrkii optimoimaan suorituskykyä kuvaavan skaalamuuttujan tuloksen, jonka maksimaaliset arvot tallentuvat algoritmin muistiin. Seuraavan kerran, kun algoritmi kohtaa aiemmin käsiteltyä havaintoa muistuttavan havainnon, maksimaalista arvoa etsitään ensi sijassa siltä piirreavaruuden alueelta, mistä aiempi optimaalinen arvo saavutettiin (Barto & Dietterich, 2004.)

2.3 Koneoppimisen keskeisiä haasteita

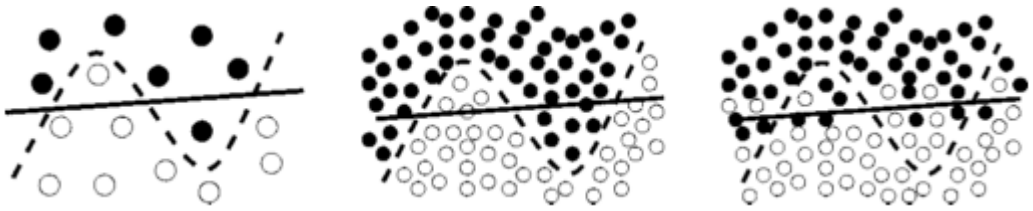
Koneoppimismallin harjoittamisessa tulee ottaa huomioon monta seikkaa. Kompleksiset mallit perinteisesti luokittelevat mallin harjoittamiseen käytettyä dataa suurella tarkkuudella, mutta tämä ei yksiselitteisesti tarkoita, että mallin suorituskyky olisi yhtä hyvä uutta dataa luokiteltaessa. Yksinkertaiset mallit suoriutuvat tyypillisesti päinvastaisesti: harjoitusdatassa virheitä sallitaan use-

ampia verrattuna kompleksiseen malliin, mutta uutta dataa käsiteltäessä tarkkuus on kompleksista suurempi (Barto & Dietterich, 2004.)

2.3.1 Yli- ja alisovittaminen

Myungia (2000) mukaillen, jos ennustustarkkuus mallin harjoitusdatassa on todella korkea, mutta suorituskyky laskee huomattavasti testausvaiheessa, mallin yleistettävyys on kehno. Tämä ongelma, eli ylisovittaminen (overfitting), tapahtuu, kun malliin sisällytetään turhan monta piirrettä, jolloin malli käytännössä oppii harjoitusdatan ulkoa. Ylisovittava malli ottaa myös herkästi huomioon harjoitusdatan kohinaa (noise), eli harhaanjohtavia ja oppimisen kannalta turhia havaintoja, kuten poikkeavia arvoja sekä väärin luokiteltua ja tulkittua dataa. On myös mahdollista, että mallista puuttuu jokin ennustus- tai luokittelutarkkuuden kannalta tärkeä piirre, jolloin havainto voidaan virheellisesti tulkita kohinaksi (Alpaydin, 2014, 31). Alisovittamisessa (underfitting) havaintoaineisto on rakenteeltaan mallia monimutkaisempi, minkä vuoksi malli ei ”taivu” selittämään havaintoaineiston monimutkaisempaa jakaumaa piirteiden liian vähäisen määrän vuoksi (Alpaydin, 2014, 39).

Kuviossa 1 vasemmanpuoleisimmasta kuvassa nähdään, kuinka liian pienestä havaintoaineistosta on vaikea sanoa kumpi malleista, lineaarinen (suora viiva) vai kompleksimpi epälineaarinen (katkoviiva), kuvaa parhaiten havaintojen todellista jakaumaa. Kun havaintoaineiston kokoa kasvatetaan riittävästi, nähdään, että keskimmäisen kuvan tilanteessa epälineaarinen malli on tosi ja lineaarinen malli alisovittuu, ja oikeanpuoleisimmassa kuvassa puolestaan lineaarinen malli on tosi ja epälineaarinen malli ylisovittuu.



Kuvio 1 Ali- ja ylisovittaminen kaksiulotteisessa piirreavaruudessa (Müller, Mika, Rätsch, Tsuda & Schölkopf, 2001, 182 kuviosta)

2.3.2 Ulotteisuuden kirous

Yksi koneoppimisen keskeisimmistä ongelmista on niin sanottu uloitteisuuden kirous (curse of dimensionality). Sen mukaan edustavaan otokseen tarvittavan datan määrä kasvaa eksponentiaalisesti suhteessa piirteiden lukumäärään (Bai, 2014). Suuremmalla määrällä piirteitä dataa on mahdollista kuvata tarkemmin ja yksityiskohtaisemmin, mutta samaan aikaan algoritmien laskennallinen

kompleksisuus kasvaa, jolloin havaintojen tekeminen datasta hidastuu ja vaikeutuu (Zanero & Serazzi, 2008).

2.3.3 Piirteen- ja mallin valinta

Blumin ja Langleyn (1997) mukaan attribuutilla tai muuttujalla tarkoitetaan koneoppimisen kontekstissa jotain dataa kuvaavaa ominaisuutta, jolla on arvo. Piirteet (feature) voivat olla yksittäisiä muuttujia tai niiden yhdistelmiä, ja niiden tehtävänä on erotella havainnot toisistaan. Vaativissa tehtävissä, kuten luonnollisen kielen prosessoinnissa algoritmit joutuvat käsittelemään valtavia määriä piirteitä: käsinkirjoitetun tekstin luokittelussa havaintoja kuvaavia piirteitä voi hyvinkin olla kymmenistä tuhansista jopa miljooniin asti. Kaikki piirteet eivät kuitenkaan erottele havaintoja toisistaan yhtä tehokkaasti, vaan lopulta vain pieni piirteiden osajoukko on tarpeen sisällyttää malliin (Blum & Langley, 1997; Bai, 2014.) Siksi piirteenvalinta (feature selection) on välttämätöntä kompleksisuuden vähentämiseksi ja toivotun suorituskyvyn saavuttamiseksi. Valituista piirteistä voidaan rakentaa myös useita malleja, joita voidaan tutkia ja vertailla keskenään esimerkiksi suorituskyvyn ja piirteiden määrän suhteen. Koneoppimismallien yksi tärkeä tavoite on olla yleistettäviä, jotta niiden tuloksiin voidaan luottaa myös tulevaisuudessa tuntematonta dataa käsiteltäessä. Mallin yleistettävyys voidaan varmistaa keskittymällä pieneen, mutta havaintoja hyvin erittelevään piirteiden osajoukkoon (Blum & Langley, 1997.) Myungin (2000) mukaan yksi matemaattisen mallinnuksen keskeisistä haasteista onkin löytää tasapaino mallin selittävyyden ja kompleksisuuden suhteen.

3 KYBERTURVALLISUUS

Tämä luku käsittelee kyberturvallisuuteen liittyviä keskeisiä käsitteitä, kyberhyökkäysten luokittelutapoja sekä kyberuhkien eri muotoja. Kyberhyökkäyksiltä puolustautuminen ja uhkatekijöiden kontrolloiminen ovat hyvin keskeinen osa kyberturvallisuutta, mutta tämän näkökulman tarkastelu rajataan tutkielman ulkopuolelle. Luku 3.1 erottaa kyberturvallisuuden siihen läheisesti liittyvistä käsitteistä. Luvussa 3.2 tutkitaan lähdekirjallisuudesta löytyviä erilaisia kyberhyökkäysten luokittelutapoja.

3.1 Kyber-, tieto- ja ICT-turvallisuus

Kyberturvallisuus, tietoturvallisuus, sekä viestiliikenne- ja tietoverkkoturvallisuus (myöh. ICT) ovat keskenään läheisiä ja ajoittain toistensa korvikkeena käytettyjä käsitteitä, joten niiden yhteys ja keskinäiset erot on syytä tuoda esiin. Suomen kyberturvallisuusstrategiassa (Puolustusministeriö, 2013) tietoturvallisuus määritellään niinä järjestelyinä, joiden avulla taataan tiedon luottamuksellisuus (confidentiality), eheys (integrity) ja saatavuus (availability). Tiedon – analogisen tai digitaalisen – tulee siis olla ulkopuolisten tavoittamattomissa, yhdenmukaista alkuperäisen tiedon kanssa ja saatavilla haluttuna ajankohtana.

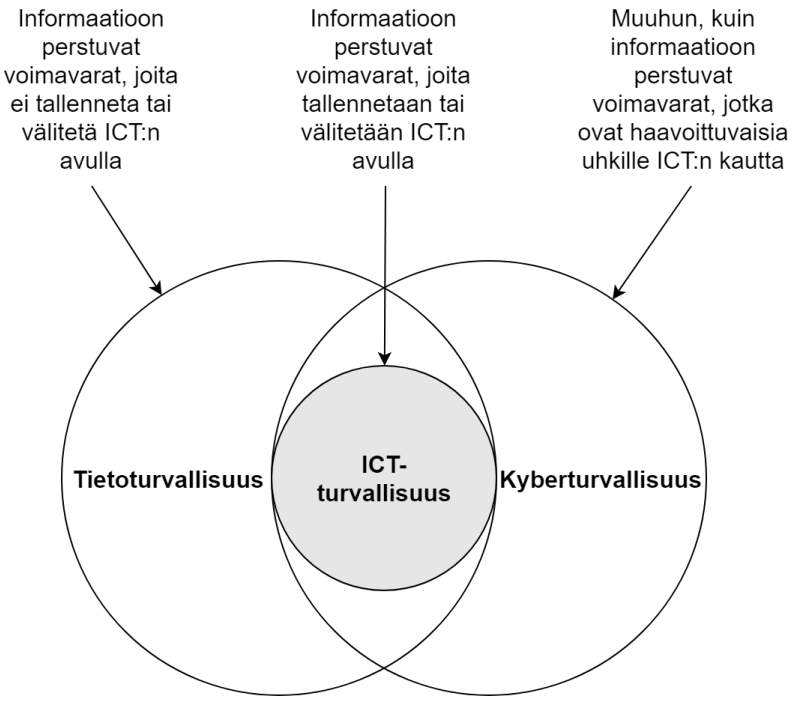
ICT-turvallisuudessa pyritään turvaamaan digitaalisen informaation lisäksi myös kaikki sen käsittelyssä osallisena olevat resurssit ja prosessit. Sähköiseen tietojärjestelmään tallennetun informaation ei kuitenkaan voida katsoa olevan yksistään turvassa, ellei koko järjestelmän teknologinen infrastruktuuri ole suojattu. ICT-turvallisuuden toteutuminen on siis edellytys digitaalisen tiedon tietoturvallisuudelle. (Von Solms & Van Niekerk, 2013)

Kyberturvallisuudella puolestaan tarkoitetaan kybertoimintaympäristön, siinä toimivien ihmisten ja laitteiden, sekä sen kautta tavoitettavien voimavarojen turvallisuutta. (Von Solms & van Niekerk, 2013; Puolustusministeriö, 2013). Kyberturvallisuuden pelikenttänä toimiva kybertoimintaympäristö käsittää

kaikki internetiin yhdistetyt laitteet ja niiden kautta tavoitettavan digitaalisen maailman palveluineen (Limnéll ym., 2014).

Tietoturvallisuus vaarantuu, kun yhtä tai useampaa sen perusperiaatteista – luottamuksellisuus, eheys ja saatavuus – loukataan. Keskiössä on informaatio, jonka väärinkäytön aiheuttama vahinko ihmisiin on kuitenkin aina välillistä. Esimerkiksi liikesalaisuuksien vuotaminen kilpailijalle voi aiheuttaa vahinkoa menetetyn markkinavaltin tai rahan kautta (Von Solms & Van Niekerk, 2013). Von Solms ja van Niekerk (2013) korostavat, että tietoturvallisuudessa ihminen on pikemminkin vain yksittäinen osa koko tietoturvaprosessia, kun taas kyberturvallisuudessa ihmiset voivat olla konkreettisesti hyökkäyksen kohteena ja omalla toiminnallaan jopa osallistua hyökkäykseen tietämättään.

Suomen kyberturvallisuusstrategiaan (Puolustusministeriö, 2013) kirjatun kyberuhkan määritelmän mukaan kybertoimintaympäristöön kohdistuvat uhat koskettavat myös tietoturvallisuutta. Tämä rajaus on kuitenkin suppea. Von Solms ja Van Niekerk (2013) argumentoivat, että kyberuhka ei välttämättä vaaranna tietoturvaa (ks. kuvio 2) Esimerkiksi vedenjakelun katkaiseminen hyökkääjän toimesta ja nettikiusaaminen ovat kyberturvallisuutta koskettavia uhkia, mutta eivät loukkaa tietoturvallisuuden perusperiaatteita. Kyberturvallisuudessa vahinko on siis mahdollista suoraan kohdistaa myös suoraan ihmisiin ilman, että tiedon luottamuksellisuus, eheys tai saatavuus kärsii. Kun Suomen kyberturvallisuusstrategiaa tarkastellaan Von Solmsin ja Van Niekerkin (2013) tutkimuksen valossa, voidaan tulkita, että kyberturvallisuus on rinnastettu ICT-turvallisuuden kanssa.



Kuvio 2 Tieto-, ICT- ja kyberturvallisuuden välinen suhde (suom. Von Solms & Van Niekerk, 2013, 101)

3.2 Kyberhyökkäysten luokittelutapoja

Jotta voidaan puhua kyberhyökkäyksestä, hyökkäyksen tavoitteena tulee olla Hathawayn ym. (2012) mukaan tietokoneverkon, kuten internetin tai yrityksen sisäisäiseen käyttöön rajatun verkon, toiminnan heikentäminen. Väyliä, joita pitkin järjestelmään päästään sisään ja toteuttamaan hyökkäys, kutsutaan hyökkäysvektoreiksi. Kyberhyökkäyksen yksi tunnuspiirteistä on ihmisiin pohjautuvien hyökkäysvektorien hyödyntäminen järjestelmässä olevien haavoittuvuuksien sijaan (Shabut, Lwin & Hossain, 2016). Suosituimpiin hyökkäysvektoreihin kuuluvat muun muassa sähköpostien tietojenkalasteluviestit, haittaohjelmat ja hajautetut palvelunestohyökkäykset (Fraley & Cannady, 2017). Symantecin (2018) raportin mukaan laajimmin käytetty hyökkäysvektori viime vuonna oli kohdistettuihin tietojenkalasteluviesteihin lukeutuva ”spear-phishing” -hyökkäys, joissa tavoitteena on saada uhri avaamaan saastutettu sähköpostin liitetiedosto tai klikkaamaan haitallista linkkiä. Toiseksi käytetyimpiä olivat ”myrkytetyt keitaat” (watering hole) eli hyökkäykset, joissa kohdeorganisaation henkilöstön käyttämille nettisivuille asetetaan haitallisia linkkejä, jotka aktivoituvat vasta, kun ennalta määrättyjen IP-osoitteiden havaitaan vierailevan sivustolla (Symantec, 2018).

Hathaway ym. (2012) tunnistavat yleisen tason luokittelutavaksi syntaktiset ja semanttiset kyberhyökkäykset. Syntaktisilla hyökkäyksillä pyritään aiheuttamaan toimintahäiriöitä verkossa saastuttamalla tietokoneen käyttöjärjestelmä erilaisilla haittaohjelmilla, esimerkiksi troijanhevosella, eli hyötyohjelman sisään piilotetulla viruksella. Syntaktisissa hyökkäyksissä toimintahäiriöt ovat päällepäin näkyviä ja ilmeisiä. Sen sijaan semanttisissa hyökkäyksissä järjestelmä vaikuttaa ulospäin toimivan normaalisti, mutta sen käsittelemän informaation eheyttä on rikottu. (Hathaway ym., 2012)

Uman ja Padmavathin (2013) tutkimuksen mukaan kyberhyökkäyksiä voidaan myös luokitella lakiperustaisesti, tarkoitusperän, osallistumisen tason, laajuuden ja käytetyn verkkotyypin mukaan. Tässä tutkielmassa keskitytään tarkastelemaan hyökkäysten tarkoitusperiä ja lakiperustaista luokittelua.

3.2.1 Hyökkäysten tarkoitusperäinen luokittelu

Uma ja Padmavathi (2013) tunnistavat kolme yleistä luokkaa kyberhyökkäysten tarkoitusperäiseen luokitteluun: Tiedustelu-, käyttöoikeus- ja palvelunestohyökkäykset. **T**iedusteluhyökkäysten (reconnaissance attack) tavoitteena on kartoittaa luvattomasti järjestelmän laajuutta ja sen tarjoamia palveluja etsien potentiaalisia murtautumiskohtia, joita voidaan hyödyntää tulevaisuudessa. Esimerkiksi verkon analysoijat eli nuuskijat (sniffer) salakuuntelevat ja tallentavat verkon yli lähetettäviä datapaketteja myöhempää analysointia varten ja portin skannauksessa hyökkääjä lähettää sarjan yhteyspyyntöjä yrittäen oppia, mitä portteja kukin palvelu hyödyntää.

Käyttöoikeushyökkäyksessä (access attack) hyökkääjä pyrkii sisään käyttäjätunnuksilla ja salasanoilla suojattuun järjestelmään, jota hän ei ole valtuutettu käyttämään. Järjestelmään voidaan murtautua muun muassa autentikointi- ja webpalveluiden haavoittuvuuksia hyödyntämällä sekä päästä käsiksi käyttäjätileille, luottamuksellisiin tietokantoihin ja muuhun arkaluonteiseen informaatioon. Esimerkiksi käyttöoikeushyökkäyksiin luettavassa mies välissä -hyökkäyksessä (man-in-the-middle) hyökkääjä asettuu kahden kommunikoivan osapuolen väliin, sieppaa ja halutessaan muokkaa lähetettyjä viestejä tavoitteensa mukaisesti. Osapuolet luulevat olevansa toistensa kanssa suoraan yhteydessä, ja toisiinsa luottaessaan, lähettävät hyökkääjälle arkaluonteista tietoa, kuten palvelun käyttäjätunnuksia tai henkilötietoja. (Uma & Padmavathi, 2013)

Palvelunestohyökkäysten (denial of service) tavoitteena on häiritä palvelun normaalia käyttöä hidastamalla sen toimintaa merkittävästi tai kaatamalla sitä ylläpitävä tietojärjestelmä tai -verkko. Järjestelmään tallennettua informaatiota voidaan myös vioittaa tai poistaa ja sitä kautta estää palvelun toimiminen. Eräs tehokas ja laajasti käytetty toteutustapa on hajautettu palvelunestohyökkäys (distributed denial of service), jossa hyökkääjä kokoaa haittaohjelman avulla lukuisista tietokoneista etäohjattavan bottiverkon, jota ohjaamalla kohde voidaan hukuttaa yhteyspyyntöjen tulvaan. (Uma & Padmavathi, 2013)

3.2.2 Lakiperustainen luokittelu

Eri maiden lakiteksteihin ja valtionhallintojen dokumentteihin kirjattuja kyberuhkien muotoja ovat Uman ja Padmavathin (2013) mukaan kyberrikollisuus, -vakoilu, -terrorismi ja -sodankäynti. Osa lähdekirjallisuudesta löytyneistä luokitteluista, kuten Limnéll ym. (2014) sekä Puolustusministeriö (2013), lisää edellä mainittujen joukkoon myös kyberaktivismin, eli ”haktivismin”. Vaikka uhat tunnistetaan globaalisti, yksiselitteistä kansainvälistä määritelmää ei monelle termille ole. Sisäministeriön (2017) tietoverkkorikollisuutta käsittelevässä julkaisussa kyberrikollisuus käsitetään kybertoimintaympäristössä tapahtuvana tai sitä hyödyntävinä rikollisena toimintana. Kyberrikosten kohteena ovat yksilöt ja yritykset esimerkiksi aineellisten oikeuksien tai yksityisyyden loukkausten muodossa (Limnéll ym., 2014).

Kybervakoilu on toimintaa, jossa yritykset tai valtiolliset toimijat pyrkivät murtautumaan esimerkiksi kilpailevan yrityksen tai armeijan tietojärjestelmään ja keräämään sieltä luottamuksellista ja salaista tietoa. Määritelmä ottaa huomioon myös kriisitilanteeseen valmistautuvan ja ennakoivan tiedustelutoiminnan kriittisen infrastruktuurin haavoittamiseksi tai lamauttamiseksi (Sisäministeriö, 2017; Limnéll ym., 2014.)

Kyberterrorismi luokitellaan hyökkäyksinä, joiden tavoitteena on laaja-alaisten häiriöiden ja sekasorron aiheuttaminen kybertoimintaympäristön välityksellä. Kyberterrorismin kohteena ovat muun muassa kriittiset infrastruktuurit ja valtioiden tietojärjestelmät (Limnéll ym., 2014) Kybersodankäynnissä osapuolet ovat kumpikin valtiollisia toimijoita, joiden suorittamat toimet joko ta-

pahtuvat sodankäynnin yhteydessä tai ovat verrattavissa aseelliseen hyökkäykseen. (Uma & Padmavathi, 2013)

Limnéll ym. (2014) määrittelevät haktivismin kybertoimintaympäristössä tapahtuvana aktivismina, joka toteuttaa henkilön tai ryhmän ideologista, poliittista tai sosiaalista motiivia. Haktivistit toimivat tyypillisesti lain harmaalla alueella ja pyrkivät saavuttamaan laajaa mediahuomiota muun muassa palvelunestohyökkäyksillä, virtuaalisella sabotaasilla, kaappaamalla ja sotkemalla nettisivuja, sekä varastamalla ja levittämällä luottamuksellista tietoa. Teot ovat luonteeltaan väkivallattomia, mutta osittain rangaistavia. Haktivismille tyypillisenä vaikutuskeinona on kohteensa parodioiminen ja asettaminen naurunalaiseksi.

4 KONEOPPIMINEN KYBERTURVALLISUUDEN KONTEKSTISSA

Tässä luvussa tarkastellaan koneoppimisen roolia kyberturvallisuudessa. Pääpaino tarkastelussa on koneoppimiseen kohdistetuilla hyökkäyksillä ja koneoppimismenetelmien hyödyntämisessä kyberhyökkäyksissä Luku 4.1 antaa hyvin tiivistetyn yleiskuvan koneoppimisella tuetuista kyberpuolustusjärjestelmistä. Luku 4.2 esittelee koneoppimiseen kohdistuvien hyökkäysten taksonomian ja luokitteluun perustuvia konkreettisia esimerkkejä. Luvussa 4.3 tarkastellaan, kuinka koneoppimista voidaan hyödyntää osana hyökkäystä.

4.1 Koneoppimisalgoritmit kyberhyökkäysten havaitsemisessa

Tunkeilijan havaitsemisjärjestelmät (Intrusion Detection System) eli IDS-järjestelmät tarkkailevat järjestelmään tulevaa ja sieltä lähtevää dataa pyrkien löytämään potentiaalisia hyökkäyksiä. Ne ovat kiinteä osa kybertuvallisuusjärjestelmiä ja auttavat löytämään merkkejä tiedon kopioimisesta, muokkaamisesta ja tuhoamisesta sekä luvattomasta käytöstä. IDS-järjestelmillä on pääasiallisesti kolme eri muotoa: väärinkäyttöön perustuvat, poikkeavuuksien eli anomalioiden havaitsemiseen perustuvat ja hybridimenetelmät. (Barreno, Nelson, Sears, Joseph & Tygar, 2006; Buczak & Guven, 2015)

Väärinkäyttöön perustuva arviointi pohjautuu tunnettujen hyökkäysten tunnusmerkeistä koostettuun listaukseen, johon käsiteltävää dataa verrataan. Ne ovat tehokkaita havaitsemaan aiemmin kohdattuja hyökkäyksiä tuottamatta valtavasti vääriä hälytyksiä (Buczak & Guven, 2015). Sommer ja Paxson (2010) mainitsevatkin, että väärinkäyttöön perustuvien järjestelmien perustavanlaatuisena ongelmana kuitenkin on niiden kyvyttömyys havaita uusia hyökkäyksiä – koneoppimisen avulla pystytään parhaiten tunnistamaan samankaltaisuuksia suhteessa aiemmin käsiteltyyn dataan, eikä luokittelemaan oikein täysin ennalta tuntemattomia havaintoja. Buczakin ja Guvenin (2015) mukaan anomalioiden havaitsemiseen tukeutuvilla järjestelmillä ei ole edellä mainittua ongelmaa. Ne

tuntevat ympäristökohtaisesti normaalin aktiviteetin ja käsittelevät kaikki siitä poikkeavat tapahtumat hyökkäyksinä. Menetelmän etuna on mahdollisuus havaita uusia hyökkäyksiä, mutta haittapuolena suuri määrä vääriä hälytyksiä, kun järjestelmä luokittelee aiemmin tuntematonta harmitonta käyttäytymistä anomaliaksi. Hybridimenetelmät yhdistävät kahden edellä mainitun lähestymistavan ominaisuuksia: tunnettujen hyökkäysten havaitsemista pyritään tehostamaan samalla, kun vähennetään ennalta tuntemattomien hyökkäysten tulkitsemista normaaliksi aktiviteetiksi (Buczak & Guven, 2015.)

Tunkeilijan havaitsemisjärjestelmien lisäksi koneoppimista on hyödynnetty kyberpuolustuksessa muun muassa bottiverkkojen havaitsemisessa (Livadas, Walsh, Lapsley & Strayer, 2006), tunkeutumisenestojärjestelmissä (Buczak & Guven, 2015), roskapostisuodattimissa (Barreno ym., 2006) sekä keskitetysti loki- ja tapahtumatietoja tallentavissa SIEM-järjestelmissä (Security Information and Event Management) (Feng, Wu & Liu, 2017).

4.2 Koneoppiminen hyökkäyksen kohteena

Barreno ym. (2006) ovat luoneet taksonomisen luokittelun koneoppimisalgoritmeihin ja niiden päälle rakennettuihin tietojärjestelmiin kohdistettuja kyberhyökkäyksiä varten (ks. taulukko 1). Hyökkäyksiä voidaan tarkastella kolmesta näkökulmasta: vaikutustapa, tietoturvallisuusloukkauksen tyyppi ja tavoiteltu tarkkuus. Vaikutustavaltaan hyökkäykset jaotellaan kausatiivisiin (causative) ja tutkiviin (exploratory) hyökkäyksiin, tietoturvallisuusloukkauksen pohjalta eheys- ja saatavuushyökkäyksiin sekä tarkkuutensa mukaan kohdistettuihin (targeted) ja ei-kohdistettuihin (indiscriminate) hyökkäyksiin.

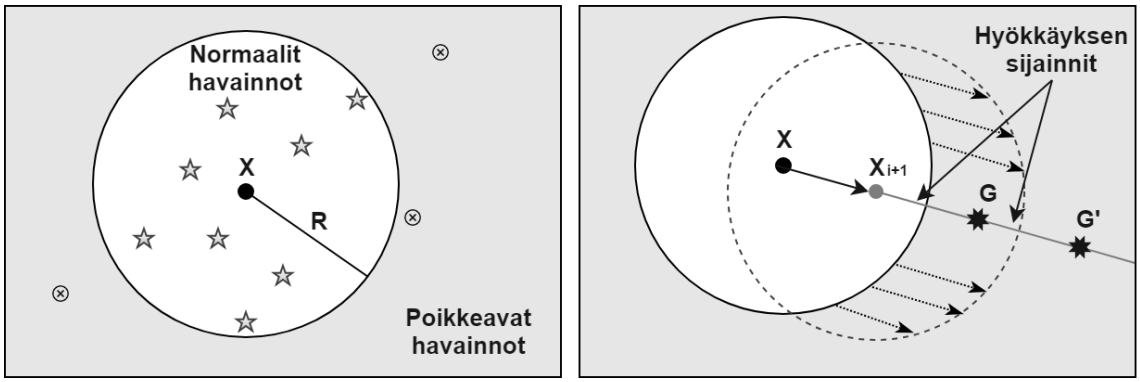
Kausatiivisissa hyökkäyksissä hyökkääjällä on mahdollisuus vaikuttaa rakennettavan luokittimen harjoitusdataan. Tavoitteena on johtaa oppimista harhaan niin, ettei malli opi erottelemaan havaintoja tarkoituksenmukaisesti. Tämä voi tapahtua konkreettisesti esimerkiksi vääristämällä harjoitusdatan selitteitä tai tilastollista jakaumaa, jolloin järjestelmä väärinluokittelee hyökkääjän syötteen sallien tunkeutumisen järjestelmään. Tutkivat hyökkäykset eivät yritä vaikuttaa mallin opetukseen, vaan hyödyntää sen olemassa olevia puutteellisuuksia. Ne ajoittuvat siis vaiheeseen, kun malli on jo koulutettu ja niiden tehtävänä on kerätä tietoa algoritmin senhetkisestä tilasta sekä toiminnasta. Tutkivat hyökkäykset pyrkivät paljastamaan hyökkäysvektoreita väärin luokiteltujen havaintojen avulla. (Barreno ym., 2006)

		Eheys	***Saatavuus***
Kausatiivinen	*Kohdistettu*	Salli tarkkaan määritetty tunkeutumisen mahdollistava syöte	Luo riittävästi virheitä, jotta järjestelmästä tulee käyttökelvoton yhdelle ihmiselle tai palvelulle
	Ei-kohdistettu	Salli ainakin yksi tunkeutumisen mahdollistava syöte	Luo riittävästi virheitä, jotta koneoppimismallista tulee käyttökelvoton
Tutkiva	*Kohdistettu*	Löydä pienestä potentiaalisten syötteiden joukosta järjestelmän sallima tunkeutumisen mahdollistama syöte	Löydä joukko havaintoja, jotka malli luokittelee väärin.
	Ei-kohdistettu	Löydä mikä tahansa järjestelmän sallima tunkeutumisen mahdollistava syöte	

Taulukko 1 Luokittelu koneoppimiseen kohdistetuille hyökkäyksille (suom. Barreno ym., 2006, 3)

Tiedon eheyttä loukatessa hyökkääjä onnistuu syöttämään haitallista dataa, jonka luokitin merkitsee virheellisesti normaaliksi (false negative). Saatavuushyökkäyksissä myös harmiton data pyritään saamaan luokiteltua haitalliseksi (false positive), jolloin luokittimen toiminnasta tulee niin epävarmaa, ettei palvelun tai järjestelmän normaali käyttö ei ole mahdollista. (Barreno ym., 2006, Nelson ym. 2008). Kun hyökkääjän aikeena on heikentää luokittelua tarkasti määriteltyjen havaintojen osalta, puhutaan kohdistetusta hyökkäyksestä. Ei-kohdistettujen hyökkäysten tavoite on löyhempi ja tärkeämpää on luokittimen suorituskyvyn yleinen heikentäminen (Barreno, Nelson, Joseph & Tygar, 2010).

Jatkuvasti oppivat online-luokittimet soveltuvat hyvin dynaamiseen ympäristöön, kuten sähköpostien luokitteluun, jossa syötteet muuttuvat ajan kuluessa (Barreno ym., 2006). Luokitin uudelleenharjoitetaan säännöllisesti päivitetyllä harjoitusdatalla. Mikäli dataa ei esiprosessoida ennen uutta harjoitussykliä, kaikki havainnot päätyvät harjoitusdataan. Tämän vuoksi online-luokittimet ovat alttiimpia kausatiivisille hyökkäyksille. (Nelson ym. 2008) Kuvio 3 esittää kohdistettua kausatiivista eheyshyökkäystä, jota voitaisiin käyttää muun muassa edellä mainittuun roskapostisuodattimeen. Havaintojakauman keskiarvosta X ulottuva kiinteä säde R määrittää alueen piirreavaruudessa, jossa sen sisällä olevat havainnot (☆) luokitellaan normaaleiksi ja ulkopuolelle jäävät havainnot (⊗) poikkeaviksi. Luokittimen suorituskyvyn tilasta tietoinen hyökkääjä voi saavuttaa tavoitepisteidensä G ja G' väärinluokittelun lähettämällä riittävän monta sähköpostia, joiden saamat arvot piirreavaruudessa sijoittuvat päätösrajan sisäpuolelle tavoitepisteiden suunnassa.



Kuvio 3 Esimerkki hyökkäyksestä online-luokittimeen (suom. Barreno ym., 2006, 7 kuviosta). Vasemmalla lähtötilanne ja oikealla hyökkäyksen aiheuttama päätösrajan siirtyminen.

Nelsonin ym. (2008) tutkimuksessa roskapostisuodattimeen kohdistetiin kausatiivisen saatavuushyökkäys, jonka tarkoituksena oli päinvastaisesti normaalien sähköpostien väärinluokittelu roskapostiksi. Siinä hyökkäyssähköposteihin sisällytettiin monia, harmittomissa sähköposteissa yleisesti käytettyjä sanoja, jotta niiden esiintyminen yhdistettiin herkemmin roskaposteihin. Ei-kohdistetussa hyökkäyksessä tavoitteena oli saada mikä tahansa normaali sähköposti luokiteltua roskapostiksi suodattimen toimesta ja saada uhri poistamaan suodatin käytöstä. Kohdistettu hyökkäys taas rajasi kapeamman ryhmän sähköposteja, joita uhrin ei haluttu näkevän. Esimerkiksi yritys voisi hyödyntää edellä mainittua hyökkäystä, jotta asiakas ei vastaanota kilpailevan yrityksen tarjousta.

Ei-kohdistetut hyökkäykset onnistuivat yhdellä prosentilla roskapostisuodattimen harjoitussyötteistä aiheuttamaan väärinluokittelun 36%:lle uusista normaaleista sähköposteista. Kohdistetussa hyökkäyksessä kohdeviestin väärinluokittelussa onnistuttiin 60% kerroista, kun viestin tunnusmerkeistä 30% oli tiedossa (Nelson ym., 2008.)

4.3 Koneoppiminen kyberhyökkäyksen apuvälineenä

Niin sanotussa vihamielisessä (adversarial) koneoppimisessa hyökkäävä osapuoli tekee hienovaraisia muokkauksia aitoihin syötteisiin ja rakentaa siten hyökkäysesimerkkejä (adversarial example). Hyökkäysesimerkit vaikuttavat ihmisen tarkastelemana normaaleilta, mutta tehty muutos on riittävän suuri, että malli luokittelee syötteen väärin. Hyökkäysesimerkkien rakentamiseen on olemassa erilaisia algoritmeja ja lähestymistapoja (mm. Szegedy ym., 2015; Papernot ym., 2016b; Shi ym., 2017), joiden avulla havaintoon voidaan rakentaa minimaalisin väärinluokitteluun johtava muutos. Szegedy ym. (2015) huomasivat tutkimuksessaan hyökkäysesimerkkien olevan tyypillisimmin juuri alisovittuvien mallien ongelma.

Papernot, McDaniel ja Goodfellow (2016a) esittelevät mustan laatikon lähestymistavan hyökkäysesimerkkien laatimiseen, jossa tutkittavan mallin toi-

mintaa tai sen arkkitehtuuria ei ennalta tunneta, ja siitä voidaan selvittää vain havainnoille annettavia luokka-arvoja. Hyökkäyksen kohteena on liikennemerkkejä luokitteleva malli, jota voitaisiin hyödyntää esimerkiksi itseohjaavissa autoissa. Menetelmä hyödyntää syviä neuroverkkoja, jotka pilkkovat luokitteluongelman pieniin, keskenään verkottuneisiin yksinkertaisiin päätöksiin eli neuroneihin, joiden välisillä yhteyksillä on eri painoarvoja riippuen päätöksen merkityksestä lopputulokseen. Ulostulokerroksella luokitin kertoo jokaisen havainnon todennäköisyyden kuulua kuhunkin luokkaan (Papernot ym., 2016a; Shi ym., 2017.)

Papernotin ym. (2016a) luoman neuroverkon tarkoituksena on rakentaa funktionaalisesti vastaava luokitin ilman ohjattujen oppimismenetelmien vaatimaa valtavaa harjoitusdatan määrää. Neuroverkolle syötettävän harjoitusdatan perustana on hyvin pieni edustava otos, jonka havaintojen luokat tiedustellaan online-luokittimelta. Aineistoa kasvatetaan synteettisesti, jotta tutkittavan luokittimen päätösraja (decision boundary) saadaan arvioitua niin, ettei lähetettyjen kyselyjen määrä kasva epäilyttäväksi. Papernot ym. (2016a) testasivat menetelmäänsä Googlen ja Amazonin koneoppimisrajapinnoilla. Raportoitu onnistumisprosentti Googlen luokitinta vastaan oli 96,19% ja Amazonin luokitinta vastaan 88,94%.

Shin ym. (2017) tutkimuksessa hyödynnettiin samaa lähestymistapaa algoritmin varastamiseen. Kohteena olivat kahdella eri koneoppimisalgoritmilla harjoitetut tekstiä kategorisoivat luokittimet, joiden funktionaalisuudet onnistuttiin kopioimaan 91- ja 84-prosenttisesti vain kymmenellä havainnolla harjoitettuna. Yli 97% vastaavuus saavutettiin, kun neuroverkko koulutettiin puolikkaalla harjoitusaineistolla.

5 YHTEENVETO

Tutkimuskysymykset, joihin tutkielma pyrki vastaamaan, olivat: *"millä tavoin koneoppiminen voi olla kyberhyökkäyksen kohteena?"* ja *"voiko koneoppimismenetelmiä hyödyntää kyberhyökkäyksissä?"*. Lähdekirjallisuus osoitti, että koneoppimista on mahdollista johtaa harhaan ja väärinkäyttää usealla eri tavalla. Kausatiivisilla hyökkäyksillä kohteena olevasta järjestelmästä voidaan tuottaa hyödyttömän epätarkka ja koneoppimismallin toimintaa tutkivilla hyökkäyksillä voidaan löytää hyödynnettäviä haavoittuvuuksia. Online-luokittimet vuotavat herkästi tietoa toiminnastaan ja käyttämästään harjoitusdatasta, jolloin hyökkääjän on mahdollista selvittää mallin päätösraja (Ateniese ym., 2015). Hyökkääjä voi myös varastaa verkossa saatavilla olevan algoritmin kouluttamalla oman mallinsa keräämään kriittistä tietoa kohteen toiminnasta. Kyberpuolustusjärjestelmien luonteesta johtuen, suuri osa koneoppimista hyödyntävistä hyökkäysmenetelmistä rajoittuu syötteiden väärinluokitteluun.

Kyberturvallisuus on reaktiivista: järjestelmiä ja ohjelmia pidetään turvallisina niin kauan, kunnes uusia haavoittuvuuksia hyödynnetään hyökkäyksissä. Kyberturvallisuusjärjestelmät osaavat jo puolustautua lukuisilta erilaisilta hyökkäyksiltä, mutta taustalla piilee loputon kamppailu uusia haavoittuvuuksia etsiviä kyberrikollisia vastaan. Kyberpuolustusjärjestelmien kehittyessä yhä tarkemmiksi on odotettavissa, ettei ihmisiin pohjautuvien haavoittuvuuksien suosio tule laskemaan. Lisäksi, koneoppimisella tuettujen käännöspalveluiden parantuessa, tulevaisuudessa nähdään yhä uskottavampia ja tehokkaampia tietojenkalasteluhyökkäyksiä.

Yhtenä suurimpana kansainvälisenä uhkatekijänä, potentiaalisen vahingon laajuuden vuoksi, pidetään hyökkäyksiä kriittiseen infrastruktuuriin (mm. Sisäministeriö, 2017; Von Solms & Van Niekirk, 2013). Myös Symantec (2018) tunnistaa kriittiseen infrastruktuuriin kohdistettujen hyökkäysten suosion kasvun ja odottaa trendin kiihtyvän entisestään vuonna 2018. Koneoppimismenetelmien kehittyessä niin puolustusmekanismit kuin kyberhyökkäyksetkin tehostuvat ja saavat uusia muotoja. Shin ym. (2017) ja Papernotin ym. (2016a) tutkimukset osoittavat, että koneoppimisen käyttö kyberhyökkäyksissä on hälyttävän tehokasta jo hyvin pienellä harjoitusaineistolla. Tämän vuoksi on odotet-

tavissa, että koneoppimismenetelmien käyttö myös kyberhyökkäyksen välineenä yleistyy lähitulevaisuudessa. Tärkeitä jatkotutkimuskohteita tulevaisuudessa ovatkin koneoppimisen väärinkäyttömahdollisuuksien kartoittaminen sekä koneoppimismenetelmien hyödyntäminen haavoittuvuuksien havaitsemisessa. Koneoppimisen turvaaminen ja kehittäminen ovat avainroolissa tulevaisuuden kybertoimintaympäristön suojaamisessa.

LÄHTEET

Alpaydin, E. (2016). *Machine learning: The New AI* (3. uud. painos). London, England: MIT Press.

Alpaydin, E. (2014). *Introduction to Machine Learning* (2. uud. painos). London, England: The MIT Press, 26–27, .

Ateniese, G., Felici, G., Mancini, L. V., Spognardi, A., Villani, A. & Vitali, D. (2015). *Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers.* International Journal of Security and Networks, 10(3), 137-150.

Bai, E. W. (2014). *Big data: The curse of dimensionality in modeling.* Proceedings of the 33rd Chinese Control Conference, CCC 2014, 6–13.

Barreno, M., Nelson, B., Joseph, A. D. & Tygar, J. D. (2010). *The security of machine learning.* Machine Learning, 81(2), 121–148.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D. & Tygar, J. D.(2006). *Can machine learning be secure?* Proceedings of the 2006 ACM Symposium on Information, computer and communications security, 16-25.

Barto, A. & Dietterich, T. (2004). *Reinforcement learning and its relationship to supervised learning.* Handbook of Learning and Approximate Dynamic Programming, 47–64.

Blum, A. L. & Langley, P. (1997). *Artificial Intelligence Selection of relevant features and examples in machine*, *97*(97), 245–271.

Buczak, A. & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, PP(99), 1.

Er, M. J., Ave, N. & Wang, N. (2016). *Deep Semi-supervised Learning Using Multi-Layered Extreme Learning Machines*. The 6th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, 457–462.

Fraley, J. B. & Cannady, J. (2017). *The promise of machine learning in cybersecurity.* Conference Proceedings - IEEE SOUTHEASTCON.

Feng, C., Wu, S., & Liu, N. (2017). *A user-centric machine learning framework for cyber security operations center.* IEEE International Conference on Intelligence and Security Informatics, 173–175.

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2014). *Explaining and Harnessing Adversarial Examples*. 2015 International Conference on Learning Representations, 1–11. ArXiv.

Iglesias, F. & Zseby, T. (2015). *Analysis of network traffic features for anomaly detection*. Machine Learning, (December 2013), 59–84.

Kotsiantis, S. B. (2007). *Supervised machine learning: A review of classification techniques*. Informatica Ljubljana, 31(3), 249-268.

Lewis, J. (2018). Economic Impact of Cybercrime—No Slowing Down. Center for Strategic and International Studies, McAfee. Haettu 14.5.2018

osoitteesta https://www.mcafee.com/us/resources/reports/restricted/economic-impact-cybercrime.pdf.

Limnéll, J., Majewski, K. & Salminen, M. (2014). *Kyberturvallisuus, 29, 113-150.* Jyväskylä: Docendo.

Livadas, C., Walsh, R., Lapsley, D., & Strayer, W. T. (2006). Using machine learning techniques to identify botnet traffic. Proceedings - Conference on Local Computer Networks, LCN, (December), 967–974.

Michalski, R., Carbonell, J. & Mitchell, T. (2013). *Machine Learning: An Artificial Intelligence Approach*. Springer Publishing Company.

Myung, J. I. (2000). *The importance of complexity in model selection*. Journal of Mathematical Psychology, 44(1), 190–204.

Müller, K. R., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001). *An introduction to kernel-based learning algorithms.* IEEE Transactions on Neural Networks, 12(2), 181–201.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I. P., Saini, U., … Xia, K. (2008). *Exploiting machine learning to subvert your spam filter.* In Proceedings of the First Workshop on Large-Scale Exploits and Emerging Threats, (7. artikkeli).

Papernot, N., McDaniel, P. & Goodfellow, I. (2016a). *Practical Black-Box Attacks against Machine Learning.* Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security. ArXiv.

Papernot, N., Mcdaniel, P., Jha, S., Fredrikson, M., Celik, Z. B. & Swami, A. (2016b). *The limitations of deep learning in adversarial settings.* Proceedings - 2016 IEEE European Symposium on Security and Privacy, 372–387.

Portugal, I., Alencar, P. & Cowan, D. (2017). *The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review*. Expert Systems with Applications, 97, 205–227.

Puolustusministeriö (2013). *Suomen kyberturvallisuusstrategia.* Forssa print.

Samuel, A. (1959). *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development.

Shi, Y., Sagduyu, Y. & Grushin, A. (2017). *How to Steal a Machine Learning Classifier with Deep Learning.* 2017 IEEE International Symposium on Technologies for Homeland Security.

Sisäministeriö (2017). *Tietoverkkorikollisuuden torjuntaa koskeva selvitys.* Sisäministertön julkaisu, 14. Sisäministeriö, Helsinki.

Sommer, S. & Paxson, V. (2010). *Outside the Closed World: On Using Machine Learning For Network Intrusion Detection.* 2010 IEEE Symposium on Security and Privacy, 305-316.

Symantec Corporation. (2018). *ISTR Internet Security Threat Report, 23*. Haettu 10.5.2018 osoitteesta http://images.mktgassets.symantec.com/Web/Symantec/%7B3a70beb8-c55d-4516-98ed-1d0818a42661%7D_ISTR23_Main-FINAL-APR10.pdf?aid=elq_16379

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). *Going deeper with convolutions.* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1–9.

Uma, M. & Padmavathi, G. (2013*). A Survey on Various Cyber Attacks and their Classification.* International Journal of Network Security, 15, 390-396.

Von Solms, R. & Van Niekerk, J. (2013). *From information security to cyber security.* Computers and Security, 38, 97–102.

Zanero, S. & Serazzi, G. (2008). *Unsupervised learning algorithms for intrusion detection.* Network Operations and Management Symposium 2008 NOMS 2008 IEEE, 1043–1048.