Syväväärennösten luomat tietoturvauhat sosiaalisen median alustoilla

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Toukokuu 2026
Riku Auranen

TURUN YLIOPISTO
Tietotekniikan laitos

Riku Auranen: Syvävääärennösten luomat tietoturvauhat sosiaalisen median alustoilla

TkK-tutkielma, 22 s.
Tietotekniikka
Toukokuu 2026

Syvävääärennösteknologian nopean kehityksen myötä sosiaalisen median alustoille on syntynyt uusia tietoturvahaasteita. Syvävääärennösteknologia perustuu neuroverkkoihin ja niiden sovelluksiin, joiden avulla voidaan luoda aidolta vaikuttavaa synteettistä mediaa. Syvävääärennöksiä voidaan hyödyntää sosiaalisen manipulaation keinona erilaisissa tietoturvahyökkäyksissä sosiaalisen median käyttäjiä vastaan. Tämä tutkielma on kirjallisuuskatsaus, jonka tarkoituksena on selvittää, millaisia tietoturvauhkia syvävääärennökset tuottavat sosiaalisen median alustoilla sekä miten niiltä voidaan puolustautua.
Tutkimustuloksissa käy ilmi, että syvävääärennöksiä tuotetaan lähes ainoastaan identiteetin varastamisen sekä sosiaaliseen manipulaatioon liittyviin käyttötarkoituksiin. Tuloksiset osoittavat, että syvävääärennösten uhkia on mahdollista vähentää lisäämällä käyttäjien tietoisuutta syvävääärennöksistä ja hyödyntämällä autentikointimenetelmiä. Keskeisiä autentikointimenetelmiä ovat digitaalisten vesileimojen asettaminen sosiaalisen median julkaisuihin sekä tunnistusalgoritmien hyödyntäminen. Näihin menetelmiin liittyy haasteita, kuten niiden luotettavuus, alustojen haluttomuus puuttua ongelmaan sekä tunnistusalgoritmien hidas kehitys.

Asiasanat: syvävääärennös, sosiaalinen media, tietoturva, tietoturvauhka

Sisällys

Kuvat

Taulukot

1 Johdanto

Syväväärennös (engl. Deepfake) on tekoälysovelluksella luotu äärimmäisen aidon oloinen synteettinen valokuva, video tai ääniraita. Syväväärennös voidaan luoda syöttämällä esimerkiksi sosiaalisen median alustalta kerätty valokuva tekoälysovellukselle, joka luo tekijän haluaman kuvan syötetyn datan perusteella. Syväväärennöksissä tekoälyllä voidaan yhdistää uhrin kasvot haluttuun kontekstiin, mikä mahdollistaa useita rikollisia käyttötarkoituksia syväväärennösteknologialle. [1]

Vuonna 2024 sosiaalisen median alustoilla, kuten Facebookissa, Instagramissa ja TikTokissa oli arviolta 5.22 miljardia käyttäjää. Tämä kansainvälinen ja laaja sosiaalinen verkosto tarjoaa kyberrikollisille alustan rikollisille toiminalle. Rikollisessa toiminnassa voidaan hyödyntää syväväärennöksiä esimerkiksi sosiaaliseen manipulaatioon, joka on keskeinen käyttötarkoitus syväväärennöksille. Syväväärennöksien avulla toteutettu sosiaalinen manipulaatio luo suuren tietoturvariskin etenkin medialukutaidottomille väestönryhmälle. [2]

Tekoälyllä tuotetun median kasvava osuus sosiaalisessa mediassa korostaa tietoturvan tärkeyttä. Tietoturvalla tarkoitetaan tiedon, järjestelmien sekä palveluiden suojaamista luvattomalta käytöltä ja häiriöiltä. Sen tavoitteena on toteuttaa CIA-kolmion (engl. CIA-triad) määritelmä, jossa varmistetaan järjestelmän tiedon Luottamuksellisuus (engl. Confidentiality), Eheys (engl. Integrity) sekä Saatavuus (engl. Availability). Kyberturva puolestaan laajentaa näkökulmaa koko digitaaliseen ympäristöön ja siihen kohdistuviin uhkiin, kuten erilaisiin verkkohyökkäyksiin tai tieto-

murtoihin. Sosiaalinen median yleistyminen lisää tieto- ja kyberturvan merkitystä, sillä alustoille jaetaan suuria määriä henkilökohtaisia tietoja käyttäjistä itsestään sekä heidän tutuistaan. Tämä lisää väärinkäytön riskiä ja luo ympäristön uhille, kuten syväväärennöksille, joilla voidaan vaikuttaa tiedon eheyteen ja luotettavuuteen. [3]

Tässä tutkielmassa perehdytään syväväärennöksiin ja siihen miten ne toimivat, miten niitä voidaan havaita ja millaisia uhkia ne voivat luoda. Tutkielmassa perehdytään myös syväväärennösten luomiin tietoturvariskeihin sosiaalisen median alustoilla. Tämän työn tutkimuskysymykset ovat:

- Tk1: Mihin tarkoituksiin syväväärennöksiä käytetään sosiaalisen median alustoilla ja millaisia uhkia ne aiheuttavat?
- Tk2: Miten syväväärennöksiä voidaan tunnistaa sosiaalisessa mediassa?
- Tk3: Millaisilla keinoilla voidaan puolustautua syväväärennöksiin liittyviltä tietoturvauhilta sosiaalisessa mediassa?

Tämä kandidaatintutkielma on toteutettu kirjallisuuskatsauksena. Lähteenä on käytetty pääasiassa tieteellisiä ja akateemisia artikkelita. Akateemisten lähteiden lisäksi tietoa on kerätty IBM:n ja Euroopan rauhaninstituutin julkaisuista. Taulukon 4.1 akateemisten lähteiden tiedonhaku on suoritettu kolmesta tietokannasta: IEEE Xplore, ScienceDirect ja Web Of Science. Hakusanoina on käytetty ”deepfake”, ”synthetic media”, ”social media platform”, ”cybersecurity”, ”cyber security” ja ”cybersecurity threat”. Hakulauseen tuloksesta valittiin tutkielmaa tukevat artikkelit, joita rajattiin ensin otsikon sekä myöhemmin abstraktin ja sisällön perusteella. Taulukkoon 4.1 valikoitui 16 artikkelia, joilla vastataan tutkimuskysymyksiin. Akateemisten tietolähteiden kriteereissä on huomioitu artikkelin julkaisuvuosi, sillä syväväärennösteknologia on nopeasti kehittyvä konsepti, minkä vuoksi tiedon ajankohtaisuus on keskeinen kriteeri.

Tutkielman toinen luku käsittelee syväväärennösteknologian taustaa. Luvussa esitellään teeman kattotermejä sekä syväväärennöksen tuottamiseen käytettyjä yleisiä teknologioita. Kolmannessa luvussa käsitellään syväväärennösten yleisiä käyttökohteita ja niiden vaikutuksia. Neljäs luku toimii tutkielman tuloslukuna, jossa puretaan taulukon 4.1 tulokset teemojen mukaisesti. Viides luku toimii tutkielman pohdintaosuutena. Tutkielman kuudennessa luvussa vastataan tutkimuskysymyksiin aikaisempien lukujen perusteella ja esitetään yhteenveto tutkielmasta.

2 Syväväärennös ja käytetyt teknologiat

2.1 Syväväärennöksen toiminta

Syväväärennös termi muodostuu englannin kielen sanoista ”deep learning” ja ”fake”, joilla tarkoitetaan tekoälysovelluksella tuotettua synteettistä sisältöä. Syväväärennös voidaan luoda esimerkiksi syöttämällä kaksi erilaista valokuvaa, ääniraitaa tai videota tekoälysovellukseen, joka yhdistää nämä datat synteettiseksi mediaksi. [4] Syväväärennökset saivat alkunsa, kun Reddit-verkkoyhteisössä käyttäjä nimimerkiltään ”Deepfakes” loi vuonna 2017 keskustelualueen, mihin käyttäjä jakoi synteettistä pornografista materiaalia julkisuuden henkilöistä [5]. Tämän jälkeen syväväärennöksille on muodostunut erilaisia käyttötarkoituksia. Syväväärennöksille ajankohtaisia käyttötarkoituksia ovat esimerkiksi poliittinen vaikuttaminen väärennettyjen vidoiden ja valokuvien avulla sekä valheellisen median ja misinformaation levittäminen. Syväväärennöksillä voidaan vaikuttaa voimakkaasti yhteiskuntaan nykypäivänä, sillä syväoppivia tekoälysovelluksia on paljon tarjolla julkisesti kaikkien saatavilla, minkä takia syväväärennöksiä voidaan tehdä helposti ilman syvempää tietoteknistä osaamista. [6]

Tekoäly (engl. artificial intelligence) on tietojenkäsittelytieteen osa-alue, jonka tavoitteena on kehittää tietojärjestelmiä, jotka kykenevät simuloimaan inhimillisiä

tehtäviä. Tälläisiä tehtäviä ovat esimerkiksi oppiminen, suunnittelu, kuvantunnistus ja äänentunnistus. Näillä tehtävillä pyritään saamaan tietokone ihmisen ajattelukyvyn tasolle, jotta tietokoneilla saataisiin suoritettua korkeamman tason tehtäviä eri elämän osa-alueilla, kuten esimerkiksi työelämässä. Tekoälyn kehitys on viime vuosina nopeutunut merkittävästi suurien investointejen, datamäärien saatavuuden ja laskentatehon kasvun myötä. [7]

Syväoppiminen (engl. Deep learning) on koneoppimisen osa-alue, joka perustuu keinotekoisiin neuroverkkoihin (engl. artificial neural network), jotka kykenevät oppimaan tietojen esitystapoja. Toisin kuin yleisesessä koneoppimisessa, syväoppiminen oppii piirteitä automaattisesti raakadatan perusteella ilman, että niitä tarvitsee määritellä käsin. Syväoppimismenetelmiä voidaan soveltaa esimerkiksi monikerroksisiin syväoppimisarkkitehtuureihin, kuten konvoluutioneuroverkoihin (engl. convolutional neural network), jotka oppivat automaattisesti piirteitä raakadatan, kuten käsittelemättömien valokuvien perusteella. Tämä tekee syväoppimisesta erityisen tehokkaan menetelmän monimutkaisten tehtävien suorittamisessa, kuten kuvien ja videoiden käsittelyssä. [8]

2.2 Syvävääarennösten teknologia

Neuroverkot (engl. Neural network) ovat keskeinen tekijä syvävääarennösteknologiassa. Neuroverkkojen toimintaperiaate vastaa aivojen biologisen tason toimintaa. Aivot koostuvat hermosoluista, jotka vastaanottavat ja välittävät sähköisiä ja kemiallisia signaaleja aksonien kautta. Neuroverkot muodostuvat toisiinsa kytketyistä solmuista, jotka käsittelevät tietoa. Neuroverkot toimivat vastaanottamalla syötteitä, käsittelemällä ne toisiinsa kytkeytyneiden osien kautta ja tuottamalla tuloksia näiden laskentaprosessien perusteella. Neuroverkon perusrakenne koostuu syötekerroksesta, yhdestä tai useammasta piilokerroksesta sekä ulostulokerroksesta. Piilokerrosten määrä vaihtelee verkoston tyypin mukaan. Neuroverkot pyrkivät oppimaan

datasta ja parantamaan ennustetarkkuuttaan asteittain. Kun neuroverkko saavuttaa optimaalisen suorituskyvyn, siitä tulee tärkeä tekoälyn toiminnan mahdollistava työkalu, joka mahdollistaa monimutkaisienkin ongelmien ratkaisemisen. [9]

Generatatiivisen kilpailevan verkon (engl. Generative adversarial network, GAN) toiminta perustuu kahden kilpailevan neuroverkon toimintaan, jotka ovat generoiva verkko ja luokitteleva verkko. Generoivan verkon tehtävänä on yrittää huijata luokittelevaa verkkoa tuottamalla synteettistä sisältöä, kun taas luokittelevan verkon tehtävänä on arvioida tuotetun sisällön aitoutta vertailemalla oikeata otosta synteettiseen otokseen. GAN-verkon tavoitteena on, että generoiva verkko oppii tuottamaan tarpeeksi laadukkaita väärennöksiä koulutusdatasta, jotta luokitteleva verkko ei kykenise enää erottamaan väärennettyjä kuvia tai videoita aidosista. [10]

Enkoodaus-Dekoodaus-verkot (engl. Encoder-Decoder-network, ED) verkot koostuvat vähintään kahdesta osasta, jotka ovat enkoodaaja ja dekoodaaja. Enkooderi muuntaa syötteen tiivistettyyn esitysmuotoon, jonka perusteella dekooderi tuottaa ulostulon. Syväväärennösteknologiassa voidaan hyödyntää useita enkoodereita ja dekoodereita sekä muokata enkoodattuja esityksiä halutun lopputuloksen saamiseksi. Mikäli enkooderi ja dekooderi ovat rakenteeltaan symmetrisiä ja verkon tavoitteena on tuottaa ulostulo, joka vastaa alkuperäistä syötettä, jolloin verkkoa kutsutaan autoenkooderiksi (engl. autoencoder). [10] Autoenkoodereihin perustuvissa syväväärennösmenetelmissä lähde- ja kohdekuvat käsitellään yhteisen enkooderin avulla, joka oppii piirteitä lähdekuvasta. Tämän jälkeen dekooderit rekonstruoivat kuvat joko lähde- tai kohdehenkilön piirteillä. Kun lähdekuvan esitys syötetään kohdekuvan dekooderille, syntyy kuva, jossa kohdekuva sisältää lähdekuvan visuaalisia piirteitä. Tämä menetelmä on yksi yleisimmistä tavoista uottaa syväväärennyskuvia. [1] Autoenkoodausverkoista käytetään myös variaatioautoenkoodereita (variational autoencoder, VAE). Variaatioautoenkooderit soveltuvat paremmin syväväärennösten generointiin kuin perinteiset autoenkooderit, sillä niiden latenttitila on erotel-

lumpi. Tämän vuoksi enkoodaukset mahdollistavat sujuvamman interpoloinnin ja tehokkaamman kvuien muokkaamisen. [10]

Diffuusiomallit (engl. Diffusion models) ovat syväväärennösten generoimiseen hyödynnettyjä teknologioita. Diffuusiomalli perustuu kohinanpoistoon perustuvaan diffuusiotodennäköisyysmalliin (engl. denoising diffusion probabilistic model, DDPM), joka on parametrinen Markovin ketju, jossa kuvia parannetaan vähentämällä niistä kohinaa vaiheittain. Diffuusiomallit koostuvat siis kahdesta päävaiheesta. Ensimmäisessä prosessissa kuvatietoihin lisätään kohinaa niin, että kuvasta tulee täysin satunnaista kohinaa. Toisessa prosessissa kohinaa poistetaan vaihe vaiheelta, jolloin saadaan tuotettua korkealaatuisia kuvatiedostoja. [11]

Konvoluutioneuroverkot (engl. Convolutional neural network, CNN) ovat syväoppimismalleja, jotka ovat suunniteltu käsittelemään rakenteellista ruutumaista dataa, kuten digitaalisia valokuvia [12]. CNN:n konvoluutiokerrokset toimivat suodattimilla, joita siirrellään syötteen yli tuottamaan abstrakti piirrekartta (engl. feature map). Verkkojen syvetessä poolauskerroksia käytetään vähentämään esityksen ulottuvuuksia, kun taas ylinäytteistyskerrokset (engl. up-sampling layers) kasvattavat niitä uudelleen. Näiden kerrosten avulla voidaan rakentaa autoenkooderi pohjaisia CNN-verkkoja, joita hyödynnetään erityisesti kuvankäsittelyssä kuten syväväärennöksissä. [10]

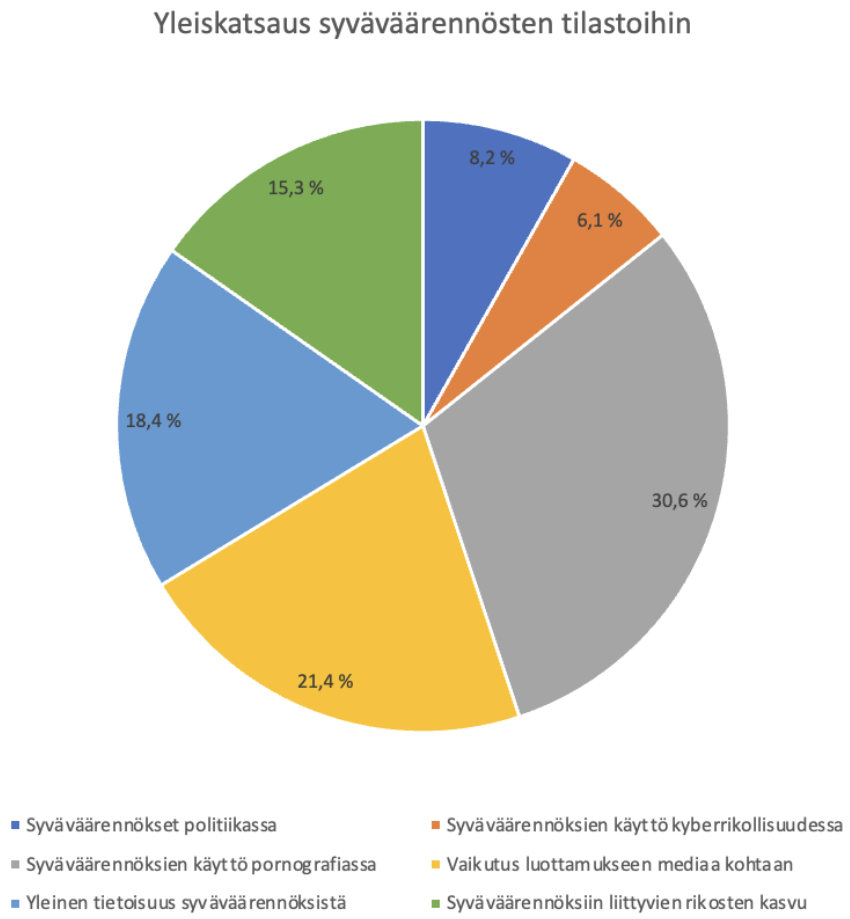
Tietoaineistot (engl. Dataset) ovat kokoelma kerättyä dataa jaoteltuna erilaisiin taulukoihin tai tiedostomuotoihin. Tietoaineistoja käytetään koneoppimismallien ja tekoälyn kouluttamiseen. [13] Syväväärennösten generointiin ja havaitsemiseen on olemassa useita valmiita tietoaineistoja. Generointiin voidaan hyödyntää esimerkiksi FaceSwap tietoaineistoa, jossa hyödynnetään GAN-verkkoa ja autoenkooderia. FaceSwap perustuu autoenkooderimalliin, jossa kuva ensin muunnetaan tiivistettyyn matemaattiseen malliin, minkä jälkeen se rakennetaan uudelleen siten, että lopputuloksena on alkuperäisen kuvan ilme ja rakenne, mutta kohteelle saadaan uusi

identiteetti. Syvävääreennösten havainnointiin voidaan hyödyntää modernia Celeb-DF tietoaineistoa, joka sisältää korkealaatuista videomateriaalia. Parantunut visuaalinen laatu tekee manipulaation havaitsemisesta haastavaa, mikä lisää aineiston soveltuvuutta syvävääreennösten tunnistusalgoritmien kouluttamiseen ja arviointiin. [14]

3 Syväväärennösten käyttökohteet

Syväväärennöksillä on useita käyttötarkoituksia, jotka voidaan luokitella positiivisiksi ja negatiivisiksi käyttötarkoituksiksi. Esimerkki positiivisesta käyttötarkoituksesta on syväväärennösten hyödyntäminen jälkiäänityksiin elokuvateollisuudessa, sillä syväväärennösteknologialla on mahdollista yhdistää jälkiäänitykset videoon sulavasti. Vaikka syväväärennöksillä on positiivisia käyttötarkoituksia, sitä kuitenkin käytetään pääosin haitallisiin ja epäeettisiin tarkoituksiin. [10] Syväväärennösteknologia on voimakas työkalu valheellisen tiedon levittämiseen, koska sillä voidaan pyrkiä rikkomaan yleistä luottamusta yksilö -ja yhteiskunnallisella tasolla. Syväväärennöksillä voidaan vaikuttaa yksityishenkilön elämään ja maineeseen esimerkiksi identiteetin varastamisella, jolloin hänestä voidaan tehdä ei-suostumuksellista synteettistä mediamateriaalia. [15] Yhteiskunnallisella tasolla syväväärennökset voivat esimerkiksi vaikuttaa mielipiteesiin poliittisissa ympäristöissä [5].

Sosiaalissa manipuloinnissa (engl. Social engineering) voidaan hyödyntää syväväärennöksiä, kun uhria johdatellaan tekemään päätöksiä tunteisiin tai auktoriteettiin vetoamalla [16]. Syväväärennösteknologialla voidaan vahvistaa ilmiötä tuottamalla aidon oloista video- tai äänimateriaalia, jolla voidaan helpommin tehdä manipulaatiosta uskottavampaa. Manipulaatio voi perustua kiireellisyyden tunteeseen, luottamussuhteeseen tai auktoriteettiasemaan. Sosiaalisessa manipulaatiossa uhrille voidaan esittää uskottava viesti esimerkiksi perheenjäseneltä, jolloin uhrin rationaalinen harkintakyky heikentyy. [5]



Kuva 3.1: Tilastoja syväväärennösten käyttökohteista [17]

Syväväärennösten hyödyntäminen harhaanjohtavan informaation luomisessa vaikeuttaa aidon tiedon ja manipuloidun sisällön erottamista toisistaan, mikä hämärtää totuuden ja valheellisen tiedon välistä rajaa. Syväväärennösten avulla voidaan luoda täysin keksittyjä tai muokattuja tapahtumakuvauksia, joilla yritetään johtaa lukijoita harhaan. Tämän seurauksena virheellisen tai täysin keksityn tiedon ja uutisten määrä voi kasvaa jatkuvasti. Pitkällä aikavälillä tämä heikentää luottamusta suuriin medialähteisiin sekä julkisuuden henkilöihin. [18] Sosiaalisen median alustat tarjoavat vaihtoehtoisen uutisten lähteen, jossa valheellisten uutisten leviäminen on huomattavasti helpompaa ja nopeampaa kuin perinteisten uutiskanavien sivustoilla. Valheelliset uutiset ovat nopean leviämiskyvyn vuoksi alttiita erilaisille ongelmille, kuten vihapuheelle, huhujen levittämiselle, rasismille sekä ennakkoluuloille.

Erityisesti kehittyvissä maissa tämä muodostaa merkittävän haasteen, sillä sosiaalisen median kautta leviävän valheellisen tiedon on havaittu joissakin tapauksissa edistäneen yhteiskunnallisia konflikteja ja pahimmassa tapauksessa aiheuttanut sotatoimia maassa. [19], [20]

Syvävääreennösten nopea kehittyminen aiheuttaa ongelmia politiikassa ja äänestyskäyttäytymisessä, sillä syvävääreennösten avulla voidaan manipuloida vaikutusvaltaisten poliittisten henkilöiden puheita, lausuntoja ja julkisia esiintymisiä. Tälläinen toiminta voi uhata demokraattisen järjetelmän legitimiteettisyyttä. Manipuloidun materiaalin levitys saattaa muokata äänetäjien mielipiteitä, mikä saattaa vaikuttaa vaalituloksiin. [18] Esimerkiksi Slovakian vaaliehdokas Michal Šimečka joutui syvävääreennöksen uhriksi. Syvävääreennöksessä Šimečka väittää peukaloineen vaalituloksia, vaikka ei ole sitä tehnyt. Tämän vuoksi Slovakialaiset alkoivat epäilemään hänen luotettavuuttaan ja tuomitsemaan hänen korruptiotaan, mikä vaikutti lopulliseen vaalitulokseen. [5] Valheellisella medialla voidaan aiheuttaa tilanteita, joissa poliittinen tai muuten vaikutusvaltainen henkilö vaatii mielenosoituksia tai jopa väkivaltaisuuksia [15].

Häirintä, kiusanteko sekä identiteetin varastaminen ovat myös keskeisiä syvävääreennöksien käyttötarkoituksia. Vuonna 2017 internetissä levisi lukuisia syvävääreennösteknologioilla luotuja videoita, joissa naisnäyttelijöiden kasvoja oli yhdistetty pornograafisiin videoihin. Videoilla on yritetty kiusata ja tuhota julkisuudessa esiintyviä näyttelijöitä ja heidän maineitaan. [21] Häirintä ei kuitenkaan rajoitu vain näyttelijöihin tai julkisuuden henkilöihin, sillä yhä useammin uhriksi joutuu tavallinen sosiaalisen median käyttäjä. Vuonna 2023 Etelä-Koreassa toteutettiin miehille kohdistettu kysely, jossa todettiin että 63 % vastanneista haluaisivat käyttää syvävääreennysteknologiaa, jotta voisivat riisua oikeassa elämässä tuntemiaan naisia. Vain 6 % kyslyyn vastanneista vastasi, että ei haluaisi käyttää syvävääreennösteknologiaa tähän tarkoitukseen. [22] Etelä-Koreassa syvävääreennöksiä on jaettu Telegram alus-

talla, jossa syvävääreennösten tekijä on syvävääreennöksen lisäksi jakanut uhrin oikeita henkilötietoja kuten nimen, puhelinnumeron, kotiosoitteen tai työpaikan, mikä on aiheuttanut uhrille huolta omasta turvallisuudestaan. Monet uhrit tuntevat rikoksentekijän entuudestaan. Esimerkiksi nuori Etelä-Korealainen peruskoulun opettaja joutui syvävääreennöksen kohteeksi, kun hänen luokkakaverinsa peruskoulusta julkaisi syvävääreennöksen hänestä Telegram alustalla avainsanalla ”tuttavien nöyryytys”. [22]

Syvävääreennöksiä voidaan hyödyntää huijausten ja petosten toteuttamisessa. Tunnetuin ensimmäisistä syvävääreennöksiä hyödyntävistä taloudellisista huijauksista on menetelmä, jossa huijarit tekevät äänisyvävääreennöksen ison energiayhtiön emoyrityksen toimitusjohtajasta. Äänisyvävääreennöksessä annettiin ymmärtää, että toimitusjohtaja pyysi työntekijältä puhelimitse, että tämän pitäisi siirtää 243,000 yhdysvaltain dollaria alihankkijalle, mutta rahat menivätkin huijarille. [5] Syvävääreennösteknologiaa voidaan käyttää myös pienempiin huijauksiin. Perusidea huijauksissa on, että luodaan syvävääreennösvideo tai ääniraita, jolla pyritään huijaamaan uhria, joka voisi lopulta maksaa huijarille. Syvävääreennöksillä voidaan myös kiristää uhria siten, että luodaan kiireellinen ongelma, jonka uhri saa raukeamaan maksamalla kiristäjän osoittamalla tavalla. [18] Huijaus voi olla esimerkiksi syvävääreennöksellä luotu ääniraita, jolla yritetään vakuuttamaan uhri, että hänen läheinen on siepattu ja hänet saa elävänä takaisin, jos maksaa rikolliselle tietyn summan rahaa [21].

4 Syväväärennösten tietoturvauhat sosiaalisessa mediassa

Syväväärennökset ja tekoälyllä generoidut synteettiset video- ja valokuvamateriaalit kehittyvät nopealla vauhdilla, mikä lisää haitallisen materiaalin leviämistä sosiaalisen median alustoilla. Haitallisen synteettisen materiaalin generoimiseen tarvittava data kerätään usein sosiaalisen median käyttäjien julkaisuista. Generoitu materiaali on erittäin aidon oloista, minkä vuoksi sitä hyödynnetään usein sosiaalisen manipulaation välineenä sekä kiusantekoon. Tässä luvussa käsitellään koko tutkielman keskiössä olevaa kysymystä: Millaisia tietoturvauhkia syväväärennökset luovat sosiaalisen median alustoilla, ja ketkä ovat erityisen alttiita niille. Lisäksi käsitellään sitä, miten syväväärennöksiä voidaan tunnistaa sekä miten niiltä voidaan puolustautua. Useimmat sosiaalisen median käyttäjistä tiedostavat synteettisen median olemassaolon, mutta monet eivät kuitenkaan tiedä miten niitä luodaan, tai eivät kykene erottamaan synteettistä mediaa oikeasta mediasta. Tässä luvussa analysoidaan kirjallisuuskatsauksen perusteella valittuja artikkeleita, jotka ovat koottu taulukkoon 4.1.

Taulukko 4.1: Tulostaulukko

Kirjoittaja & vuosi	T1: Käyttötarkoitus	T2: Some-uhat	T3: Tunnistaminen	T4: Puolustautuminen	T5: Tietoturva
Afroz et al., (2025) [5]	X		X		
Ali et al., (2022) [21]	X		X	X	
Ayub Khan et al., (2024) [23]	X	X	X	X	X
Chawla & Baraskar (2024) [12]	X	X	X		
Frolov et al., (2022) [4]	X		X	X	
Ji (2024) [22]	X	X		X	
Kasera et al., (2025) [9]	X	X	X	X	X
Khan et al., (2025) [24]	X	X		X	X
Laczi & Póser., (2024) [25]	X	X		X	
Mustak et al., (2023) [26]	X	X		X	
Narayanan et al., (2021) [27]		X	X		
Nowakowski (2024) [16]		X	X	X	X
Preu et al., (2022) [28]			X	X	
Saha et al., (2024) [1]	X				
Saheb et al., (2024) [29]	X	X		X	
Thuraisingham (2020) [30]	X	X	X	X	X

4.1 Syväväärennösten käyttötavat ja uhat sosiaalisessa mediassa

Aineiston perusteella syväväärennöksillä on useita käyttötarkoituksia sosiaalisessa mediassa, joista suurin osa liittyy haitalliseen toimintaan. Lisäksi aineistossa on käsitelty positiivisiakin käyttötarkoituksia kuten sitä, miten syväväärennöksiä voidaan hyödyntää markkinoinnissa kaupallisissa käyttökohteissa, tai miten syväväärennöksiä voidaan hyödyntää psykoterapiassa. Markkinoinnissa voidaan käyttää syväväärennösteknologialla tehtyjä sosiaalisen median vaikuttajia, joilla voidaan edustaa yrityksen brändiä eri alustoilla ilman, että yrityksen tarvitsee hyödyntää oikeita vaikuttajia. [25], [26] Kuitenkin lähes kaikissa artikkeleissa korostuu syväväärennösten negatiiviset käyttötarkoitukset. Aineiston mukaan syväväärennöksiä käytetään erityisesti kiusantekoon, kiristämiseen, valheellisen median levittämiseen, mainehaitan tekemiseen sekä huijaamiseen eri tavoilla. [5], [21], [22]

Useissa artikkeleissa kerrotaan miten syvävääreennöksiä käytetään huijauksiin, petoksiin ja tiedonkalastamiseen, joissa manipuloitua kuva- tai videomateriaalia käytetään uhrin harhauttamiseen [5], [21], [24], [27]. Tällaisissa tapauksissa rikollinen esiintyy usein toisena henkilönä, kuten uhrille tuttuna henkilönä. Uhrille tuttu henkilö voi olla esimerkiksi esihenkilö tai sukulainen, joista rikollinen luo uskottavaa synteettistä materiaalia. Synteettisellä materiaalilla rikollinen pyrkii huijaamaan uhria ja saamaan taloudellista hyötyä. Käyttäjämäärät sosiaalisen median alustoilla ovat kasvaneet suuriksi, minkä vuoksi rikollisten on nykyään helpompi löytää uhri syvävääreennöshuijauksille verrattuna menneisiin vuosiin. [26]

Lisäksi aineistossa keskeisenä käyttötarkoituna on kerrottu olevan misinformaation, valheellisten median ja uutisten levittäminen hyödyntäen syvävääreennöksiä [1], [5], [21], [23], [29]. Manipuloidulla medialla voidaan vaikuttaa yleiseen yhteiskunnalliseen mielipiteeseen, jolloin syvävääreennöksillä voidaan pyrkiä vaikuttamaan esimerkiksi vaalien lopputulokseen [1], [9]. Syvävääreennöksillä voidaan myös vaikuttaa uutissivustojen mediasisältöjen sekä yleisesti sosiaalisen median sisältöjen luotettavuuteen. Esimerkiksi, jos luotettavaksi luokitellulle alustalle julkaistaan synteettistä mediaa, niin uutissivuston luotettavuus heikkenee [23]. Suurimmilla sosiaalisen median alustoilla julkaisujen validiteettiä ei kontrolloida. Tämän vuoksi alustoilla on levinnyt runsaasti autenttiseksi luonnehdittua materiaalia, vaikka se olisikin ollut synteettistä mediaa kuten syvävääreennös, mikä heikentää median luotettavuutta. [23]

Useat tutkimukset tuovat esiin myös identiteetin varastamisen, häirinnän sekä kiusanteon keskeisenä käyttätarkoituksena [5], [21], [22], [25]. Syvävääreennösten avulla voidaan luoda mediasisältöä, jossa henkilö esitetään sanomassa tai tekemässä asioita, joita hän ei ole oikeasti tehnyt. Syvävääreennösten ensimmäisiä käyttökohteita oli ei-suostumuksellinen pornograafisen sisällön tuottaminen julkisuuden henkilöistä, mikä voidaan luokitella identiteenin väärentämiseksi. [5] Syvävääreennöstek-

nologialla tuotettu ei-suostumuksellinen pornograafinen sisältö ei aina kuitenkaan kohdistu pelkästään julkisuuden henkilöihin, vaan se voi kohdistua myös tavallisiin ihmisiin ja jopa alaikäisiin ihmisiin [5], [22], [25]. Identiteetin väärentämistä voidaan käyttää myös huijaustarkoitusiin, jolloin rikollinen voi esiintyä kenä tahansa henkilönä [5]. Aineistojen mukaan sosiaalisen median alustoilta voidaan kerätä melkein kenestä tahansa koulutusaineistoa syväväärennösteknologioille, jolloin identiteetin väärentäminen onnistuu helposti [9], [22].

4.2 Syväväärennösten tunnistaminen ja torjunta sekä tietoturva

Syväväärennösten tunnistaminen perustuu aistinvaraiseen visuaaliseen havainnointiin sekä syviä neuroverkostoja hyödyntäviin tunnistamisalgoritmeihin. Aineiston perusteella syväväärennöksiä voidaan tunnistaa syvä- ja koneoppimsalgoritmien ja biometrisen analyysen avulla. Äänisyväväärennöksen tunnistamiseen voidaan käyttää syväoppimisjärjestelmää, joka etsii äänestä epäsäännöllisyyksiä. Epäsäännöllisyyksiä voivat olla esimerkiksi äänimallit tai taustamelu. Järjestelmää voidaan kouluttaa tunnistamaan normaali ääni ja vertaamaan sitä syötteeseen, missä on epäilty syväväärennös. [5] Videosyväväärennöksiä voidaan analysoida kolmiulotteisen mittakaavan avulla, jolloin voidaan havaita epäsäännöllisyyksiä kasvonpiirteiden dynamiikassa. Kasvojen analysointiin hyödynnetään valmiita tietotaineistoja, kuten FaceForensics++. [5], [9] Kuvasyväväärennösten tunnistamista voidaan analysoida syöttämällä syväväärennös ja oikea kuva samasta henkilöstä syväoppimisalgoritmille. Syväoppimislagoritmi vertailee syötettä oikeaan kuvaan, jolloin tunnistustulos on tarkasti onnistunut. Vertailua voidaan suorittaa myös valmiiden tietoaineistojen avulla, mutta tällöin tulos ei välttämättä onnistu yhtä tarkasti. [5], [9] Aineiston perusteella syväväärennösten tunnistamismenetelmät kehittyvät nopeasti [12].

Aineistossa aistinvarainen visuaalinen havainnointi luetellaan keskeiseksi tunistusmenetelmäksi. Tunnistusmenetelmässä pyritään keskittymään syvävääreennösvideossa tai valokuvassa esiintyvien kasvojen yksityiskohtiin, jolloin materiaalista saattaa paljastua epäsäännöllisyyksiä. [21], [23] Usein epäsäännöllisyyksiä ilmenee syvävääreennösvideon tai valokuvan kasvoissa poskien, leuan sekä otsan alueilla, jolloin mainitut alueet voivat vaikuttaa luonnottoman sileiltä tai ryppyisiltä. Lisäksi syvävääreennösteknologiat usein epäonnistuvat tuottamaan realistisia kasvojen karvoituksia sekä huomioimaan silmälasien vaikutusta silmien alueella. Videosyvävääreennöksissä useasti huulien sekä silmien liikkeet ovat epäsäännöllisiä, mistä voidaan päätellä videon olevan syvävääreennös. [4], [9], [23]

Tunnistamisen lisäksi aineistossa korostetaan ennaltaehkäisevien menetelmien merkitystä. Useissa artikkeleissa esitetään tapoja autentikoida sosiaaliseen mediaan julkaistavien sisältöjen aitoutta. Yleisimpiä ehdotuksia olivat digitaalisten vesileimojen käyttäminen kaikissa julkaisuissa. [4], [9], [21] Aineiston perusteella toinen yleinen tunnistamiseen liittyvä menetelmä on, että materiaali syötetään manuaalisesti koneoppimisalgoritmille, joka tunnistaa epäsäännöllisyydet materiaalissa. Tällöin sosisaalisen median alustan työntekijä tai mielellään kolmannen osapuolen yritys kykenisi autentikoimaan materiaalin aitouden [21], [26]. Tunnistamisen lisäksi artikkeleissa kerrotaan ennaltaehkäiseviä menetelmiä valheellisen median uhille. Keskeisin menetelmä on erityisen alttiiden väestönryhmien perehdyttäminen liittyen syvävääreennösten uhkiin ja niiden torjumiseen. [25] Koulutusta pitäisi keskittää erityisesti haavoittuvassa asemassa oleville väestönryhmille kuten nuorille ja vanhuksille, sillä tutkimusten mukaan nuoret aikuiset kykenevät tunnistamaan syvävääreennyksiä paremmin kuin lapset ja vanhemmat ihmiset. [25], [28] Tämän lisäksi aineistossa kerrotaan myös ennaltaehkäiseksi menetelmäksi digitaaliesn jalanjäljen omatoimisen hallitsemisen sekä yleinen digitaalisesta hygieniasta huolehtimisen [16], [22].

Aineiston perusteella syvävääreennökset luovat merkittäviä tietoturvauhkia, jotka voivat kohdistua yksityishenkilöihin sekä yrityksiin. Aineistossa tietoturvaan liittyvät teemat koostuivat sosiaalisen manipulaation, tiedon luotettavuuteen sekä identiteetin varastamiseen liittyviin riskeihin. [5], [16] Useissa julkaisuissa käsitellään tietoturvasta tiedon eheyden heikentymisen näkökulmasta. Valheellinen media ja misinformaatio esittävät valheellista tietoa oikean tiedon muodossa, jolloin oikean ja valheellisen tiedon erottelu muuttuu vaikeammaksi. [5], [23] Aineistossa tuodaan esiin syvävääreennöksen vaikutuksia luottamuksellisuuteen identiteetin varastamisen ja väärinkäytön kautta. Uhrien kuvia ja ääntä voidaan käyttää luvatta, mikä voi johtaa yksityisyyden loukkauksiin. [25], [30] Syvävääreennösten keskeisin huijaukseen liittyvä tekniikka on sosiaalinen manipulaatio, jossa syvävääreennösteknologiaa hyödynnetään ihmisen harhauttamiseen ja luottamuksen hyväksikäyttämiseen [5], [16]. Lisäksi aineistossa käsitellään, kuinka sosiaaliseen mediaan jaetaan herkästi ajankohtaista sijaintitietoa, jolloin jakaja voi houkutella varkaita asunnolleen, kun tiedetään hänen olevan muualla kuin kotonaan [30].

5 Pohdinta

Tässä tutkielmassa tarkasteltiin syväväärennösteknologian luomia tietoturvauhkia sosiaalisen median alustojen käyttäjillä. Lisäksi tarkasteltiin syväväärennösten tunnistamista sekä niiltä puolustautumista. Aineistossa tuodaan ilmi, että sosiaalinen media on erinomainen alusta jakaa synteettistä mediaa suurille ihmismäärille. Tilanteesta ongelmallisen luo sosiaalisen median alustojen haluttomuus omaksua syväväärennöksien ja synteettisen median tunnistamiseen liittyviä autentikointitekniikoita, mikä kasvattaa syväväärennösten luomia tietoturvauhan riskiä.

Tutkielmassa käy ilmi, että syväväärennösten ensimmäisiä käyttötarkoituksia on ollut ei-suostumuksellisen pornografian tekeminen julkisuudenhenkilöistä. Syväväärennökset ovat saaneet nopeasti uusia käyttötarkoituksia tämän myötä, kuten esimerkiksi valheellisen median, misinformaation sekä huijausten muodossa.

Kirjallisuuskatsauksen perusteella yleisimmät sosiaalisen median alustat ovat keskeinen tekijä syväväärennösten vahingolliseen käyttöön. Sosiaalisen median alustat ovat lisäksi erinomainen paikka jakaa valheellisia uutisia sekä misinformaatiota, sillä alustotoilla tiedon autentikointi on heikkoa sekä tiedon leviäminen voi olla hyvinkin nopeaa. Tarkasteltu kirjallisuus [25], [27] viittaa siihen, että keskimäärin nuoret, lapset sekä yli 65 vuotiaat ihmiset ovat alttiimpia uhille sosiaalisessa mediassa. Tästä voi muodostua ongelma, kun heidän sosiaalisen median käyttäjilleen syötetään valheellisen median sekä misinformaation sisältöä, mikä voi vaikuttaa esimerkiksi heidän äänestämiskäyttäytymiseene. Syväväärennöksiä generoivia tekoäly-

malleja voidaan kouluttaa sosiaalisesta mediasta haravoitujen mediasisältöjen avulla. Haravoitu mediasisältö voi olla sattumanvaraisesti valittua julkisista profiileista, mutta sitä voidaan myös kohdistaa yksityisille profiileille, mikäli datan kerääjälle on annettu pääsy käyttäjän lataamille julkaisuille. Erityisen suuri uhka muodostuu alaikäisten sosiaalisen median profiileille, koska heidän julkaisemiaan mediasisältöjä voidaan käyttää koulutusmateriaalina ilman heidän suostumustaan.

Kirjallisuuden perusteella syväväärennöksiä hyödyntämällä toteutetut sosiaalisen manipuloinnin hyökkäykset ovat keskeinen tietoturvauhka sosiaalisen mediassa. Syväväärennöksillä toteutetuissa sosiaalisen manipulaation hyökkäyksissä rikollinen hyödyntää uhrista tai uhrin tuntemasta henkilöstä julkaistua dataa, jolloin hyökkääjä saa tehtyä uskottavan syväväärennöksen manipulaatiotaan varten.

Aineistossa todetaan, että syväväärennöksien tunnistamiseen on kehitetty menetelmiä, mutta ne eivät pysy generoinnissa hyödynnettävien GAN-verkkojen sekä autoenkoodereiden kehityksen perässä. Tämä kasvattaa kuilua syväväärennösten generoinnin sekä tunnistamisen välillä, mikä kasvattaa syväväärennösten käyttötarkoitusten onnistumismahdollisuuksia. Tehokkaiden koneellisten tunnistusmenetelmien puutteessa sosiaalisen median käyttäjiä tulisi perehdyttää syväväärennösten käyttötarkoituksista sekä niiden luomista tietoturvauhista. Kirjallisuuden perusteella opettamalla yleisiä puolustusmenetelmiä voidaan ehkäistä syväväärennösten käyttötarkoitusten toimivuutta. Tutkimusaineiston mukaan käyttäjien tietoisuuden kasvattaminen ja alustojen sääntöjen muuttaminen läpinäkyvämmän sisällön suuntaan ovat keskeisiä menetelmiä syväväärennöksiltä puolustautumiseen.

6 Yhteenveto

Tutkielmassa tarkasteltiin syväväärennösten luomia tietoturvauhkia sosiaalisen median alustoilla. Tavoitteena oli selvittää, miten syväväärennösteknologian kehittyminen näkyy sosiaalisessa mediassa sekä millaisia uhkia se luo käyttäjille. Syväväärennösten yleistyminen korostaa erityisesti medialukutaidon merkitystä, koska synteettisen median haitallisia vaikutuksia voidaan vähentää. Tämän vuoksi tutkielmassa käsiteltiin myös syväväärennösten tunnistamiseen ja niiltä puolustautumiseen liittyviä menetelmiä.

(Tk1) Tutkielmassa havaittiin useita eri käyttötarkoituksia syväväärennöksille, joiden vaikutukset korostuvat erityisesti sosiaalisen median alustoilla. Syväväärennöksiä hyödyntäen voidaan toteuttaa sosiaalisen manipuloinnin hyökkäyksiä sosiaalisen median käyttäjiin, jolloin käyttäjältä yritetään huijata rahaa tai kalastella tietoa. Sosiaalisen manipulaation lisäksi sosiaalisessa mediassa voidaan levittää valheellisia uutisia sekä misinformaatiota. Yleisin käyttötarkoitus syväväärennöksille on identiteetin varastaminen, jolloin hyökkääjä yrittää luoda mainehaittaa tai kiusata uhria tekemällä hänestä synteettistä mediaa ja levittämällä sitä eri sosiaalisen median alustoilla.

(Tk2) Syväväärennösten tunnistaminen on keskeinen tutkimusaihe autenttisen sosiaalisen median sisällön kannalta. Syväväärennöksiä voidaan tunnistaa erilaisten tunnistusalgoritmeihin sekä omaan arvioon perustuen. Tunnistusalgoritmit tarvitsevat tunnistamiseen vertailukelpoista dataa, jonka saaminen kaikista sosiaalisen me-

dian käyttäjistä on hyvin haastavaa. Tämä luo myös tietoturvahaasteen tunnistusdatan luotettavan säilönnän kannalta, sillä data voi sisältää biometrisiä tunnisteita, mikä luo riskin datan väärinkäytölle.

(Tk3) Syväväärennöksiltä voidaan puolustautua monilla erilaisilla tavoilla. Syväväärennöksiltä puolustautumisessa korostuu ennaltaehkäisevät menetelmät. Keskeinen menetelmä on sosiaalisen median käyttäjien perehdyttäminen syväväärennöksiin sekä niiden aiheuttamiin uhkiin. Lisäksi keskeiseksi puolustusmenetelmäksi on harkittu kuvien autentikointia kolmannen osapuolen toimesta, jolloin kuvan aitous voidaan todentaa esimerkiksi virtuaalisella vesileimalla.

Tulevaisuudessa syväväärennösteknologian arvioidaan kehittyvän nopeasti nykyistä realistisemmaksi. Tämä luo haasteen tunnistusalgoritmien kehitykselle, joiden kehitys ei ole nykyään yhtä nopeaa kuin generoinnin. Tämän vuoksi sosiaalisen median käyttäjille tulisi esittää puolustuskeinoja syväväärennösten riskien varalta.

Tämän tutkielman perusteella voidaan ehdottaa jatkotutkimuskohteita. Jatkotutkimusta voidaan tehdä syväväärennösten tunnistusalgoritmien tehokkuudesta eri sosiaalisen median alustoilla. Toinen tutkimuskohde on syväväärennoksiltä puolustautumiseen käytettyjen menetelmien tarkempi määrittely ja kartoitus. Tutkimalla kyseisiä aiheita voidaan pyrkiä luomaan tietoturvallisempi ympäristö sosiaalisen median käyttäjille.

Lähdeluettelo

[1] M. Saha, P. K. Pagadala ja M. T. Basu, ”Cyber Deceptions: Unveiling the Threat of Deepfakes in Digital Realms”, teoksessa *2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, lokakuu 2024, s. 113–116. DOI: 10.1109/ICCCMLA63077.2024.10871909.

[2] A. A. AlAmeeri ja M. B. AlMourad, ”Impact of Social Engineering on Social Media Users”, teoksessa *2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, huhtikuu 2025, s. 1–5. DOI: 10.1109/ITIKD63574.2025.11004806.

[3] H. Taherdoost, ”Cybersecurity vs. Information Security”, *Procedia Computer Science*, vol. 215, s. 483–487, 2022. DOI: 10.1016/j.procs.2022.12.050.

[4] D. B. Frolov, D. D. Makhaev ja V. V. Shishkarev, ”Deepfakes and Information Security Issues”, teoksessa *2022 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, syyskuu 2022, s. 147–150. DOI: 10.1109/ITQMIS56172.2022.9976507.

[5] K. Afroz, S. Chittam ja K. Roy, ”Understanding the Threat of Political Deepfakes”, teoksessa *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*, helmikuu 2025, s. 1–6. DOI: 10.1109/ICAIC63015.2025.10848832.

[6] C. Yavuz, "A Multidisciplinary Look at History and Future of Deepfake With Gartner Hype Cycle", *IEEE Security & Privacy*, vol. 22, nro 3, s. 50–61, toukokuu 2024. DOI: 10.1109/MSEC.2024.3380324.

[7] Z. Han, "The Application of Artificial Intelligence in Computer Network Technology", teoksessa *2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, lokakuu 2021, s. 632–635. DOI: 10.1109/AINIT54228.2021.00127.

[8] S. Patel, "An Overview and Application of Deep Convolutional Neural Networks for Medical Image Segmentation", teoksessa *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, helmikuu 2023, s. 722–728. DOI: 10.1109/ICAIS56108.2023.10073857.

[9] G. Kasera, M. Solanki, H. Kaur ja K. Shah, "A Detailed Exploration to Deepfake: A Cybersecurity Threat", teoksessa *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, tammikuu 2025, s. 1–6. DOI: 10.1109/SCEECS64059.2025.10940812.

[10] Y. Mirsky ja W. Lee, "The Creation and Detection of Deepfakes: A Survey", *ACM Comput. Surv.*, vol. 54, nro 1, 7:1–7:41, tammikuu 2021. DOI: 10.1145/3425780.

[11] M. Zubair ja S. Hakak, "Exploring the Landscape of Compressed DeepFakes: Generation, Dataset and Detection", *Neurocomputing*, vol. 619, s. 129 116, 2025. DOI: 10.1016/j.neucom.2024.129116.

[12] V. Chawla ja T. Baraskar, "Comparative Analysis of DeepFake Detection Methods", teoksessa *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, elokuu 2024, s. 1–8. DOI: 10.1109/ASIANCON62057.2024.10838081.

[13] A. Badman ja M. Kosinski. ”What is a dataset”. Toukokuu 2026. url: https://www.ibm.com/think/topics/dataset.

[14] R. Lakhchaura, P. Madan ja A. Vishnoi, ”A Review of Deepfake Technology: Advancements in Detection and Generation Methods, Datasets and Tools”, teoksessa *2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, marraskuu 2025, s. 1–5. DOI: 10.1109/ICCCA66364.2025.11325599.

[15] N. Negi, D. Kaushal, S. Bahl ja A. Mer, ”Deepfake Technology: An In-depth Review of the Emerging Threats Posed by Artificial Intelligence to Society”, teoksessa *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, joulukuu 2024, s. 270–274. DOI: 10.1109/EmergIN63207.2024.10961120.

[16] W. Nowakowski, ”Social Engineering Analysis Framework: A Comprehensive Playbook for Human Hacking”, *IEEE Access*, vol. 13, s. 18 827–18 849, 2025. DOI: 10.1109/ACCESS.2025.3532999.

[17] S. Singh ja A. Dhumane, ”Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges”, *MethodsX*, vol. 15, s. 103 632, joulukuu 2025. DOI: 10.1016/j.mex.2025.103632.

[18] M. Wazid, A. K. Mishra, N. Mohd ja A. K. Das, ”A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact”, *Cyber Security and Applications*, vol. 2, s. 100 040, tammikuu 2024. DOI: 10.1016/j.csa.2024.100040.

[19] M. G. Yigezu, M. A. Mehamed, O. Kolesnikova, T. K. Guge, A. Gelbukh ja G. Sidorov, ”Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media”, teoksessa *2023 International Conference on Informa-*

tion and Communication Technology for Development for Africa (ICT4DA), lokakuu 2023, s. 171–175. DOI: 10.1109/ICT4DA59526.2023.10302243.

[20] European Institute of Peace. "Fake news misinformation and hate speech in Ethiopia: A vulnerability assessment". Toukokuu 2026. url: https://www.eip.org/wp-content/uploads/2021/04/Fake-News-Misinformation-and-Hate-Speech-in-Ethiopia.pdf.

[21] A. Ali, K. F. Khan Ghouri, H. Naseem, T. R. Soomro, W. Mansoor ja A. M. Momani, "Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security", teoksessa *2022 International Conference on Cyber Resilience (ICCR)*, lokakuu 2022, s. 1–5. DOI: 10.1109/ICCR56254.2022.9995821.

[22] S. Ji, "#MeToo in an AI-generated Deepfake Sexual Violence Era in South Korea", *Women's Studies International Forum*, vol. 112, s. 103 146, 2025. DOI: 10.1016/j.wsif.2025.103146.

[23] A. Ayub Khan et al., "Digital Forensics for the Socio-Cyber World (DF-SCW): A Novel Framework for Deepfake Multimedia Investigation on Social Media Platforms", *Egyptian Informatics Journal*, vol. 27, s. 100 502, 2024. DOI: 10.1016/j.eij.2024.100502.

[24] A. A. Khan et al., "Cybersecurity, Digital Forensics, and the IoT for Deepfake Investigation on Social Media Platforms: A Review", *Human-Centric Computing and Information Sciences*, vol. 15, s. 38, heinäkuu 2025. DOI: 10.22967/HCIS.2025.15.038.

[25] S. A. Laczi ja V. Póser, "Impact of Deepfake Technology on Children: Risks and Consequences", teoksessa *2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, syyskuu 2024, s. 215–220. DOI: 10.1109/SISY62279.2024.10737593.

[26] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman ja Y. K. Dwivedi, "Deepfakes: Deceptions, Mitigations, and Opportunities", *Journal of Business Research*, vol. 154, s. 113 368, 2023. DOI: 10.1016/j.jbusres.2022.113368.

[27] V. Narayanan, B. W. Robertson, A. Hickerson, B. Srivastava ja B. W. Smith, "Securing Social Media for Seniors from Information Attacks: Modeling, Detecting, Intervening, and Communicating Risks", teoksessa *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, joulukuu 2021, s. 297–302. DOI: 10.1109/TPSISA52974.2021.00053.

[28] E. Preu, M. Jackson ja N. Choudhury, "Perception vs. Reality: Understanding and Evaluating the Impact of Synthetic Image Deepfakes over College Students", teoksessa *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, lokakuu 2022, s. 0547–0553. DOI: 10.1109/UEMCON54665.2022.9965697.

[29] T. Saheb, M. Sidaoui ja B. Schmarzo, "Convergence of Artificial Intelligence with Social Media: A Bibliometric & Qualitative Analysis", *Telematics and Informatics Reports*, vol. 14, s. 100 146, 2024. DOI: 10.1016/j.teler.2024.100146.

[30] B. Thuraisingham, "The Role of Artificial Intelligence and Cyber Security for Social Media", teoksessa *2020 Ieee 34th International Parallel and Distributed Processing Symposium Workshops (Ipdpsw 2020)*, Los Alamitos: IEEE Computer Soc, 2020, s. 1116–1118, ISBN: 978-1-7281-7445-7. DOI: 10.1109/IPDPSW50202.2020.00184.