Tampereen yliopisto

Ville Kivioja

MENETELMÄT SYVÄVÄÄRENNÖS-KUVIEN TUNNISTAMISEEN SOSIAALI-SESSA MEDIASSA

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Joulukuu 2025

TIIVISTELMÄ

Ville Kivioja: Menetelmät syväväärennöskuvien tunnistamiseen sosiaalisessa mediassa
Kandidaattitutkielma
Tampereen yliopisto
Tietojenkäsittelytieteiden tutkinto-ohjelma
Joulukuu 2025

Tekoälyteknologioiden valtavirtaistumisen myötä 2020-luvulla syväväärennössisältöjen määrä verkossa on kasvanut räjähdysmäisesti. Syväväärennöksillä (eng. deepfake) tarkoitetaan syväneuroverkoilla tuotettuja valheellisia sisältöjä. Syväväärennösten avulla on mahdollista vaikuttaa voimakkaasti luottamukseen verkon sisältöjä kohtaan, kohdentaa hyökkäyksiä henkilöitä kohtaan, aiheuttaa suuria taloudellisia menetyksiä ja rapauttaa luottamusta demokraattisiin prosesseihin.

Tässä tutkielmassa tarkastellaan syväväärennöskuvien tuottamis- ja tunnistusmenetelmiä sosiaalisessa mediassa. Tutkimuksessa relevanteimmiksi arkkitehtuurimalleiksi syväväärennöskuvien tuottamiseen tunnistettiin GAN-, VAE- ja diffuusiomallit ja syväväärennöskuvien yleisimmin käyttämiksi tekniikoiksi kasvojen vaihtaminen, kasvojen uudelleenesitys, kokonaan synteettisten kasvojen luonti ja kasvojen ominaisuuksien muokkaaminen. Tunnistuskeinoista keskeisimmiksi nousivat tunnistaminen koneoppimisteknologioilla, tunnistaminen fysikaalisten ja fysiologisten epäjohdonmukaisuuksien avulla ja tunnistaminen digitaalisten tunnisteiden avulla. Lisäksi tutkielmassa tunnistettiin tärkeäksi tietoisuuden lisääminen tekoälyteknologioista ja syväväärennöksistä. Tulosten perusteella mikään keino ei yksinään riitä vastaamaan syväväärennöskuvien alati kasvaviin vaikutuksiin, vaan ratkaisevaa tulee olemaan eri keinojen yhdistely.

Avainsanat: syväväärennökset, syväneuroverkot, kuvat, tunnistaminen, sosiaalinen media

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

TEKOÄLYN KÄYTTÖ OPINNÄYTTEESSÄ

Opinnäytteessäni on käytetty tekoälysovelluksia:

☐ Ei
☒ Kyllä

Ilmoitukseni mukaan olen käyttänyt opinnäytteessäni tutkielmaprosessin aikana seuraavia tekoälysovelluksia: ChatGPT, Google Gemini

Tekoälysovellusten nimet ja versiot: GPT-5, 2.5 Flash

Käyttötarkoitus: Käytin ChatGPT -sovellusta tutkielman kielenhuollossa ja apuna termien kääntämisessä englannista suomeksi. Termit tarkastettiin tekemällä hakuja saaduilla termeillä, varmistuen näin niiden asianmukaisuudesta. Käytin Google Gemini -sovellusta luodakseni esimerkkikuvan kokonaan synteettisen henkilökuvan luonnista.

Osiot, joissa tekoälyä on käytetty: Tekoälyä hyödynnettiin koko tutkielmassa, koko kirjoittamisprosessin aikana.

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien osat, joissa on hyödynnetty tekoälyä, ja hyväksyn vastuun mahdollisista eettisten ohjeiden rikkomuksista.

SISÄLLYSLUETTELO

1 Johdanto 1

2 Tutkimusmenetelmä 3

3 Syväväärennöskuvien tuottamisessa hyödynnettävät arkkitehtuurimallit ja tekniikat 4

3.1 GAN-malli 4

3.2 VAE-malli 5

3.3 Diffuusio-malli 5

3.4 Syväväärennöskuvien hyödyntämät tekniikat 5

4 Menetelmät syväväärennöskuvien tunnistamiseen 8

4.1 Tunnistaminen koneoppimisteknologioilla 8

4.2 Tunnistaminen fysikaalisten ja fysiologisien epäjohdonmukaisuuksien avulla 10

4.3 Digitaaliset tunnisteet 10

4.4 Tietoisuuden lisääminen tekoälyteknologioista ja syväväärennöksistä 11

5 Keskustelu 13

6 Yhteenveto 15

Lähdeluettelo 16

1 Johdanto

Viime vuosina tekoälyteknologiat ovat nousseet yleiseen tietoisuuteen ja niiden käyttö on kasvanut merkittävästi. Osasyinä käytön kasvuun on viime vuosina tapahtunut nopea kehitys tuotettujen kuvien realistisuudessa, teknologioiden yleistymisessä ja helppokäyttöisyydessä (Lyu 2024; Jheelan & Pudaruth 2025). Kuvia luovia tekoälyteknologioita pystytään hyödyntämään esimerkiksi videoiden suoratoistopalveluiden optimoinnissa, kustannusten alentamisessa viihde- ja mainosteollisuuksissa ja erilaisissa kuntoutusmenetelmissä lääketieteessä (Lyu 2024). Samaan aikaan syväoppimisteknologiaa hyödynnetään yhä enenevissä määrin haitallisen sisällön levittämisessä.

Syväväärennökset (eng. deepfake) tarkoittavat syväneuroverkoilla (eng. deep neural network, DNN) tuotettua synteettistä multimediasisältöä, jolla pyritään välittämään harhaanjohtavaa tai virheellistä informaatiota (Lyu 2024; Negi ja muut 2024). Syväväärennöksiksi tunnistetuissa sisällöissä on tapahtunut eksponentiaalista kasvua. Joulukuun 2018 ja kesäkuun 2020 välillä syväväärennössisältöjen määrä oli noussut 7964 tapauksesta 49081 tapaukseen (Pandey ja muut 2024). Vuonna 2023 syväväärennössisältöjä oli löydetty verkosta noin 500 000 kappaletta ja vuonna 2025 määrän ennustetaan nousevan noin kahdeksaan miljoonaan (Negreiro 2025).

Syväväärennössisällöillä on mahdollista vaikuttaa merkittävästi verkon informaatioympäristöön (Kharvi 2024; Lyu 2024). Tekoäly- ja syväväärennössisältö sosiaalisessa mediassa luovat haasteita medialukutaidolle (Kharvi 2024; Lyu 2024), faktojen tarkastamiselle (Kharvi 2024), demokratian prosessien luotettavuudelle (Kharvi 2024; Lyu 2024) ja yksityisyydensuojalle (Kharvi 2024). Syväväärennöskuvilla ja -videoilla voidaan luoda valheellisia representaatioita yksilöistä, aiheuttaen kohteelle mainehaittaa ja psykologista kärsimystä (Lyu 2024; Negi ja muut 2024). Yksilön itsensä lisäksi nämä valheelliset representaatiot aiheuttavat vahinkoa kognitiiviselle turvallisuudelle: syväväärennösvideot ja -kuvat luovat tilanteen, jossa ihmiset ovat kykenemättömiä erottamaan todellista informaatiota valheellisesta. Syväväärennöskuvilla on mahdollista vaikuttaa suoraan ihmisten käyttäytymiseen. Luomalla syväväärennöskuvia käsillä olevasta kriisistä pystytään aiheuttamaan pörssiosakkeiden kurssiromahduksia tai väärennetyillä sosiaalisen median kuvilla identiteettivarkauksia. Valheellisen sisällön ohella korkeatasoiset syväväärennökset mahdollistavat sen, että myös täysin aidot kuvat ja videot voidaan kyseenalaistaa väärennöksinä. (Lyu 2024.)

Koska uhka syväväärennöskuvien vaikutuksille kasvussa, pidän aiheen tutkimista tärkeänä. Tutkielmassani tarkastelen erityisesti syväväärennöskuvien tuottamismenetelmiä ja mahdollisia keinoja ilmiön torjumiseksi sosiaalisessa mediassa.

Tutkimuskysymyksiäni ovat:

1. Millä menetelmillä syväväärennöskuvia tuotetaan sosiaaliseen mediaan?
2. Millä menetelmillä syväväärennöskuvia voidaan tunnistaa?

Johtuen kandidaatintutkielman rajallisesta laajuudesta, olen pyrkinyt rajaamaan tarkastelun kaikkein keskeisimpiin teknologioihin syväväärennöskuvien tuottamisessa ja tunnistamisessa. Tutkielmassa käyn mallit läpi yleisluontoisesti ja ilman esimerkiksi niiden toimintaan liittyviä matemaattisia kaavoja.

2 Tutkimusmenetelmä

Toteutin tutkimuksen kirjallisuuskatsauksena vuoden 2025 syyskuun ja marraskuun välisenä aikana. Käytin tiedonhaussa tietokantoja ACM Digital library, Proquest computer science database ja Andor. Rajasin haun aikavälin käsittämään 2020 jälkeen ilmestyneet julkaisut. Perustelen aikarajausta teknologioissa viime vuosina tapahtuneella nopealla kehityksellä.

Muodostin hakulausekkeet yhdistelmällä seuraavia:

1. **Syväväärennökset:** käytin hakutermeinä ”deep fake” ja ”deepfake” sekä näiden permutaatioita asteriskilla (*) ottaen huomioon tapaukset, joissa termi esiintyi osana jotain muuta sanaa, esimerkiksi ”deepfake-image” ja monikkomuodot.
2. **Sosiaalinen media:** käytin kaikissa hakulausekkeissa vakiintunutta muotoa ”social media”.
3. **Kuvat:** käytin hakutermeinä ”image”, ”photo” ja ”picture”, sekä näiden permutaatioita asteriskin kanssa.
4. **Uhat:** käytin hakutermeinä ”threat”, ”risk”, ”harm” ja ”impact”, sekä näiden monikkomuotoja hyödyntäen asteriskia.

Rajasin hakutulokset englanninkielisiin. Asetin ehdoksi artikkelin valitsemiselle tutkimusaineistoon sen, että kustakin kategoriasta täytyy esiintyä jokin hakutermi artikkelin tiivistelmässä. Tämän jälkeen seuloin hakutulokset manuaalisesti lukemalla artikkelien tiivistelmät. Otin julkaisun mukaan, jos se käsitteli syväväärennöskuvien vaikutuksia ja/tai tunnistamismenetelmiä. Valitsin mukaan aineistoon julkaisut, jotka käsittelivät syväväärennöksiä sekä kuvien että videoiden näkökulmasta, kun taas pelkästään videoita käsittelevät artikkelit hylkäsin. Edellä kerrotun lisäksi syväneuroverkkojen toimintaa taustoittamaan etsin tietoa Andorista hakulausekkeella ”Deep neural network image generation”. Lopuksi mukaan otettiin artikkeleita, joiden tuloksia oli käsitelty muissa tutkimuksissa, jotta voitiin tehdä primääriviittauksia alkuperäisiin lähteisiin.

3 Syvävääärennöskuvien tuottamisessa hyödynnettävät arkkitehtuurimallit ja tekniikat

Valokuvien väärentämisessä ei sinällään ole mitään uutta ja sitä on tapahtunut aina valokuvauksen keksimisestä saakka. Digitaaliset kuvankäsittelyohjelmistot ovat helpottaneet käsittelyä entisestään ja madaltaneet taloudellista ja taidollista kynnystä kuvien manipuloinnille. Syvävääärennösteknologioiden poikkeuksellisuus verrattuna aikaisempiin tekniikoihin perustuu niiden tarkkuuteen jäljitellä haluttuja ominaisuuksia. (Masood ja muut 2022.)

Syvävääärennöskuvien tuottaminen perustuu syväneuroverkkojen hyödyntämiselle (Pandey 2024). Syväneuroverkot perustuvat ajatukseen jäljentää ihmisaivojen toimintaa kouluttamalla malli tuottamaan toivotunlaisia tuloksia annetuilla syötteillä. Näissä malleissa aivojen neuronien biokemialliset prosessit korvataan matemaattisin funktioin. Neuroverkot koostuvat syöte- ja ulostulokerroksesta sekä kerroksittaisista neuroneista. Syöte annetaan neuroverkolle vektorimuotoisena, jonka jälkeen kussakin kerroksessa olevat neuronit laskevat vektorissa oleville arvoille uudet arvot funktioidensa mukaisesti ja lähettävät sen eteenpäin seuraavaan kerrokseen, kunnes lopulta ulostulokerroksessa annetaan tulos laskutoimituksista. (Drori 2024, s.4–9.) Syvävääärennöskuvien tuottamiseen köytettävät syväneuroverkot koostuvat tyypillisesti kymmenestä yli sataan tasosta (Pandey ja muut 2024). Kuvien tapauksessa ominaisuuksia, johon laskettavan muodon arvot perustuvat, käytetään ilmaisua piilevät piirteet (eng. latent features) (Masood ja muut 2022).

Seuraavissa aliluvuissa on esitelty keskeisimmät arkkitehtuurimallit syvävääärennöskuvien tuottamiseen ja tekniikat, joita tavanomaisimmin hyödynnetään syvävääärennöskuvissa.

3.1 GAN-malli

GAN-malli (eng. generative adversarial network model) koostuu kahdesta erillisestä syväneuroverkosta. Nämä ovat generaattori (eng. generator) ja diskriminaattori (eng. discriminator). Generaattori luo kuvia kasvoista sattumanvaraisesta kohinasta (eng. noise), kun taas diskriminaattori pyrkii jaottelemaan sille syötettyjä kuvia väärennöksiin ja aitoihin kuviin ja antaa tästä palautetta generaattorille. Generaattori pyrkii prosessissa parantamaan luomiensa kuvien laatua saadun palautteen pohjalta, ja diskriminaattori pyrkii tunnistamaan kuvia yhä tarkemmin. Lopulta generaattorin luoma väärennetty kuva vastaa

piirteiltään niin hyvin todellista kuvaa, että diskriminaattori ei pysty erottamaan sitä aidosta. (Pandey ja muut 2024; Lyu 2024; Jheelan & Pudaruth 2025.)

3.2 VAE-malli

VAE-malli (eng. variational auto-encoder model) koostuu kahdesta erillisestä syväneuroverkosta, enkooderista (eng. encoder) ja dekooderista (eng. decoder) (Lyu 2024.). Enkooderin tehtävänä on muuntaa kuva laskettavaan muotoon sen piirteisiin perustuen, ja dekooderin tehtävänä on puolestaan palauttaa kuva perustuen tähän muotoon (Masood ja muut 2022). Kumpikin hyödyntää koulutusdatanaan kuvia, joista osa on kuvia henkilöstä, josta halutaan tehdä syvävääreennös, ja osa henkilöstä (luovuttaja, eng. donor), jonka piirteitä halutaan sijoittaa kohdehenkilön kuvaan. Kumpikin näistä hyödyntää samaa enkooderia, mutta eri dekooderia. Tämän lisäksi muodostetaan kaksi muuta enkooderi-dekooderi paria. Tämän jälkeen syväverkkoja pyritään säätämään minimoimalla erot dekonstruoitujen kuvien ja syötekuvien piirteiden välillä, kunnes nämä erot ovat mahdollisimman pieniä. (Lyu 2024.)

3.3 Diffuusio-malli

Diffuusiomallissa (eng. diffusion model) syötekuvaan lisätään satunnaista kohinaa, jonka poistamiseen käytetään kohinanpoistamiseen koulutettua syväneuroverkkoa (eng. denoiser). Prosessia jatketaan iteratiivisesti, kunnes kohinanpoistin ei kykene enää tuottamaan satunnaisesta kohinasta eroavaa lopputulosta. Tämän jälkeen kohinanpoistin jatkaa kohinan poistamista ilman että uutta lisätään, kunnes uusi kuva on saatu luotua. (Lyu 2024.)

3.4 Syvävääreennöskuvien hyödyntämät tekniikat

Tekniikoita, joita syvävääreennöskuvat yleisimmin hyödyntävät, ovat:

- Kasvojen vaihtaminen (eng. face swapping): Kohteen kasvojen piirteet siirretään toiselle henkilölle (Masood ja muut 2022).

Kuva 1. Esimerkki kasvojen vaihtamista hyödyntävästä syväväärennöskuvasta (Masood ja muut 2022)

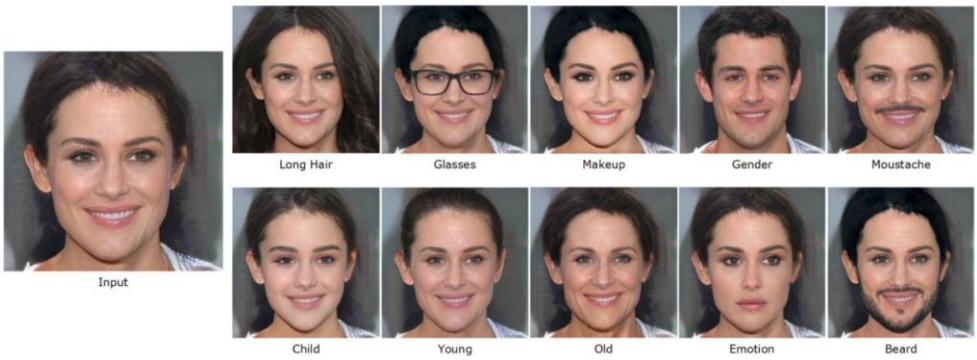
- Kasvojen uudelleenesitys (eng. face reenactment, tunnetaan myös termillä nukkemestari, eng. puppet master): Kohteelle siirretään toisen henkilön ilmeitä (Masood ja muut 2022).

Kuva 2. Esimerkki kasvojen uudelleenesitystä hyödyntävästä syväväärennöskuvasta (Masood ja muut 2022).

- koko kasvojen synteesi (eng. entire face synthesis): Luodaan täysin keinotekoiset kasvot, jotka eivät perustu kenenkään tietyn henkilön piirteisiin (Masood ja muut 2022).

Kuva 3. Esimerkkikuva koko kasvosynteesistä. Kuva luotu Googlen Gemini 2.5 Flash-mallilla kehotteella ”Olet johtavan PR-firman ammattivalokuvaaja. Luo profiilikuva kuvitteelliselle toimitusjohtajalle sosiaaliseen mediaan. Henkilöllä on itsevarma hymy. Hän on pukeutunut tummansiniseen pukuun. Tausta on toimisto, mutta sumennettu.”

- kasvojen ominaisuuksien muokkaus (eng. facial attribute manipulation): Muutetaan tiettyjä ominaisuuksia kasvoissa. Tällaisia muutoksia voisivat olla esimerkiksi ihon silottaminen, ihon värin tai sukupuolen vaihto tai esimerkiksi silmälasien poistaminen tai lisääminen kohteelle (Masood ja muut 2022).



Kuva 4. Esimerkki kasvojen ominaisuuksien muokkausta hyödyntävistä syvävääreennöskuvista (Masood ja muut 2022).

4 Menetelmät syväväärennöskuvien tunnistamiseen

Koska syväväärennösteknologiat kehittyvät nopeasti ja syväväärennöskuvat ovat piirteiltään hyvin moninaisia, luottaminen pelkästään tunnistusteknologioihin ei riitä. Seuraavassa on kerrottu tutkimuskirjallisuudessa havaituista tavoista tunnistaa, varmentaa kuvien aitous tunnistein ja lisätä tietämystä syväväärennösteknologioista niiden haittojen torjumiseksi.

4.1 Tunnistaminen koneoppimisteknologioilla

Samaan tapaan kuin syväväärennöskuvien tuottamisteknologioissa pyritään enkoodaamaan kuvan piirteitä neuroverkkoon ja sitten tuottamaan uusi kuva näiden piirteiden kautta, koneoppimisteknologioilla kuvien sisältämiä piirteitä pystytään hyödyntämään kyseisten väärennösten tunnistamisessa. Periaatteena on tällöin kouluttaa malli syväväärennöskuvilla ja aidoilla kuvilla havaitsemaan sellaisia piirteitä, jotka esiintyvät syväväärennöskuvissa mutta eivät yleensä aidoissa kuvissa. Kirjallisuuskatsauksen perusteella relevanteimmiksi syväväärennöskuvien tunnistukseen soveltuviksi koneoppimisteknologioiksi nousivat syväneuroverkoista konvoluutioneuroverkko ja muista koneoppimisteknologioista tukivektorikone ja käyn läpi niiden piirteet yleisluontoisesti. Muita tarkastellussa kirjallisuudessa mainittuja koneoppimismalleja olivat logistinen regressio (eng. logistic regression), päätöspuu (eng. decision tree), satunnaismetsä (eng.random forest), XGBoost (eng. extreme gradient boosting) ja naiivi Bayesin luokitin (eng. naive Bayes).

Konvoluutioneuroverkko (eng. convolutional neural network, CNN) on syväneuroverkko, joka koostuu konvoluutio-, yhdistämis- ja täysliitäntäkerroksista. Konvoluutioneuroverkko pyrkii ensin muuttamaan yksinkertaistettuun 1-ulotteiseen muotoon säilyttäen mahdollisuuksien mukaan kaikki kuvan luokitteluun liittyvät ominaisuudet. Konvoluutiokerros käy kuvan läpi pikselialueiden avulla, usein kooltaan 3 x 3, ja laskee kustakin skalaaritulon. Kun kuva on käyty läpi, näistä skalaarituloista muodostetaan piirreavaruus. Yhdistämiskerroksen tehtävänä on tiivistää piirreavaruus mahdollisimman tiiviiseen muotoon mahdollisimman laskennan optimoimiseksi. Täysliitäntäkerros on mallissa neuroverkko, joka toteuttaa kuvien luokittelun tuotetun piirreavaruuden avulla. (Mallet ja muut 2023.) Koska neuroverkkojen toimivuus syväväärennöskuvien tunnistuksessa on riippuvainen koulutusdatan laadusta, koulutusdatan laadun heikentyminen tarkoittaa myös entistä suurempaa todennäköisyyttä virhetunnistuksille. Vastustavilla hyökkäyksillä (eng. adversarial attacks) tarkoitetaan sellaisten syötteiden saattamista jotain kautta osaksi neuroverkon koulutusdataa, joko kohdistetusti tai kohdistamattomasti, jotka

aiheuttavat merkittävää väärinluokittelua datassa, mutta joita ihminen ei pysty erottamaan tavanomaisista syötteistä (Huang & Wu 2024).

Tukivektorikone (eng. support vector machine, SVM) on malli, jossa annetaan ulottuvuus (eng. dimension) kullekin kuvasta tarkasteltavalle ominaisuudelle. Kuvasta enkoodatut arvot sijoitetaan näihin ulottuvuuksiin ja arvot jaetaan joukkoihin hypertasojen (eng. hyper-level) avulla siten, että joukkojen tukivektorien ja hypertason väliin jää mahdollisimman suuri etäisyys. Nämä asetetut hypertasot toimivat siten erottimina aitojen ja väärennettyjen kuvien välillä. Malletin ja kumppaneiden tutkimuksessa konvoluutiotioneuroverkkoa hyödyntävä malli kykeni parempaan GAN-mallilla tuotettujen kasvokuvien tunnistamiseen saavuttaen 88,33 prosentin tunnistustarkkuuden verrattuna tukivektorikoneen saavuttamaan 81,69 prosenttiin. (Mallet ja muut 2023.)

Syvävääärennöskuvien tunnistamista sosiaalisessa mediassa vaikeuttavat tavat, joilla palvelut poistavat kuvien metatietoja sekä pakkaavat ja pienentävät niiden tarkkuutta. Tämä voi sekä häivyttää merkkejä mahdollisesta manipuloinnista että aiheuttaa vääriä positiivisia tunnistuksia, mikäli tunnistusmallia ei olla säädetty huomioimaan tätä ilmiötä. (Masood 2022.) Pandey ja kumppanit (2024) kehittivät perinteisiä koneoppimisalgoritmeja hyödyntävän menetelmän eri tasoisten, matalaresoluutioisten ja korkearesoluutioisten syvävääärennösten tunnistamiseen. Heidän menetelmässään eriresoluutioisista aidoista ja syvävääärennetyistä kuvista koostuva koulutusdata jaetaan ensin matala- ja korkearesoluutioiseen dataan. Tämän jälkeen luodaan kaksi heikosti oppivaa (eng. weak learner) syväneuroverkkoa kouluttamalla toinen matalaresoluutioisella ja toinen korkearesoluutioisella koulutusdatalla. Lopuksi kummatkin mallit yhdistetään yhdeksi malliksi. He testasivat tutkimuksessaan osana mallin toimintaa seuraavia koneoppimisalgoritmeja: logistista regressiota, tukivektorikonetta, päätöspuuta, satunnaismetsää, XGBoost:a ja naiivia Bayesin luokitinta. Näistä parhaiten syvävääärennöskuvien tunnistuksessa joukosta, joka sisälsi vain matalaresoluutioisia kuvia, suoriutui tukivektorikone 98,27 prosentin tunnistustarkkuudella ja joukosta korkearesoluutioisia kuvia satunnaismetsä 99,97 prosentin tarkkuudella. Lisäksi tutkimuksessa hyödynnettiin äänestysluokittelijaa (eng. voting classifier), eli mallia, jossa hyödynnetään useampaa edellä mainittua koneoppimisalgoritmia, ja käytetään näiden tuottamista tunnistuksista saatua keskiarvoa tai enemmistöä lopullisessa tunnistuksessa. Tutkimuksessa äänestysluokittelija saavutti 98,72 prosentin tarkkuuden syvävääärennösten tunnistuksessa joukosta, joka sisälsi sekä matala- että korkearesoluutioisia kuvia. (Pandey ja muut 2024.)

Koneoppimisteknologioiden avulla saavutettujen tunnistustulosten logiikka eroaa merkittävästi tavasta, jolla ihminen pyrkisi tunnistamaan kuvan syvävääärennökseksi, minkä

vuoksi Wang ja kumppanit (2025) pitävät tärkeänä pystyä selittämään luokittelutuloksen perusteet tavallisille ihmisille. Luotettavuuden tunteen lisäksi läpinäkyvämpi luokittelu voisi auttaa myös asiantuntijoita tunnistamaan piileviä eroja aitojen ja synteettisten kuvien välillä. (Wang ja muut 2025.) Selitettävällä tekoälyllä (eng. explainable AI, XAI) on mahdollista tarkastella koneoppimisteknologioiden tunnistusprosessia lähemmin sen tuottamien visualisointien, kuvan osioiden korostamisella (eng. masking) ja tuottamalla numeraalista dataa tunnistusprosessista ja tunnistusmallin käyttämistä painotuksista eri piirteiden havaitsemista kohtaan (Jheelan & Pudaruth 2025).

4.2 Tunnistaminen fysikaalisten ja fysiologisien epäjohdonmukaisuuksien avulla

Syväväärennöskuvia on mahdollista tunnistaa tarkastelemalla kuvassa esiintyviä fysikaalisia tai fysiologisia epäjohdonmukaisuuksia. Näistä ensimmäiseen joukkoon kuuluvat esimerkiksi valonlähteisiin ja varjoihin liittyvät erikoisuudet tai artefaktit, esimerkiksi varjon suuntautuminen väärin, kun taas toiseen joukkoon kuuluvat ihmisen kasvoissa tai muualla kehossa esiintyvät anomaliat, esimerkiksi symmetriassa, pupillien muodossa ja iiristen värissä. Ongelmana tällaisessa tunnistamisessa on, että sen toimivuus on riippuvainen kasvojen suuntauksesta kuvassa ja valaistuksesta. Esimerkiksi heijastukset voivat tuottaa vääriä positiivisia tuloksia aidoista kuvista, jos näitä ei olla otettu huomioon ennalta. Lisäksi haasteena fysikaalisten ja fysiologisten ominaisuuksien avulla tunnistamisessa on, että syväväärennösteknologiat kehittyvät hyvin nopeaan tahtiin. Mallien kehitys on vähentänyt virheiden määrää ja olemassa olevat virheet ovat yhä hienovaraisempia ja siten vaikeammin havaittavia. (Wang ja muut 2025.)

4.3 Digitaaliset tunnisteet

Lohkoketjuja on mahdollista hyödyntää kuvien autenttisuuden varmentamisessa tallentamalla niihin kuvan luomiseen liittyviä tietoja kuten tarkan kuvan luomisajankohdan ja tekijän identiteettiin liittyviä tietoja (Negi ja muut 2024). C2PA-standardia (eng. Coalition for Content Provenance and Authenticity) tukevat laitteet voivat tallentaa kuvaamisajankohdan ja kunkin pikselin sijainnin kuvassa ja tämä tieto voitaisiin tallentaa digitaaliseen allekirjoitukseen. Allekirjoitus taas puolestaan vietäisiin lohkoketjuun, josta sisällön käyttäjä pystyisi varmentamaan kuvan aitouden. (Kharvi 2024.)

PUF-tunnisteissa (eng. physical unclonable functions) hyödynnetään alkuperäisen kuvausvälineen valmistusprosessissa syntyviä yksilöllisiä ominaisuuksia. Näitä ominaisuuksia voivat olla esimerkiksi yksilölliset erot eri sirujen välillä signaalin etenemisnopeudessa, ja niitä hyödyntäen laite pystyy antamaan digitaalisena sormenjälkenään satunnaisen bittijonon. Bathalapallin ja muiden (2024) tutkimuksessa PUF:a hyödynnettiin

kasvon piirteiden koordinaattien kanssa muodostettaessa kuvalle tunniste. Heidän tutkimuksensa ei ota kantaa siihen, mihin PUF tulisi tallentaa myöhempää kuvan autenttisuuden tarkastamista varten. (Bathalapalli ja muut 2024.)

Vesileimoissa (eng. watermark) kuvan luomiseen liittyviä metatietoja tai muuta dataa piilotetaan kuvaan itsessään niin, että ihmissilmä ei pysty havaitsemaan kyseistä tietoa, mutta ohjelman on mahdollista lukea kyseiset tiedot kuvasta (Negi ja muut 2024). Vesileimoja on myös mahdollista hyödyntää siten, että kuvan metatietojen sijaan kuvaan upotetaan ominaisuuksia, jotka aiheuttavat kuvan laadun heikkenemisen silloin, kun kuva pyritään dekoodaamaan piirreverkosta takaisin kuvaksi. Toimintaperiaate tässä SLIC-menetelmässä (eng. secure learned image codec) perustuu samaan ideaan kuin vastustavissa hyökkäyksissä: hienovaraisten ominaisuuksien avulla kuva luokittuu neuroverkossa väärin, ja siten vaikeutetaan neuroverkon toimintaa, vaikka kuva näyttäisi edelleen ihmissilmälle täysin normaalilta. (Huang & Wu 2024.)

Vesileimojen heikkoutena on, että kuvan muuntuminen voi heikentää tunnisteen toimivuutta merkittävästi. Erityisen alttiita vesileimat ovat niin kutsutuille affiinisille muutoksille (eng. affine transform), esimerkiksi kuvan eri alueiden skaalaamiselle ja venyttämiselle. Muita vesileimojen toimivuutta heikentäviä kuvan muuntelutapoja ovat muun muassa kuvan rajaaminen (eng. cropout), kuvan vaalentaminen (eng. lightening), gauss-suomennus, kuvan terävöittäminen (eng. sharpening), histogrammitasaus (eng. histogram equalization) ja kompressointi JPEG-kuvaksi (eng. JPEG compression). (Huang & Wu 2024).

4.4 Tietoisuuden lisääminen tekoälyteknologioista ja syväväärennöksistä

Julkisilla käsityksillä syväväärennössisällöistä on merkitystä sille, kuinka hyvin näitä sisältöjä pystytään erottelemaan muusta sisällöstä. Näin ollen tietoisuuden lisääminen syväväärennösteknologioista on tärkeää pyrittäessä torjumaan niiden haittoja. (Kharvi 2024.) Ali ja kumppanit (2021) tarkastelevat tutkimuksessaan luomiensa oppimisaktiviteettien vaikutusta yläkoululaisten (eng. middle school) ymmärrykseen generatiivisesta tekoälystä. Aktiviteetteja oli yhteensä 5 jaoteltuna eri osa-alueisiin: tutustuminen generatiivisiin tekoälymalleihin, GAN-mallin komponenttien ymmärtäminen, GAN-mallin käyttökohteiden ymmärtäminen, syväväärennöskuviin tutustuminen ja niiden tunnistaminen ja sisällön luominen generatiivisilla malleilla. Ennen aktiviteetteja ja aktiviteettien jälkeen teetettyjen kyselyiden perusteella aktiviteetit lisäsivät lasten ymmärrystä generatiivisten mallien teknisistä komponenteista, sovelluskohteista ja niiden eettisistä ulottuvuuksista. (Ali ja muut 2021.)

Kuten aiemmin todettiin, fysikaalisten ja fysiologisten piirteiden avulla tunnistaminen on vaikeutunut merkittävästi tekoälymallien kehittymisen myötä. Tällä on ollut vaikutusta erityisesti siihen, kuinka hyvin ihminen pystyy tunnistamaan syväväärennöskuvia aidoista kuvista. Lagon ja muiden (2022) tutkimuksessa havaittiin, että osallistujat pystyivät tunnistamaan syväväärennöskasvokuvan 80 prosenttia testauskuvista, joissa oli hyödynnetty vuonna 2017 sen aikaista GAN-mallia. Sen sijaan, kun osallistujia pyydettiin tunnistamaan kasvokuvia, jotka olivat luotu vuosien 2018 ja 2019 malleilla, tunnistusprosentti heikkeni merkittävästi ollen viimeisenä mainitussa vain 26 prosenttia. Lisäksi osallistujat pitivät 2019 mallin luomia kuvia realistisempina kuin aitoja kuvia. (Lago ja muut 2022.) Siten pidän erityisen tärkeänä juuri tietoisuutta ilmiöstä ja kriittistä suhtautumista sosiaalisessa mediassa vastaan tuleviin kuviin. Mikäli henkilöllä ei ole ymmärrystä ilmiöstä ja sen yleisyydestä, hänen mahdollisuutensa edes löytää keinoja kuvien tunnistamiseen syväväärennöksiksi ovat olemattomat.

5 Keskustelu

Syväväärennöskuvien luoma uhka on todellinen ja kasvava niiden määrän jatkuvan lisääntymisen ja laadun paranemisen vuoksi. Syväväärennöskuvia tuottavien syväneuroverkkojen keskeisimmiksi arkkitehtuureiksi tunnistettiin tutkimuskirjallisuuden perusteella GAN-, VAE ja diffuusiomallit ja tekniikoiksi kasvojen vaihtaminen, kasvojen uudelleenesitys, koko kasvojen synteesi ja kasvojen ominaisuuksien muokkaus. Tunnistusmenetelmiksi havaittiin tunnistaminen koneoppimisteknologioilla, digitaalisten tunnisteiden avulla, fysikaalisten ja fysiologisten epäjohdonmukaisuuksien avulla ja tietoisuuden lisääminen syväväärennösteknologioista.

Tunnistaminen koneoppimismenetelmillä sekä fysikaalisten ja fysiologisten epäjohdonmukaisuuksien avulla tarjoavat mahdollisuuden yrittää tunnistamista ilman kuviin sisällytettävää tai muualle varastoitavaa dataa. Koneoppimismenetelmät vaativat itsessään paljon työtä ja asiantuntijuutta, jotta mallit saadaan koulutettua tunnistamaan syväväärennöskuvia. Lisäksi edellä mainitut tunnistusmenetelmät vaativat jatkuvaa kehittämistä pystyäkseen vastaamaan syväväärennöskuvien tuottamisteknologioiden ja tunnistamista häiritsevien menetelmien, kuten vastustavien hyökkäysten, kehitykseen. Vesileimat tarjoavat tästä riippumattoman lähestymistavan tunnistukseen, mutta ovat taas alttiita kuvan muutoksille. Ulkoisiin säilöihin, esimerkiksi lohkoketjuihin, sijoitettavat tunnisteet tarjoavat mahdollisuuden välttää vesileimojen sudenkuopat, mutta vaativat taas standardeista sopimisen ja infrastruktuurin luonnin tunnisteita varten. Kirjallisuuskatsauksen perusteella näyttäisi selvältä, että mikään keino yksinään ei riitä vastaamaan syväväärennöskuvien luomaan uhkaan, vaan olennaista tulee olemaan eri menetelmien yhdistely.

Tutkimusprosessin aikana tutkielman aiheessa tapahtui pientä siirtymää. Alun perin tarkoituksenani oli tarkastella laajemmin syväväärennöskuvien vaikutuksia sosiaalisessa mediassa ja tähän olin alun perin suunnitellut kolmannen tutkimuskysymyksen. Kirjoittamisprosessin aikana rajaus kuitenkin siirtyi käsittelemään yhä lähemmin nimenomaan syväväärennöskuvien tuottamis- ja tunnistamistekniikoita. Haasteena alkuperäisessä asetelmassa oli pyrkiä pitämään aihe riittävästi tietojenkäsittelytieteen tutkimuksessa eikä laajentaa sitä sosiologisten tieteiden suuntaan. Olkoonkin siten, että tutkielmani pyrkii etsimään nimenomaisesti keinoja syväväärennöskuvien vaikutusten minimoimiseen, kun taas jätän näiden vaikutusten syvällisemmän luonteen ymmärtämisen sosiaali- ja yhteiskuntatieteiden tutkijoiden vastuulle. On kuitenkin huomattava, että aiheeni on vahvasti näiden vaikutusten ymmärtämisen motivoima ja tiedonhakulausekkeeni heijastelevat tätä voimakkaasti. Jälkiviisaana nyt olisin muuttanut tiedonhakulausekkeitani palvelemaan

paremmin nykyisiä tutkimuskysymyksiäni. Koska tietämykseni generoivista tekoälyteknologioista lisääntyi merkittävästi tutkimusprosessin aikana, saattaisin nyt myös lähestyä tiedonhakua hyödyntäen esimerkiksi arkkitehtuurimalleja hakulausekkeissa.

Tutkielma ei ota kantaa siihen, määritelläänkö tekoälyteknologioilla tuotettu kuva oikeasta ihmisestä aina syväväärennökseksi vai määritteleekö esimerkiksi käyttötarkoitus valheellisen informaation levittämiseen tämän. Esimerkiksi vuonna 2023 kuvanluontiohjelma Midjourneyn Reddit -keskustelualueelle lisättiin kuva paavi Francisista kalliinoloisessa valkoisessa toppatakissa, mutta kuva lähti leviämään myöhemmin Twitterissä. Näistä alkuperäisessä julkaisukanavassa kuvan tiedettiin olevan keinotekoinen, kun taas Twitterissä kuvaa erehdyttiin luulemaan aidoksi. (Novak 2023.) Toisaalta vaikka tekoälyteknologioilla tuotettu kuva olisikin tarkoitettu taiteeksi ja myöhemmin päätyisi muuhun käyttöön, sen tunnistamiseen pätisivät samat lainalaisuudet kuin syväväärennöksiksi alun perin luotujen kuvien tunnistamisessa. Kuva voitaisiin yrittää edelleen tunnistaa koneoppimisteknologioiden, fysikaalisten ja fysiologisten poikkeavuuksien ja tunnisteiden tai niiden puuttumisen avulla.

Tutkielma vastasi kysymykseen, kuinka syväväärennöskuvia voitaisiin tunnistaa, mutta ei vastannut siihen, miten esimerkiksi lainsäädännöllä voitaisiin minimoida syväväärennöskuvien aiheuttamia riskejä tai minkälaista vastuuta tekoälyteknologioita kehittäviltä yrityksiltä tulisi vaatia. Koska kyseessä ovat nykyisessä muodossaan verrattain uusi ilmiö, lainsäädännöllä on ollut haasteita vastata sen vaikutuksiin. EU hyväksyi kesäkuussa 2024 ensimmäisenä maailmassa tekoälyä koskevan lainsäädännön. Asetuksen mukaan generatiivisia tekoälymalleja ei luokitella suurriskisiksi järjestelmiksi, mutta niiden tulee noudat-taa EU:n tekijänoikeuslainsäädäntöä, pystyä ilmaisemaan tieto tekoälyllä tuotetusta sisällöstä, estää laittoman sisällön tuottaminen ja julkaista yhteenvetoja koulutukseen käytetystä tekijänoikeuksin suojatusta datasta. (Euroopan parlamentti 2023.) Esimerkiksi Google lisää nykyään omilla tekoälyteknologoillaan tuotettuihin sisältöihin SynthID -vesileiman ja tarjoaa mahdollisuuden kyseisen vesileiman hyödyntämiseen kumppaneilleen (Google DeepMind 2025).

6 Yhteenveto

Tässä tutkielmassa tarkastelin syväväärennöskuvien tuottamismalleja ja tekniikoita, vaikutuksia sosiaalisessa mediassa ja keinoja vaikutusten minimoimiseen tunnistusmenetelmien ja tietoisuuden lisäämisen avulla. Kokonaisuutena syväväärennöskuvat ovat ja tulevat olemaan merkittävä haaste sosiaaliselle medialle niiden määrän kasvaessa ja laadun parantuessa. Tunnistusmenetelmien on kehityttävä erottamaan poikkeavuuksia verrattuna aitoihin kuviin ja toisaalta pystyttävä sietämään yrityksiä vaikeuttaa tunnistamista esimerkiksi vastustavilla hyökkäyksillä. Koska tunnistusmenetelmät eivät pysty antamaan itsessään täyttä varmuutta kuvan aitoudesta, ihmisten tietoisuudella ilmiöstä on merkitystä syväväärennöskuvien vaikutusten minimoimisessa.

Lähdeluettelo

Ali, S., DiPaola, D., Lee, I., Hong, J., & Breazeal, C. (2021). Exploring Generative Models with Middle School Students. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445226

Bathalapalli, V. K. V. V., Yanambaka, V. P., Mohanty, S., Kougianos, E. (2024). PUFshield: A Hardware-Assisted Approach for Deepfake Mitigation Through PUF-Based Facial Feature Attestation. *Proceedings of the Great Lakes Symposium on VLSI 2024,* (pp. 676–681). ACM. https://doi.org/10.1145/3649476.3660394

Drori, I. (2023). *The science of deep learning (First edition.)*. Cambridge University Press.

Euroopan parlamentti. (2023). *EU:n tekoälysäädös on ensimmäinen laatuaan.* Euroopan parlamentti. https://www.europarl.europa.eu/topics/fi/article/20230601STO93804/eu-n-tekoalysaados-on-ensimmainen-laatuaan. Haettu 9.11.2025.

Google DeepMind. (2025). SynthID: A tool to watermark and identify content generated through AI. https://deepmind.google/models/synthid/. haettu 21.11.2025.

Huang, C., Wu, J. (2024). SLIC: Secure Learned Image Codec through Compressed Domain Watermarking to Defend Image Manipulation. *In Proceedings of the 6th ACM International Conference on Multimedia in Asia (MMAsia '24)*. Association for Computing Machinery, New York, NY, USA, Article 1, 1–7. https://doi-org.libproxy.tuni.fi/10.1145/3696409.3700161

Jheelan, J., Pudaruth, S. (2025). Using Deep Learning to Identify Deepfakes Created Using Generative Adversarial Networks. *Computers (Basel)*, 14(2), 60. https://doi.org/10.3390/computers14020060

Kharvi, P. L. (2024). Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media. *IEEE Security & Privacy*, 22(4), 115–122. https://doi.org/10.1109/MSEC.2024.3405963

Lago, F., Pasquini, C., Bohme, R., Dumont, H., Goffaux, V., & Boato, G. (2022). More Real Than Real: A Study on Human Visual Perception of Synthetic Faces. *IEEE Signal Processing Magazine*, 39(1), 109–116. https://doi.org/10.1109/MSP.2021.3120982

Lyu, S. (2024). DeepFake the menace: mitigating the negative impacts of AI-generated content. *Organizational Cybersecurity Journal: Practice, Process and People*, *4(1)*, 1–18. https://doi.org/10.1108/OCJ-08-2022-0014

Mallet, j., Pryor, l., Dave, r., Vanamala, M. (2023). Deepfake Detection Analyzing Hybrid Dataset Utilizing CNN and SVM. *In Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence* (ISMSI '23). Association for Computing Machinery, New York, NY, USA, 7–11. https://doi-org.libproxy.tuni.fi/10.1145/3596947.3596954

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2022). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, 53(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

Negi, N., Kaushal, D., Bahl, S., & Mer, A. (2024). Deepfake Technology: An In-depth Review of the Emerging Threats Posed by Artificial Intelligence to Society. *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, 270–274. https://doi.org/10.1109/EmergIN63207.2024.10961120

Negreiro, M. (2025). Children and deepfakes. European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775855/EPRS_BRI(2025)775855_EN.pdf (Haettu 5.10.2025)

Novak, M. (2023). That Viral Image Of Pope Francis Wearing A White Puffer Coat Is Totally Fake. *Forbes.* https://www.forbes.com/sites/mattnovak/2023/03/26/that-viral-image-of-pope-francis-wearing-a-white-puffer-coat-is-totally-fake/ (Haettu 8.11.2025)

Pandey, M., Singh, S., Malik, A., & Kumar, R. (2024). Detecting low-resolution deepfakes: an exploration of machine learning techniques. *Multimedia Tools and Applications, 83(25)*, 66283–66298. https://doi.org/10.1007/s11042-024-18235-7

Wang, X., Tsai, T., Lin, L., Guo, L., Hu, S., Chang, M.., Atrey, P., Lyu, S. (2025). Spotting the Fakes: A Deep Dive into GAN-Generated Face Detection. *ACM Trans. Multimedia Comput. Commun. Appl. 21, 7, Article 193 (July 2025)*. https://doi.org/10.1145/3742786