Syväväärennösten tunnistaminen koneoppimismenetelmillä

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Maaliskuu 2026
Mikko Kaijala

TURUN YLIOPISTO
Tietotekniikan laitos

Mikko Kaijala: Syväväärennösten tunnistaminen koneoppimismenetelmillä

TkK-tutkielma, 17 s.

Maaliskuu 2026

Syväväärennös eli deepfake on tekoälyn avulla tuotettu tai muokattu video, jota hyödynnetään yhä enemmän misinformaation ja huijausten yhteydessä. Tämän tutkielman tavoitteena on tarkastella, millaisia poikkeamia voidaan hyödyntää syväväärennösten tunnistuksessa sekä arvioida koneoppimisen mallien tarkkuutta tunnistuksessa. Tutkielma toteutettiin kirjallisuuskatsauksena, jossa analysoitiin neljää syväväärennösten tunnistukseen keskittyvää tutkimusta. Tulokset osoittavat, että koneoppimisen mallit pystyvät tunnistamaan syväväärennöksiä korkealla tarkkuudella hyödyntämällä tilallisia, ajallisia ja biometrisiä poikkeamia. Mallien suorituskyky riippuu kuitenkin käytetystä aineistosta ja menetelmistä.

Asiasanat: syväväärennös, deepfake, koneoppiminen, syväoppiminen, misinformaatio, kuvatunnistus, videotunnistus

Sisällys

1 Johdanto

Digitaalisesta misinformaatiosta on tullut niin haitallista, että World Economic Forum (WEF) on luokitellut misinformaation yhdeksi pääuhkaksi yhteiskunnalle [1]. Kun valetieto saa sosiaalisessa mediassa edes vähänkään vahvistusta, sitä uskovien tahojen uskomuksia on hyvin vaikeaa korjata [2]. Misinformaatio voi vaikuttaa ihmisten uskomuksiin politiikasta, terveydestä, ympäristöstä ja yleisesti yhteiskunnastamme [3]. Tilannetta pahentaa ennestään se, kuinka vakuuttavissa muodoissa misinformaatio esiintyy. Yksi näistä esiintymismuodoista on syvävääreennös tai deepfake, eli tekoälyllä manipuloitu video tai kuva, joka pyrkii esittämään todellista henkilöä [4].

Syvävääreennökset ovat vuosien aikana kehittyneet niin paljon, että niitä on yhä vaikeampi erottaa aidoista videoista. Niitä hyödynnetään nykyään huijauksissa, rikoksissa, poliittisessa vaikuttamisessa ja yleisen mielipiteen muovaamisessa [5]. Kyseisen teknologian käyttö haitallisiin tarkoituksiin on jyrkässä kasvussa. Yrityksen Sumsub vuonna 2024 tehdyn identiteettihuijausraportin mukaan syvävääreennöshuijausten määrä nelinkertaistui vuodesta 2023 ja 7% kaikista huijausyrityksistä vuonna 2024 olivat deepfake-pohjaisia. Suurimmat nousut syvävääreennöshuijauksissa olivat Lähi-idässä jossa oli 643% kasvu, Afrikassa jossa oli 393% kasvu ja Latinalaisessa Amerikassa ja Karibian alueella jossa oli 255% kasvu. [6]

Vuonna 2021 tehdyssä tutkimuksessa nimeltä 'deepfake detection by human crowds, machines and machine-informed crowds' [7], testattiin ihmisten kykyä tun-

nistaa syväväärennöksiä. Testaus suoritettiin esittelemällä kaksi videota - yksi syväväärennös ja toinen aito video, joista heidän piti valita kumpi videoista oli syväväärennös. Osallistujista, jotka läpäisivät tarkastusvaiheen, tunnisti syväväärennöksiä 66% tarkkuudella. Rekrytoimattomilla osallistujilla oli hieman parempi tarkkuus, 69%. Vaikka tulokset ovat suhteellisen vahvoja, ne kertovat kuinka kehittyneitä syväväärennökset ovat jo. Ajan myötä syväväärennösten realistisuus vain parantuu, ja on ennustettu, että tulevaisuudessa GAN-mallien (Generative Adversarial Networks) eli syväväärennöksiä tuottavien mallien avulla voidaan luoda yhä vakuuttavampia syväväärennöksiä, joissa voidaan jopa vaihtaa henkilöiden koko vartaloita toiseen [4]. On siis hyvin todenäköistä, että ihmisten kyky erottaa syväväärennöksiä aidoista videoista todenäköisesti heikkenee merkittävästi tulevaisuudessa. Tämän vuoksi tarvitaan luotettavia menetelmiä syväväärennösten tunnistukseen.

Tämän tutkielman tarkoitus on tutkia, mitä poikkeamia voidaan hyödyntää syväväärennösten tunnistuksessa ja kuinka tehokkaita erilaiset koneoppimisen mallit ovat tunnistuksessa. Syväväärennösten haittavaikutukset ovat jyrkässä kasvussa ja syväväärennösten luomisen menetelmät paranevat vain vuodesta toiseen tuottaen yhä realistisempia väärennöksiä [4]. Tunnistukseen tarvitaan siis tarkkoja tunnistusmenetelmiä, vähentääkseen syväväärennösten aiheuttamia haittavaikutuksia.

Tämän työn **tutkimuskysymykset** ovat:

TK1. Mitä poikkeamia voidaan hyödyntää syväväärennösten tunnistuksessa?

TK2. Kuinka tarkkoja koneoppimisen mallit ovat syväväärennösten tunnistuksessa?

Tässä tutkielmassa tarkastellaan syväväärennösten tunnistusta koneoppimisen menetelmien avulla kirjallisuuskatsauksen muodossa. Luvussa 2 esitellään syväväärennöksiin ja niiden tuottamiseen liittyvä taustatieto. Luvussa 3 analysoidaan neljä syväväärennösten tunnistukseen keskittyvää tutkimusartikkelia, joissa hyödynnetään erilaisia koneoppimisen ja syväoppimisen malleja. Artikkelien pohjalta tarkas-

tellaan käytettyjä aineistoja, hyödynnettyjä poikkeamatyyppejä sekä mallien toimintaa ja tunnistustarkkuuksia. Luvussa 4 pohditaan tuloksia ja niiden merkitystä syvävääreennösten tunnistuksen kannalta, ja luvussa 5 esitetään työn keskeiset johtopäätökset.

Menetelmät ja tiedonhaku

Tutkielma toteutettiin kirjallisuuskatsauksena. Tietokannat, joissa haku toteutettiin ovat IEEE, SSRN, ACM, PNAS. Nämä valikoituvat tietokannoiksi, sillä haku tuotti näissä tietokannoissa merkittävästi enemmän olennaisia artikkeleita, kuin muissa tietokannoissa. Käytetyt hakutermit olivat ("deep fake"OR deepfake OR deep-fake) AND ("deepfake detection OR misinformation").

Tavoitteena oli muodostaa yleiskuva syvävääreennösten tunnistukseen käytettävistä koneoppimismenetelmistä sekä arvioida niiden tarkkuutta. Kirjallisuuskatsaus valittiin menetelmäksi, koska syvävääreennösten tunnistuksesta on julkaistu viime vuosina monia tutkimuksia.

Hakutulokset sisälsivät useita syvävääreennöksiä käsitteleviä tutkimuksia, joista merkittävä osa keskittyi syvävääreennösten luontiin tai syvävääreennösten yleiseen tarkasteluun. Tutkielmaan valittiin neljä artikkelia laajemmasta joukosta sillä perusteella, että ne edustivat erilaisia koneoppimisen arkkitehtuureja ja tunnistusmenetelmiä sekä raportoivat selkeästi mitattavat tunnistustarkkuudet. Valinnan tavoitteena ei ollut kattaa kaikkia aiheesta julkaistuja tutkimuksia, vaan muodostaa monipuolinen kokonaiskuva keskeisistä lähestymistavoista syvävääreennösten tunnistuksessa. Tutkielmaan valittiin lopulta neljä tutkimusartikkelia, jotka täyttivät seuraavat kriteerit:

1. Tutkimus keskittyi nimenomaan syvävääreennösten tunnistukseen
2. Menetelmänä oli koneoppiminen tai syväoppiminen
3. Tutkimuksessa raportoitiin selkeät tunnistustarkkuudet.

2 Taustatietoa

Tässä luvussa esitellään syväväärennöksiin liittyvät keskeiset käsitteet ja teknologiat. Luvun tarkoitus on kuvata, mitä syväväärennökset ovat ja millaisiin menetelmiin niiden tuottaminen perustuu. Tämä taustatieto tukee seuraavassa luvussa tarkasteltavien syväväärennösten tunnistusmenetelmien ymmärtämistä ja auttaa hahmottamaan, millaisiin teknisiin ominaisuuksiin tunnistus perustuu.

2.1 Syväväärennökset ilmiönä

Syväväärennös tai **deepfake** on tekoälyn avulla luotu tai muokattu videoväärennös, joka pyrkii näyttämään aidolta videolta [4]. Syväväärennösten kohteena on usein poliittinen tai muu julkinen henkilö, joiden puhetta ja kasvon eleitä pyritään imitoimaan neuroverkkojen avulla.

Termiä 'deepfake' käytettiin ensimmäisen kerran vuonna 2018, kun anonyymi Reddit-käyttäjä loi foorumin, joka oli tarkoitettu syväväärennösten luontiin ja jakeluun. Foorumilla jaettiin neuroverkkojen avulla luotuja face-swap eli naamanvaihtovideoita, joissa vaihdettiin julkisuuden henkilöiden kasvoja aikuisviihteeseen. [5] Teknologia oli silloin vielä hyvin alkeellista ja väärennöksistä huomasi selvästi, että ne eivät olleet aitoja. Nykypäivänä kuitenkin syväväärennösten tuotanto on niin kehittynyttä, että niitä on hankala erottaa aidoista videoista. Kehityksen myötä tätä teknologiaa on myös aloitettu käyttämään entistä haitallisempiin tarkoituksiin. Esimerkiksi syväväärennöksillä voidaan levittää valetietoa, manipuloida yleistä mie-

lipidettä ja mustamaalata yksittäisiä henkilöitä [5]. Kyseistä teknologiaa hyödynnetään myös rikollisuudessa, ja Sumsubin vuonna 2023 tehdyn raportin mukaan AI-pohjaiset identiteettivarkaudet kuuluivat viiden yleisimmän identiteettivarkauden tyypin joukkoon [8].

2.2 Syväväärennösten luontimenetelmät

Väärennösten tuottamiseen hyödynnetään neuroverkkoja, jotka analysoivat suuria määriä dataa, kuten kuvia, videoita ja ääninäytteitä. Analysoinnin kautta neuroverkko oppii imitoimaan henkilön kasvon eleitä, ääntä ja muita maneereja. Syväväärennöksiä tuotetaan pääosin Generative Adversarial Networks -malleilla (GAN), jotka muodostuvat kahdesta kilpailevasta neuroverkosta: generaattorista ja diskriminaattorista. Generaattorin tehtävänä on luoda mahdollisimman aidon näköistä dataa, kuten kuvia tai videoita, jonka jälkeen diskriminaattori arvioi, onko tuotettu media aito vai väärennös. Tämä vuorovaikutus on jatkuvaa ja toistuu, kunnes generaattori kykenee 'huijaamaan' diskriminaattoria. Vuorovaikutuksen myötä generaattori oppii tuottamaan erittäin realistista sisältöä, jota ihminen ei välttämättä osaa erottaa väärennökseksi. [4]

Autoencoder tai **autoenkoodaaja** on toinen tapa luoda syväväärennöksiä. Autoenkoodaajia käytetään kasvojen vaihtamiseen eli face-swap-teknologiassa. Malli koostuu yhteisestä enkoodajasta ja kahdesta dekoodaajasta. Enkoodaajan tehtävänä on tiivistää kuvasta kaikki olennaiset piirteet kompressoituun muotoon. Dekoodaajan tehtävänä on rekonstruktoida alkuperäinen syöte mahdollisimman tarkasti kompressoidusta datasta. Syy miksi syväväärennösten luonnin yhteydessä käytetään kahta dekoodaajaa on se, että toinen dekoodaaja oppii rekonstruktoimaan alkuperäisen henkilön kasvot ja toinen kohdehenkilön. Lopputuloksena on realistinen video, jossa alkuperäisen henkilön ilmeet ja liikkeet säilyvät, mutta kasvot on korvattu toisen henkilön kasvoilla. [9]

3 Mallien tarkastelu

Tässä luvussa tarkastellaan neljää eri tutkimusartikkelia, jotka käsittelevät syvävääärennösten tunnistusta eri koneoppimisen mallien avulla. Tutkimusartikkelit hyödyntävät erilaisia malleja, datakokoelmia ja tunnistuksen menetelmiä. Tarkoituksena on tutkia miten erilaiset mallit toimivat, millaisia menetelmiä ne hyödyntävät, millaisia ominaisuuksia mallit hyödyntävät tunnistuksessa ja millaisella tarkkuudella ne tunnistavat syvävääärennöksiä.

Taulukossa 3.1 esitetään tarkasteltujen tutkimusten tekijät, käytetyt koneoppimismallit sekä testauksessa hyödynnetyt aineistot. Taulukko havainnollistaa, että tarkastellut tutkimukset hyödynsivät erilaisia koneoppimismalleja ja testiaineistoja syvävääärennösten tunnistuksessa.

Taulukko 3.1: Tutkittujen artikkelien tekijät, koneoppimismallit ja testiaineistot. [10][11][12][13]

Tutkimus	**Koneoppimismalli**	**Testiaineisto**
Liwei Deng	EfficientNet-V2	FaceForensics++
Joshi Paritosh	Xception	deepfake_faces
Georgios Petmezas	CNN-LSTM-Transformer	DeepFakeDetection, Celeb-DF, FaceForensics++
Deressa Wodajo	CViT	DeepFake Detection Challenge

3.1 Aineiston esittely

Ensimmäinen tarkasteluun valittu koneoppimismalli on CNN-pohjainen EfficientNet-V2-malli. EfficientNet on Googlen kehittämä syväoppimismalli. Tarkasteluun on valittu Liwei Dengin vuonna 2022 julkaistu tutkimus "Deepfake Video Detection Based on EfficientNet-V2 Network". Deng käsittelee artikkelissaan EfficientNet-V2-verkon käyttöä syvävääreennösten tunnistamiseen. Mallin testauksessa hyödynnettiin FaceForensics++-datakokoelmaa. Datakokoelma sisältää suuren määrän sekä manipuloituja että aitoja videoita. [10]

Xception-konvoluutioneuroverkko on toinen CNN-pohjainen malli, joka valittiin tarkasteluun. Joshi Paritoshin vuonna 2024 kirjoitettu tutkimus "Deep Fake Image Detection using Xception Architecture" käsittelee Xceptionin suorituskykyä syvävääreennösten tunnistuksessa. Mallin testauksessa käytettiin deepfake_faces-nimistä datakokoelmaa, joka sisältää noin 90000 aitoa sekä manipuloitua kuvaa. [11]

CNN-LSTM-Transformer-hybridimalli on kolmas tarkasteluun valittu koneoppimismalli. Georgios Petmezasin vuonna 2024 kirjoitettu tutkimus "Video Deepfake Detection Using a Hybrid CNN-LSTM-Transformer Model for Identity Verification" käsittelee kolmen mallin yhdistelmää syvävääreennösten tunnistukseen. Mallin koulutuksessa käytettiin VoxCeleb2-datakokoelmaa, ja testauksessa hyödynnettiin FaceForensics++-, Celeb-DF- sekä DeepFakeDetection-kokoelmia. [12]

Convolutional Vision Transformer on neljäs tarkasteluun valittu koneoppimismalli. Deressa Wodajon ja Solomon Atnafun vuonna 2020 kirjoitetussa tutkimuksessa käytetään CNN- ja Vision Transformer -pohjaisten arkkitehtuurien yhdistelmää. Malli testattiin DFDC-aineistolla (Deepfake Detection Challenge), joka on yksi laajimmista saatavilla olevista deepfake-videoaineistoista. [13]

3.2 Analysoitavat ominaisuudet

Syvävääreennösten tunnistuksessa havaittavat poikkeamat on jaettu selkeyden vuoksi kolmeen kategoriaan: tilalliset, ajalliset ja biometriset poikkeamat. Tilalliset poikkeamat liittyvät yksittäiseen kuvassa näkyvään sisältöön, kuten kasvojen rakenteellisiin vääristymiin, valaistuksen epäjohdonmukaisuuksiin tai ihon tekstuurivirheisiin. Ajalliset poikkeamat ilmenevät videon edetessä, esimerkiksi kasvojen liikkeiden epäjohdonmukaisessa rytmissä tai ilmeiden käyttäytymisessä. Biometriset poikkeamat perustuvat henkilön yksilöllisiin kasvonpiirteisiin, joita voidaan tarkastella kolmiulotteisina rakenteina.

EfficientNet-V2-malli analysoi videoista eroteltuja kuvakehyksiä. Malli tunnisti tilallisia poikkeamia, kuten eroja kasvojen geometriassa, ihon tekstuurissa ja valon heijastuksessa. Tutkimuksen alussa videoista erotettiin yksittäiset kasvokuvat, jonka avulla pystyttiin opettamaan mallille aitojen kasvojen ja manipuloitujen kasvojen eroja. Yleisimmät epäjohdonmukaisuudet, jotka malli huomasi, olivat vääristyneet kasvonpiirteet ja epätarkat reunalueet kasvoissa. [10]

Xception-malli analysoi yksittäisiä kuvia, joten se kiinnitti huomiota tilallisiin poikkeamiin. Artikkelissa ei eritelty mitä konkreettisia visuaalisia piirteitä malli havaitsi. Malli on kuitenkin CNN-pohjainen ja CNN-arkkitehtuuri kykenee havaitsemaan tilallisia poikkeamia kuten ihon tekstuurin epätarkkuudet ja kasvojen muodon vääristymät. Koska malli käsittelee staattisia kasvokuvia, se ei huomioi liikettä tai käyttäytymisen jatkuvuutta. [11]

CNN-LSTM-Transformer-malli analysoi kokonaisia videoita, minkä vuoksi se kykeni tarkastelemaan poikkeamia huomattavasti laajemmin kuin muut tarkasteluun valitut mallit. Malli koostuu kolmesta eri neuroverkkokomponentista: CNN vastaa kuvakehysten visuaalisesta analysoinnista, LSTM (Long Short-Term Memory) seuraa ajallista käyttäytymistä ja liikeratoja, ja Transformer-verkko tarkastelee pitkän aikavälin riippuvuuksia kehyksien välillä. Hybridimallin avulla voidaan havaita yk-

sittäisiin kehyksiin liittyviä tilallisia poikkeamia ja ajallisia epäjohdonmukaisuuksia kasvojen liikkeissä tai ilmeiden rytmissä. Tutkimuksessa hyödynnettiin lisäksi 3D Morphable Model-mallinnusta, jonka avulla pystyttiin mallintamaan kasvojen kolmiulotteisia muotoja. Tämä mahdollistaa myös biometrisiin piirteisiin kohdistuvan tarkastelun, kuten kasvojen rakenteellisten mittasuhteiden vertaamisen. [12]

CViT analysoi videoista erotettuja kasvonkuvia, joten se painotti tilallisia poikkeamia. CNN-osa mallista keskittyy tilallisten visuaalisten piirteiden oppimiseen, kun taas Transformer-osa huomioi laajempia rakenteellisia yhteyksiä kuvien sisällä. Tämä mahdollistaa sekä yksityiskohtaisten että laajempien visuaalisten rakenteiden yhteiskäsittelyn. [13]

Taulukko 3.2 havainnollistaa, millaisia poikkeamatyyppejä kukin malli osaa tunnistaa. Poikkeamatyypit ovat jaettu kolmeen kategoriaan. EfficientNet-V2, Xception ja CViT keskittyivät tilallisiin poikkeamiin eli kuvan tai videosta otetun kehyksen virheellisiin yksityiskohtiin kasvoissa. CNN-LSTM-Transformer tai lyhyemmin CNN-Hybrid tunnisti tilallisten poikkeamien lisäksi ajallisia poikkeamia eli liikkeiden yhteydessä syntyviä poikkeamia. CNN-Hybrid myös kiinnitti huomiota biometrisiin poikkeamiin eli henkilön oikeaan 3D-kasvonmuotoon liittyviin poikkeamiin.

Taulukko 3.2: Koneoppimismallien painottamat poikkeamat syväväärennöksissä

Poikkeamatyyppi	**EfficientNet-V2**	**Xception**	**CNN-Hybrid**	**CViT**
Tilalliset poikkeamat	x	x	x	x
Ajalliset poikkeamat			x	
Biometriset poikkeamat			x	

3.3 Mallien toiminta

Dengin EfficientNet-V2-tutkimuksessa videoista eroteltiin erillisiä kuvakehyksiä. Seuraavaksi data esikäsiteltiin rajaamalla kasvot ja skaalamalla kuva yhtenäiseen ko-

koon, jonka jälkeen se oli valmis mallille syötettäväksi. Seuraavaksi data syötettiin konvoluutioverkkoon, joka totesi kuvat aidoksi tai väärennetyksi. EfficientNet-V2-verkossa käytetään compound-skaalausta (engl. Compound scaling), jossa verkon syvyyttä, leveyttä ja resoluutiota säädetään mallin suunnitteluvaiheessa ennalta määritettyjen skaalauskertoimien avulla. Tämä mahdollistaa mahdollisimman tehokkaan luokittelun ilman merkittävää lisälaskentaa. Koulutuksessa malli oppii tunnistamaan visuaalisia piirteitä, jotka ovat epätyypillisiä aidoille kasvoille. Näitä piirteitä ovat aiemmin mainitut poikkeamat eli rakenteelliset vääristymät, valaistuksen virheet ja poikkeamat ihon tekstuurissa. Kun malli on koulutettu, se pyrkii luokittelemaan kasvot joko aidoiksi tai väärennökseksi. [10]

Paritoshin Xception-tutkimuksessa analysoitiin yksittäisiä kuvia. Malli toimii depthwise separable-konvoluutiolla, joka eroaa hieman perinteisestä konvoluutiosta. Kyseinen prosessi jaetaan kahteen osaan. Aluksi suoritetaan depthwise-konvoluutio, jossa jokainen syötekanava, esimerkiksi värikanavat, käsitellään omalla suodattimella. Sen jälkeen pointwise-konvoluutiossa yhdistetään eri kanavien tiedot yhdeksi kokonaisuudeksi. Tarkoituksena on, että malli oppii ensin tilalliset piirteet, kuten reunat ja tekstuurit kanavan sisällä ja sitten vasta yhdistää kanavien väliset suhteet. Tämän menetelmän tavoitteena on vähentää laskentakuormaa verratuna perinteiseen konvoluutioon. Tunnistuksessa Xception-malli analysoi yksittäisiä kuvia konvoluutioneuroverkon avulla ja oppi tunnistamaan tilallisia poikkeamia, jotka liittyivät kuvan rakenteeseen ja tekstuuriin. Luokittelu tehtiin ilman ajallisen tiedon hyödyntämistä. [11]

Petmazasin CNN-LSTM-Transformer-mallin tutkimuksessa käytetty koneoppimisen malli koostui kolmesta eri mallista: CNN, LSTM ja Transformer. Tunnistusprosessi alkaa esikäsittelyllä, jossa erotellaan yksittäiset kehykset videoista ja rajataan kasvon alueet. Tämän jälkeen kasvoista tuotetaan 3D Morphable Models-mallin avulla numeerinen esitysmuoto, joka sisältää tiedon kasvojen muodosta, il-

meistä ja asennosta. Seuraavaksi CNN-malli hoitaa tilallisten poikkeamien tunnistuksen, kuten tekstuurivirheet ja muut epäjohdonmukaisuudet kasvon yksityiskohdissa. LSTM-mallin tehtävänä on tarkastella ajallisia virheitä, kuten epätavallisia silmänräpäyksiä tai suun liikkeitä lyhyellä aikavälillä. Transformer-malli huomioi myös ajallisia poikkeamia, mutta tarkastelee niitä laajemmin eli se huomioi videon ilmeiden ja eleiden jatkuvuuden koko aikajaksolta. CNN-hybrid-malli eroaa Xceptionista ja EfficientNetistä siten, että sen tunnistus ei perustu pelkkään aito-vai-väärennösluokitteluun. Mallin tulkinta on paljon monipuolisempaa hybridiyhdistelmän vuoksi ja 3D-datan vuoksi.[12]

Wodajon CViT-tutkimuksessa esikäsiteltiin testiaineisto erottamalla kasvokuvat videoista ja muuttamalla ne yhtenäiseen kokoon. CViT-malli koostuu kahdesta pääkomponentista: piirreoppijasta (Feature Learning) ja Vision Transformerista. Feature Learning koostuu VGG-tyyppisestä konvoluutioneuroverkosta, joka koostuu 17 kerroksesta. Sen tehtävänä on muodostaa syvällisiä piirre-esityksiä kasvokuvista. Piirre-esitykset jaetaan osiin ja syötetään Transformer-arkkitehtuurille, joka tuottaa lopullisen luokituksen aidon ja väärennetyn välillä. Koulutus tehtiin yli 300 000 kuvan aineistolla. [13]

3.4 Mallien tulokset

Dengin tutkimuksessa EfficientNet-V2-mallia käytettiin syvävääreennettyjen kasvojen luokitteluun. Tarkoituksena oli löytää luotettava ja tehokas ratkaisu väärennettyjen kasvojen tunnistamiseen yksittäisistä videoruuduista. Mallia koulutettiin ja testattiin FaceForensics++-aineistolla, joka sisälsi laajan kokoelman aitoja ja manipuloituja videoita. Tuloksissa EfficientNet-V2 saavutti 94,3% tarkkuuden, mikä osoitti sen soveltuvan hyvin suurten videopohjaisten aineistojen automaattiseen läpikäyntiin. Tutkimuksessa malli esitettiin skaalautuvana ja luotettavana vaihtoehtona syvävääreennösten tunnistukseen.[10]

Xception-mallia testattiin Paritoshin tutkimuksessa, jonka tavoitteena oli arvioida mallin kykyä erottaa aidot kasvokuvat manipuloiduista kuvista. Malli testattiin deepfake_faces-aineistolla, joka sisälsi noin 90000 kuvaa. Xception saavutti testissä 93,01% tarkkuuden. Tutkimuksessa mallia ehdotettiin käytännölliseksi ratkaisuksi tilanteisiin, joissa suuri määrä kuvia tarvitsee tarkastaa nopeasti ja automaattisesti. [11]

Petmezasin tutkimuksessa CNN-LSTM-Transformer-hybridimallia kehitettiin tunnistamaan videomuotoisia syvävääreennöksiä. Mallin tavoitteena oli huomioida kasvojen käyttäytyminen, liikkeet ja rakenteelliset piirteet. Malli koulutettiin VoxCeleb2-aineistolla ja testattiin DFD-, Celeb-DF- ja FaceForensics++-aineistoilla. Tulokset osoittivat, että hybridimalli saavutti keskimäärin 96,2% tarkkuuden. Tutkimuksessa mallin käyttöä ehdotettiin erityisesti videopohjaiseen identiteetin todentamiseen ja huijausyritysten torjuntaan, joissa pelkkä kuvatasoinen tarkastelu ei ole riittävä. [12]

Wodajon tutkimuksen tavoitteena oli kehittää mahdollisimman yleistettävissä oleva malli, joka kykenee tunnistamaan syvävääreennöksiä erilaisista aineistoista. CViT-malli testattiin DFDC-aineistolla ja sen tarkkuus oli parhaimmillaan 91,5%. Tutkimuksessa havaittiin, että virheellisesti tunnistetut tapaukset johtuivat useimmiten heikkolaatuisista tai väärin rajatuista kasvokuvista. Tulokset osoittivat, että konvoluutioneuroverkon ja Transformer-arkkitehtuurin yhdistelmä on toimiva ratkaisu syvävääreennösten tunnistamiseen. [13]

Taulukossa 3.3 esitetään eri mallien tarkkuudet, käytetyt testiaineistot sekä mittaustavat, joilla syvävääreennösten tunnistusta arvioitiin. Kaikki neljä mallia saavuttivat korkean tarkkuuden, mutta mittaustapojen ja aineistojen ero tekee suoran vertailun haastavaksi. EfficientNet-V2, Xception ja CViT arvioivat syvävääreennöksiä yksittäisistä kuvista tai kuvakehyksistä, ja tarkkuus määritettiin kuvakohtaisesti oikeiden ja väärien luokitusten perusteella. Sen sijaan CNN-LSTM-Transformer kä-

sittelee videon kokonaisuutena ja arvioi syvävääärennöksen todennäköisyyttä koko videon perusteella, mikä tekee mittauksesta erilaisen. Lisäksi CNN-LSTM-Transformer testattiin useilla aineistoilla, mikä antaa sille vahvemman pohjan. Taulukko havainnollistaa, että vaikka mittausmenetelmät eroavat, kaikki kolme mallia pystyvät suoriutumaan syvävääärennösten tunnistuksesta luotettavasti valituilla aineistoilla.

Taulukko 3.3: Mallien tarkkuudet, testiaineistot ja mittaustapa[10][11][12][13]

Malli	**Tarkkuus**	**Testiaineisto**	**Mittaustapa**
EfficientNet-V2	94,3 %	FaceForensics++	Videosta erotetut kasvokuvat
Xception	93,0 %	deepfake_faces	Yksittäiset kasvokuvat
CNN-LSTM-Transformer	96,2 %	DFD, Celeb-DF, FaceForensics++	Kokonaiset videot
CViT	91,5 %	DFDC	Videosta erotetut kasvokuvat

4 Pohdinta

Tässä luvussa tarkastellaan vielä kerran tutkimuksissa saavutettuja tuloksia ja pohditaan, miten ne vastaavat tutkimuskysymykseen TK2: Kuinka tarkkoja koneoppimisen mallit ovat syvävääreennösten tunnistuksessa. Tarkastellut mallit erosivat toisistaan niiden tunnistusmenetelmien sekä analysoitavien aineistojen suhteen, mutta ne tarjoavat hyvän yleiskuvan koneoppimiseen perustuvien mallien tunnistustavoista ja suorituskyvystä.

Kaikki tarkastellut mallit saavuttivat korkean tunnistustarkkuuden eri aineistoilla, mikä osoittaa, että koneoppimisen mallit soveltuvat hyvin syvävääreennösten tunnistukseen. Mallien väliset erot eivät liittyneet pelkästään tarkkuuteen, vaan myös siihen, millaista dataa ne analysoivat ja millaisia poikkeamia ne pystyivät hyödyntämään tunnistuksessa.

Tulosten perusteella voidaan havaita, että videopohjainen CNN-LSTM-Transformer-malli, joka hyödynsi tilallisia, ajallisia ja biometrisiä poikkeamia, saavutti tarkastelluista malleista korkeimman tarkkuuden [12]. Tämä viittaa siihen, että syvävääreennöksissä ilmenee usein ajallisia epäjohdonmukaisuuksia, kuten kasvojen liikkeisiin tai ilmeiden rytmiin liittyviä poikkeamia, joita ei ole mahdollista havaita staattisista kuvista. Ajallisen tiedon hyödyntäminen vaikuttaa olevan merkittävä tekijä tunnistustarkkuuden parantamisessa, erityisesti videomuotoisessa aineistossa.

Biometrisiin piirteisiin perustuvat menetelmät mahdollistavat henkilön kasvojen kolmiulotteisten rakenteiden hyödyntämisen tunnistuksessa. Nämä menetelmät sopi-

vat tilanteisiin, joissa tunnistettava henkilö on ennalta rekisteröity järjestelmään, ja joissa tunnistuksen tavoitteena on identiteetin todentaminen. Biometrisiin poikkeamiin perustuva tunnistus voi olla hyödyllistä esimerkiksi pankkipalveluissa tai muissa turvallisuuskriittisissä järjestelmissä, joissa virheellisen tunnistuksen seuraukset voivat olla erittäin haitallisia.

Korkea tunnistustarkkuus tietyllä aineistolla ei välttämättä tarkoita, että malli yleistyy hyvin uusiin ja erilaisiin syväväärennöksiin. Mallien tarkkuus voi heikentyä, jos testiaineisto poikkeaa koulutusaineistosta. Kyseinen ongelma on yksi keskeisimmistä haasteista syväväärennösten tunnistuksessa, sillä deepfake-teknologiat kehittyvät jatkuvasti ja tuottavat yhä realistisempaa sisältöä.

Tarkastellut mallit on testattu rajatuilla aineistoilla, jotka eivät välttämättä vastaa todellisia käyttötilanteita. Todellisessa ympäristössä videot voivat olla heikkolaatuisia, huonosti rajattuja tai sisältää häiriötekijöitä, kuten valaistusvaihteluita, jotka voivat vaikeuttaa tunnistusta. Tämän vuoksi mallien suorituskyky todellisissa sovelluksissa voi olla alhaisempi kuin tutkimusympäristössä raportoidut tulokset.

Aiemmin mainitussa tutkimuksessa, jossa ihmisten kykyä tunnistaa syväväärennöksiä testattiin, tunnistustarkkuus jäi alle 70 prosentin [7]. Ottaen tämä huomioon voidaan todeta, että koneoppimisen mallit ovat tällä hetkellä selvästi ihmisiä luotettavampia ratkaisuja syväväärennösten tunnistuksessa. Tämä korostaa automaattisten tunnistusmenetelmien merkitystä erityisesti tilanteissa, joissa käsitellään suuria määriä videomateriaalia, kuten sosiaalisessa mediassa, verkkopalveluissa ja digitaalisissa tunnistusjärjestelmissä.

Voidaan todeta, että kaikki tarkastellut mallit osoittautuivat toimiviksi ratkaisuiksi syväväärennösten tunnistuksessa, vaikka niiden vahvuudet ja käyttötarkoitukset erosivat toisistaan. Tulokset antavat yleiskuvan mallien suorituskyvystä, mutta osoittavat tunnistusmenetelmien kehittämisen olevan välttämätöntä väärennöstekniikoiden kehittyessä

5 Yhteenveto

Syvävääreennösten tunnistuksessa voidaan hyödyntää monia erilaisia poikkeamia ja saavuttaa korkea tunnistustarkkuus. Tässä tutkielmassa tarkasteltiin neljää erilaista artikkelia ja niissä esiteltyjä koneoppimisen malleja. Tutkielman tutkimuskysymykset ovat: TK1. Mitä poikkeamia voidaan hyödyntää syvävääreennösten tunnistuksessa?, TK2. Kuinka tarkkoja koneoppimisen mallit ovat syvävääreennösten tunnistuksessa?. Tarkastelun pohjalta ei voida todeta yhtä ratkaisua yksiselitteisesti parhaaksi. Jokainen malli erosi toiminnaltaan ja painotti erilaisia poikkeamia tunnistuksessa. Mallit antavat kuitenkin hyvän yleiskuvan mallien tarkkuudesta tunnistamistehtävässä.

Tutkimuskysymyksen TK1 tarkoitus oli tarkastella millaisia poikkeamia voidaan hyödyntää syvävääreennösten tunnistuksessa. Tarkasteltujen artikkelien perusteella voidaan todeta, että syvävääreennökset sisältävät useita erilaisia poikkeamia, joita koneoppimisen mallit voivat hyödyntää tunnistuksessa.

Yleisin ja helpoiten havaittavissa oleva poikkeamatyyppi olivat tilalliset poikkeamat, jotka olivat havaittavissa yksittäisissä kuvakehyksissä. Nämä poikkeamat esiintyivät esimerkiksi ihon tekstuurivirheinä, kasvojen rakenteellisina vääristyminä tai valaistuksen epäjohdonmukaisuuksina.

Videopohjaisessa tunnistuksessa oli mahdollista hyödyntää tilallisten poikkeamien lisäksi ajallisia poikkeamia, jotka ilmenivät kokonaisissa videoissa. Nämä poikkeamat esiintyivät esimerkiksi epäjohdonmukaisuuksina kasvojen liikkeissä tai ilmei-

den rytmissä.

Viimeinen tarkasteltu poikkeama oli biometriset poikkeamat, eli henkilön kasvojen yksilöllisiin ja kolmiulotteisiin rakenteisiin liittyvä poikkeama. Biometrisiä poikkeamia hyödynnetään identiteettiin perustuvassa tunnistuksessa, mikä edellyttää henkilön kasvojen kolmiulotteista mallinnusta.

Tutkimuskysymyksen TK2 tarkoitus oli antaa yleiskuva koneoppimisen mallien tarkkuudesta syvävääreennösten tunnistuksessa. Tarkasteltujen artikkelien ja mallien perusteella voidaan todeta, että koneoppimisen mallit kykenevät tunnistamaan syvävääreennöksiä suhteellisen korkealla tarkkuudella verrattuna ihmisen suorituskykyyn.

Jokainen tarkasteltu malli saavutti yli 90% tunnistustarkkuuden, vaikka ne hyödynsivät erilaisia arkkitehtuureja ja aineistoja. Korkein saavutettu tunnistustarkkuus oli 96,2%, jonka saavutti CNN-LSTM-Transformer-malli. Tarkkuuksia ei voida kuitenkaan vertailla keskenään ja valita yksiselitteisesti parasta mallia, sillä tarkastellut mallit hyödynsivät erilaisia aineistoja ja tunnistustapoja. Tulosten perusteella voidaan vastata tutkimuskysymykseen TK2 siten, että koneoppimisen mallit kykenevät tunnistamaan syvävääreennöksiä korkealla tarkkuudella valituilla aineistoilla. Tässä tutkielmassa tarkastellut mallit saavuttivat yli 90 % tunnistustarkkuuden, mikä on selvästi enemmän kuin ihmisten vastaavissa kokeissa saavuttama tarkkuus.

Tutkielman perusteella voidaan todeta, että syvävääreennösten tunnistuksessa voidaan saavuttaa korkea tarkkuus hyödyntämällä koneoppimisen malleja ja erilaisia poikkeamatyyppejä. Koneoppimisen tunnistusmenetelmät ovat tällä hetkellä selvästi ihmisiä luotettavampia. On kuitenkin huomioitava, että syvävääreennösteknologian kehittyessä tunnistettavat poikkeamat voivat vähentyä tai muuttua vaikeammin havaittavaksi. Tunnistusmenetelmien jatkuva kehitys ja mukautuminen ovat edellytyksiä sille, että tunnistustarkkuus säilyy korkeana syvävääreennöstekniikoiden kehittyessä.

Lähdeluettelo

[1] L. Howell et al., "Digital wildfires in a hyperconnected world", *WEF report*, vol. 3, nro 2013, s. 15–94, 2013, Viitattu 12.3.2026. url: https://www3.weforum.org/docs/WEF_GlobalRisks_Report_2013.pdf.

[2] M. Del Vicario et al., "The spreading of misinformation online", *Proceedings of the national academy of Sciences*, vol. 113, nro 3, s. 554–559, 2016. DOI: 10.1073/pnas.1517441113.

[3] Z. Adams, M. Osman, C. Bechlivanidis ja B. Meder, "(Why) is misinformation a problem?", *Perspectives on Psychological Science*, vol. 18, nro 6, s. 1436–1463, 2023. DOI: 10.1177/17456916221141344.

[4] M. Westerlund, "The emergence of deepfake technology: A review", *Technology innovation management review*, vol. 9, nro 11, s. 39–52, 2019. DOI: 10.22215/timreview/1282.

[5] P. Singh ja D. B. Dhiman, "Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century", *Qeios*, 2023, Preprint. Viitattu 12.3.2026. url: https://www.qeios.com/read/DPLE2L.

[6] "Sumsub Identity Fraud Report 2024", Sumsub, tekninen raportti 133, 2024, Viitattu 12.3.2026. url: https://sumsub.com/fraud-report-2024/.

[7] M. Groh, Z. Epstein, C. Firestone ja R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds", *Proceedings of the National*

Academy of Sciences, vol. 119, nro 1, e2110013119, 2022. DOI: 10.1073/pnas.2110013119.

[8] Sumsub, "Identity Fraud Report", Sum ja Substance Ltd (UK), 2023, Viitattu 12.3.2026, s. 65. url: https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/.

[9] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov ja A. S. Smirnov, "Methods of deepfake detection based on machine learning", teoksessa *2020 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus)*, Viitattu 12.3.2026, IEEE, 2020, s. 408–411. url: https://ieeexplore.ieee.org/abstract/document/9039057.

[10] L. Deng, H. Suo ja D. Li, "Deepfake Video Detection Based on EfficientNet-V2 Network", *Computational Intelligence and Neuroscience*, vol. 2022, nro 1, s. 3 441 549, 2022. DOI: 10.1155/2022/3441549.

[11] P. Joshi ja V. Nivethitha, "Deep Fake Image Detection using Xception Architecture", teoksessa *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, Viitattu 12.3.2026, 2024, s. 533–537. url: https://ieeexplore.ieee.org/abstract/document/10578398.

[12] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. Almaloglou ja D. Zarpalas, "Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification", *Multimedia Tools and Applications*, s. 1–20, 2025, Viitattu 12.3.2026. url: https://link.springer.com/article/10.1007/s11042-024-20548-6.

[13] D. Wodajo, P. Lambert, G. V. Wallendael, S. Atnafu ja H. Mareen, "Improved Deepfake Video Detection Using Convolutional Vision Transformer", teoksessa *Proceedings of the 2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024. DOI: 10.1109/gem61861.2024.10585593.