Syvävääärennösvideoiden tunnistaminen konvoluutioneuroverkkojen avulla

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tieto- ja viestintätekniikka
Toukokuu 2024
Hanna Leppänen

TURUN YLIOPISTO
Tietotekniikan laitos

Hanna Leppänen: Syvävääärennösvideoiden tunnistaminen konvoluutioneuroverkkojen avulla

TkK-tutkielma, 29 s.
Tieto- ja viestintätekniikka
Toukokuu 2024

Syvävääärennösten ja niiden tuottamiseen käytettyjen teknologioiden nopea kehittyminen on lisännyt tunnistusmenetelmien tarvetta. Etenkin rikollisissa käyttötarkoituksissa syvävääärennökset aiheuttavat konkreettisia uhkia. Syvävääärennösten tuottamiseen suosituimpia teknologioita ovat GAN-verkot, enkoodaus-dekoodaus-verkot ja diffuusiomallit. Tämän tutkielman tavoitteena on selvittää mihin syvävääärennösvideoiden tunnistus voi perustua ja millaiset tunnistusmenetelmät ovat merkittävimpiä. Tutkielmassa lisäksi arvioidaan ja vertaillaan merkittävimmiksi todettuja menetelmiä ja tulevaisuuden näkymiä.

Syvävääärennösten tunnistusmenetelmät rakentuvat pidempään tutkimuskohteena olleen kuvantunnistuksen pohjalle. Kuvien ja objektien tunnistuksessa tehokkaimmaksi on osoittautunut konvoluutioneuroverkko. Suurin osa syvävääärennösten tunnistusmenetelmistä perustuu näihin konvoluutioverkkoihin. Syvävääärennösten tunnistuksessa hyödynnetään myös takaisinkytkettyjä neuroverkkoja ja niiden sovelluksia. Tässä tutkielmassa käsitellään tarkemmin kolmea merkittävää tunnistusmenetelmää.

Tunnistusmenetelmien erot liittyvät niiden saavuttamiin numeerisiin tuloksiin, rakenteiseen ja käytettyihin tietoaineistoihin. Etenkin tietoaineiston valinta vaikuttaa vahvasti saatuihin testaustuloksiin. Tietoaineistot voidaan nähdä tunnistusmenetelmien rajoittavana tekijänä. Tunnistusmenetelmien suurimpia haasteita ovat menetelmien käytettävyys sekä tulosten yleistettävyys ja ennustettavuus. Tulevaisuudessa menetelmiä tulee kehittää näiden haasteiden ratkaisemiseksi.

Asiasanat: syvävääärennösvideo, konvoluutioneuroverkko, tietoaineisto, kuvantunnistus, tunnistusmenetelmä, yleistettävyys, ennustettavuus

Sisällys

1 Johdanto

Tekoälyllä tuotetut syväväärennökset (engl. deepfake) ovat yleistyneet disinformaation levittämisen, poliittisen vaikuttamisen, rikollisuuden ja huumorin välineinä. Syväväärennös on generatiivisella tekoälyllä tuotettu kuva, teksti, ääni tai video. Niiden tuottamiseen on kehitetty useita sovelluksia, joista monet ovat kuluttajien käytössä. Syväväärennöksiin liittyviä tieteellisiä julkaisuja on julkaistu viime vuosien aikana jopa tuhansia vuodessa [1]. Tuotettuja väärennöksiä voidaan käyttää moniin eri tarkoituksiin, mutta erityisesti vahingoittavat käyttökohteet huolestuttavat tutkijoita [1] [2]. Tekoälyn sovelluksista kuluttajia pelottavat eniten juuri harhaanjohtavat syväväärennökset [1].

Syväväärennökset huolestuttavat myös Microsoftin varapuheenjohtajaa Brad Smithiä. Hänen mukaansa syväväärennökset ovat tekoälyn huolestuttavin käyttökohde [2]. Syväväärennöksistä etenkin videot leviävät lähes huomaamattomasti sosiaalisessa mediassa muiden videoiden joukossa. Videoissa voidaan väärentää esimerkiksi poliittisten johtajien tai muiden julkisuuden henkilöiden sanoja tai tekoja. Syväväärennössovelluksella saadaan edesmennyt Kuningatar Elisabet tanssimaan pöydällä[3] tai laulaja Taylor Swiftin kasvot pornografisiin kuviin. Taylor Swiftin tapaus sai myös poliittisten johtajien huomion ja korosti lainsäädännön tarvetta, jolla krimi-

[1] https://today.yougov.com/technology/articles/46058-majorities-americans-are-concerned-about-spread-ai, luettu 11.3.2024

[2] https://www.theguardian.com/technology/2023/may/25/deepfakes-ai-concern-microsoft-brad-smith, luettu 11.3.2024

[3] https://yle.fi/a/3-11716366, luettu 13.3.2024

nalisoitaisiin vahingoittavien syväväärennösten luominen ja levittäminen [4]. Lakien säätäminen ei pysy syväväärennösteknologioiden kehityksen mukana, joten on tärkeää tarjota kuluttajille, yrityksille ja organisaatioille työkaluja syväväärennösten tunnistamiseen.

Ihmisten heikko kyky tunnistaa etenkin hyvin tuotettuja syväväärennösvideoita on myös yksi argumentti laadukkaiden tunnistusmenetelmien puolesta. Tutkittaessa ihmisten kykyä tunnistaa väärennetyt videot huomattiin selviä eroja videoiden laadun perusteella. Mitä laadukkaampi syväväärennösvideo on, sitä heikommin ihmiset tunnistavat sen väärennökseksi. Yli 75 prosenttia vastaajista luuli hyvälaatuista syväväärennösvideota aidoksi. Kiinnostava näkökulma on ihmisen ja koneen tunnistuskyvyn erot. Koneen ”katsetta” ei voi verratta ihmisen katseeseen. Ihmiselle helposti väärennökseksi tunnistettava video ei koneelle välttämättä ole sitä, vaan jopa päinvastoin. [3] Tutkimustilanteessa koehenkilöt tietävät arvioivansa videoiden aitoutta ja keskittyvät siihen. Yleensä syväväärennösvideoita kohdataan kuitenkin niitä odottamatta, jolloin tunnistuksen voidaan olettaa olevan vielä heikompaa kuin tutkimustilanteessa. Hyvälaatuisten syväväärennösvideoiden tunnistamiseen tarvitaan siis työkaluja.

Tämän kirjallisuuskatsauksen alussa on tavoitteena laajasti määritellä mitä syväväärennökset ovat ja miten niitä tuotetaan. Ennen tunnistusmenetelmien tarkempaa käsittelyä käydään läpi niiden teknisiä edellytyksiä. Tunnistusmenetelmien yleiskatsauksen jälkeen esitellään kolme erilaista tunnistusmenetelmää, niiden toiminta ja rajoitteet. Lopuksi menetelmiä vertaillaan toisiinsa ja arvioidaan tunnistusmenetelmien tulevaisuuden näkymiä. Tutkielmani tutkimuskysymykset ovat:

1. Mihin syväväärennösvideoiden tunnistus voi perustua?
2. Mitkä ovat merkittävimpien tunnistusmenetelmien toimintaperiaatteet ja miten ne eroavat toisistaan?

[4] https://www.politico.eu/article/europe-eye-fix-taylor-swift-nude-deepfake/, luettu 13.3.2024

3. Miten menetelmiä tulee kehittää tulevaisuudessa?

Tämän kirjallisuuskatsauksen aineistohaku on suoritettu systemaattisena aineistohakuna IEEE tietokannasta. Hain aineistoa ensin laajemmin kaikista tekoälyyn perustuvista syväväärennöstunnistusmenetelmistä. Aihe osoittautui nopeasti liian laajaksi, joten rajasin hakua kattamaan vain syväväärennösvideoiden tunnistukseen käytetyt metodit. Tämä rajaus ei kuitenkaan ollut riittävä, koska videoiden tunnistukseen käytetään laajasti eri tekijöihin perustuvia menetelmiä. Rajasin aineistohakua edelleen kattamaan vain konvoluutioneuroverkkojen käytön syväväärennösvideoiden tunnistuksessa.

Lopullinen käyttämäni hakulause oli *("deepfake video" detection OR "deepfake video" recognition) AND (convolutional neural network OR cnn). CNN on lyhenne sanalle konvoluutioneuroverkko (engl. convolutional neural network)*. Haku on toteutettu englannin kielellä. Hakutuloksia oli yhteensä IEEE:ssä 54, joista valitsemaani aiheeseen liittyi otsikoiden perusteella 25 aineistoa. Aiheeseen liittyvistä aineistoista valitsin tarkempaan käsittelyyn 16 aineistoa niiden abstraktien perusteella. Tiedonhaun viimeisessä vaiheessa luin artikkelit kokonaisuudessaan ja sisällytin tähän kirjallisuuskatsaukseen 5 keskeisintä aineistoa niiden sisällön, kielellisen laadun ja ajankohtaisuuden perusteella. Näiden aineistojen lisäksi olen hyödyntänyt tutkielman tukena myös muuta aineistoa IEEE:stä, Volterista ja Google Scholarista.

2 Taustaa

Syväväärennös käsitteenä esiintyi ensimmäisen kerran vuoden 2017 lopussa, kun Reddit-sivustolla käyttäjänimellä deepfakes esiintyvä käyttäjä julkaisi syväoppimisella väärennettyjä pornografisia videoita, joihin oli liitetty julkisuuden henkilöiden kasvoja. Kyseinen käyttäjä kehitti FakeApp-sovelluksen, jolla seuraavana vuonna Buzzfeed teki väärennösvideon Barack Obamasta. Mirsky ja Lee määrittelevät syväväärennöksen "syväoppimisella tuotetuksi uskottavaksi mediaksi". He korostavat artikkelissaan, että syväväärennösten tarkoitus on nimenomaan huijata ihmistä eikä konetta, kuten joillain tekoälyllä tuotetuilla haittaohjelmilla on tavoitteena. [4] Masood ja muut määrittelevät syväväärennöksen synteettiseksi tekoälyllä tuotetuksi videoksi tai kuvaksi [5]. Syväväärennökselle käsitteenä ei ole vakiintunutta määritelmää, mutta valtaosa määritelmistä ovat samankaltaisia. Englannin käsite deepfake edustaa myös tiettyä tapaa tuottaa syväväärennös, mutta sitä käytetään yleisemmin kuvaamaan kaikkia eri menetelmillä tuotettuja syväväärennöksiä [6].

Syväväärennösvideoiden luontiin voidaan käyttää monia eri menetelmiä. Masood ja muut jakavat syväväärennösvideot viiteen kategoriaan niiden tuotantotavan perusteella. Seuraavissa lähdehenkilö tarkoittaa videossa alun perin esiintyvää henkilöä ja kohdehenkilö ihmistä, jonka kasvot tai puhe videoon halutaan liittää tai kenen videossa halutaan esiintyvän. Syväväärennösvideo voidaan luoda:

1. vaihtamalla lähdehenkilön kasvot kohdehenkilön kasvoiksi,
2. muuttamalla kohdehenkilön huulten liikkeet tiettyä ääniraitaa vastaavaksi,

3. matkimalla kohdehenkilön eleitä kuten silmien liikkeitä, kasvojen ilmeitä tai pään asentoja,
4. teettämällä kokonaan uusia kasvokuvia kohdehenkilöstä,
5. muokkaamalla kohdehenkilön kasvojen tiettyjä ominaispiirteitä. [5]

Mubarak ja muut käyttävät samaa yllä näkyvää luokittelua [1]. Myös Mirsky ja Lee luokittelevat videot lähes vastaavalla tavalla. He pitäytyvät kolmessa kategoriassa: kohdehenkilön eleiden tai kehonkielen matkiminen, kasvojen tai niiden osan korvaaminen ja kasvojen piirteiden muokkaus tai kokonaan uusien kasvojen luominen. Kun syvävääreennöksiä luodaan halvoilla ja helpommin saavutettavilla ohjelmilla, niitä kutsutaan kevyiksi tai halvoiksi väärennöksiksi. [4] Tässä tutkielmassa ei käsitellä kevyitä väärennöksiä tätä mainintaa enemmän. Seuraavaksi esitellään syvävääreennösten taustalla olevia teknologioita.

2.1 Syvävääreennösten tekninen pohja

Syvävääreennösvideoita luodaan monilla neuroverkkojen sovelluksilla [4]. Syväoppiminen perustuu neuroverkkoihin. Syväoppimiselle on ominaista, että koneelle annetaan mahdollisuus oppia konsepteja niiden hierarkioiden kautta. Kone oppii tunnistamaan haastaviakin konsepteja pilkkomalla sen pienempiin yksinkertaisiin osiin.[7]

Neuroverkot rakentuvat neuroneista ja niitä yhdistävistä synapseista ihmisen aivojen tavoin. Neuronit muodostavat verkon kerrokset, joiden määrä voi vaihdella. Synapseilla on neuroverkossa painot, joita säädetään verkon koulutuksen aikana. Neuroni laskee synapsin painon perusteella painotetun keskiarvon vastaanottamistaan signaaleista ja keskiarvon ylittäessä tietyn kynnysarvon, neuroni aktivoituu ja lähettää signaalin eteenpäin. Näin toimii yksinkertaisuudessaan TLU-neuroni (engl. threshold logic unit). Kun neuroverkko koulutetaan tietoaineistolla (engl. dataset), joka sisältää jokaiselle sisääntulevalle signaalille toivotun ulostulevan signaalin, on

kyse ohjatusta oppimisesta. Neuroverkon antamaa ulostulevaa signaalia verrataan tavoitteeseen ja tarvittavien synapsien painot muokataan. Sama toistetaan eri signaalipareille ja painoja sekä aktivaation kynnysarvoja muokataan, kunnes verkko oppii tuottamaan halutulla todennäköisyydellä oikean ulostulevan signaalin. Kun neuroverkon koulutus on valmis, se osaa optimaalisessa tilanteessa luoda myös täysin uusille signaaleille tavoitteiden mukaisen tuloksen. [8]

GAN-verkko käsitteenä tulee englannin kielen sanoista generative adversarial network, joka voidaan suomentaa generatiiviseksi kilpailevaksi verkoksi. GAN-verkon toiminta perustuu kahteen erilliseen verkkoon, generoivaan ja luokittelevaan, joilla se koulutetaan tuottamaan realistisia kuvia. Generoiva verkko luo väärennöksiä, joilla se pyrkii huijaamaan luokittelevaa verkkoa, joka pyrkii erottelemaan väärennökset aidoista videoista. Verkon tavoitteena on, että generoiva verkko oppii tuottamaan niin laadukkaita väärennöksiä, ettei luokitteleva verkko enää erota väärennöksiä aidoista videoista. Kun tämä tavoite on saavutettu, luokitteleva verkko poistetaan ja generoivalla verkolla jatketaan väärennösten luomista. GAN-verkoista on edelleen kehitetty juuri syvävääreennösten luontiin tehokkaampia sovelluksia kuten pix2pix ja CycleGan. [4] Tällä hetkellä suosituimpia neuroverkkoja syvävääreennösten tuottamiseen ovat GAN-verkot ja seuraavaksi käsiteltävät ED-verkot [4] [1].

Enkoodaus-dekoodaus (ED) verkot sisältävät vähintään kaksi kerrosta, enkooderin ja dekooderin [4]. Enkooderin tehtävä on oppia esittämään verkkoon sisääntuleva data matalaulotteisemmalla abstraktiotasolla. Dekooderin tehtävä on päinvastainen eli jälleenrakentaa näistä matalan tason abstraktioista uloslähtevä data. VAE eli variational autoencoder on enkoodaus-dekoodaus-verkon eritysmuoto, joka sisältää todennäköisyysmallinnusta, jonka ansiosta verkko osaa myös luoda uutta dataa. [1] VAE-verkkoja hyödynnetään syvävääreennösten luomisessa, koska niiden latentti avaruus on erotellumpi, jolloin enkoodaus reagoi paremmin interpolointiin ja muutoksiin [4]. Latentti avaruus tarkoittaa avaruutta, jossa datapisteet on jaotel-

tu niiden samankaltaisuuden mukaan tietyille alueille. Erottelevuus viittaa näiden datapisteiden jaotteluun ja etäisyyksiin toisistaan.[1]

Diffuusiomallit (DM) ovat myös laajasti käytössä syväväärennösten luomisessa. Diffuusiomallit lisäävät kohinaa kuviin ja oppivat poistamaan tätä kohinaa myöhemmin. Kohinaa lisätään ja poistetaan kuvista asteittain. Kun malli on koulutettu se osaa tuottaa ohjausjärjestelmän avulla pelkästä kohinasta korkealaatuisia aidon näköisiä kuvia. Diffuusiomallien on huomattu suoriutuvan jopa GAN-verkkoja paremmin tietyissä käyttötarkoituksissa. [1] Saman ovat huomanneet myös ChatGPT:n kehittäneen OpenAI:n tutkijat. GAN-verkkojen heikkoutena voidaan pitää niiden huonoa skaalautuvuutta laajempiin uusiin aihealueisiin. Tämä on seurausta GAN-verkkojen oppimisen laajoista vaatimuksista. GAN-verkkojen on huomattu myös tuottavan vähemmän monimuotoisia kuvia verrattuna todennäköisyyteen perustuviin malleihin kuten diffuusiomalleihin. Diffuusiomallien on todettu jo tuottavan laadukkaampia kuvia esim. CIFAR-10 tietoaineistolla, mutta ImageNet tietoaineistoon perustuvia kuvia GAN-verkot tuottavat laadukkaammin. [9] Tietoaineistojen vaikutuksia syväväärennöksiin ja tunnistusmenetelmiin käsitellään seuraavaksi.

2.2 Tietoaineistojen merkitys

Tietoaineistot (engl. dataset) ovat laajoja kokoelmia, joilla koneoppimisen malleja koulutetaan ja testataan. Syväväärennöksiä varten on kehitetty useita kuva- ja videokokoelmia, jotka sisältävät aitoja ja väärennettyjä kuvia tai videoita. Kehitettyjen tietoaineistojen sisällöt vaihtelevat laajasti ja aineistojen laadun arviointi onkin tärkeä osa koneoppimismallien kehitysprosessia. [2] [10] Tietoaineistot on usein edelleen jaettu koulutus-, arviointi- ja testianeistoihin. Jos tietoaineiston jako on tehty huolimattomasti, on vaarana polarisaatio ja aineiston spesifien ominaisuuksien liiallinen korostuminen koulutetussa mallissa. Tietoaineistojen suuresta määrästä ja

[1] https://www.baeldung.com/cs/dl-latent-space, luettu 15.5.2024

niiden keskinäisestä erilaisuudesta seuraa se, että eri tietoaineistolla yksittäinen tunnistusmenetelmä voi saavuttaa laajasti vaihtelevia tuloksia. [2]

Videoista koostuvien tietoaineistojen kehitys on kiihtynyt viime vuosien aikana. Väärennettyjen videoiden tuottaminen perinteisillä muokkausmenetelmillä on aikaavievää, joten suurimmat tietoaineistot on tuotettu hyödyntäen tekoälyn sovelluksia syvävääreennösten luomisessa. [2] Tietoaineistojen kehityksen trendi on ollut luoda laajempia aineistoja yhä laajempien koneoppimismallien kehitykseen ja näin saavuttaa merkityksellisiä virstanpylväitä alan kehityksessä. Ongelmia on noussut esiin tietoaineistojen puolueellisuuteen, aineistojen tosielämän vastaavuuteen ja kerätyn datan dokumentaation heikkouteen liittyen. Lisäksi ongelmaksi on osoittautunut mallien kyky oppia ratkaisemaan kyseessä oleva tehtävä niin sanottuun yleiseen heuristiikkaan eikä aineistoon perustuen. Tietoaineistojen kentän tulisi laajentua huomioimaan niiden arvioinnissa myös muut mittarit virstanpylväiden ohella ja tiedon keräämiseen liittyvät sekä eettiset että juridiset näkökulmat. [10]

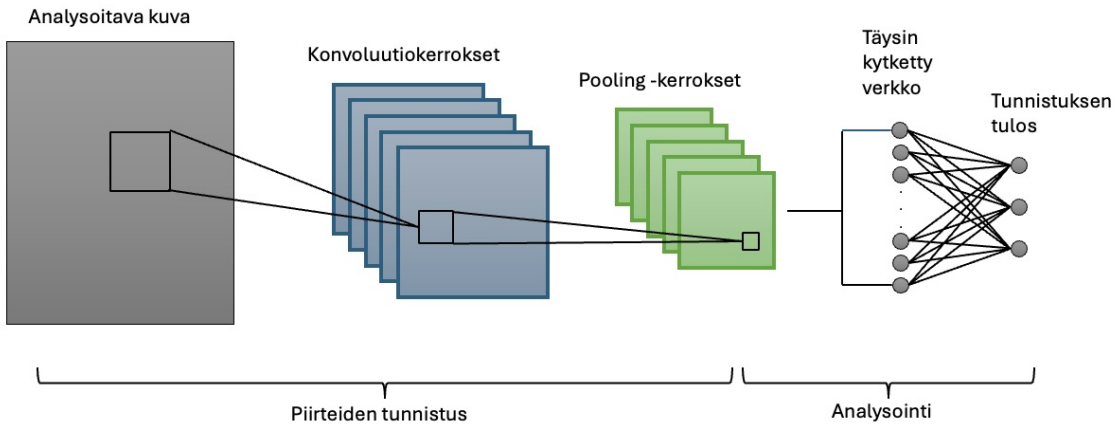
Tietoaineistojen koot vaihtelevat laajasti. FaceForensics++ sisältää tuhat Youtubesta kerättyä alkuperäistä videota, joista neljällä eri menetelmällä on tuotettu 4000 syvävääreennösvideota. Faceforensics++ on jatkokehitetty 2018 julkaistusta Faceforensics-tietoaineistosta. [11] Kokonsa puolesta pienempi, mutta myös laajasti käytetty tietoaineisto on DeepfakeTIMIT. Tietoaineisto on koottu muokkaamalla GAN-pohjaisilla menetelmillä VidTIMIT-tietoaineiston videoita. Tietoaineistoon on tuotettu matala- ja korkealaatuisia videoita, 320 kappaletta molempia. Lopullisessa tietoaineistossa erilaatuiset videot on pidetty erillään, jotta mallien tunnistuskyvystä saadaan tietoa molemmissa skenaarioissa. [12] Tietoaineistot vaihtelevat siis koon, jaottelun ja syvävääreennösten tekotavan perusteella. Paullada ja muut arvioivat artikkelissaan, että tietoaineistot voidaan nähdä koneoppimisalgoritmien kehityksen rajoittavana tekijänä [10]. Syvävääreennösten ja niiden tunnistusmenetelmien kehittyminen nykyiseen tilaansa on ollut mahdollista osittain tietoaineistojen ansiosta.

2.3 Konvoluutioneuroverkot kuvantunnistuksessa

Tietoaineistojen ohella toinen tekijä, joka on mahdollistanut erityisesti syvävääränösten tunnistusmenetelmien kehittymisen, on kuvantunnistusmenetelmät ja niiden saavuttamat tulokset. Kuvantunnistuksessa laajasti käytetty menetelmä on konvoluutioneuroverkko. Konvoluutioneuroverkon alkeellista muotoa käytettiin ensimmäisen kerran 1980-luvulla käsin kirjoitettujen numeroiden tunnistuksessa. Tämän jälkeen verkoissa hyödynnettiin vastavirta-algoritmia, joka mahdollisti mallien kouluttamisen ilman vaativaa esikäsittelyä [13] [14]. Konvoluutioverkon kerrosten kasvattamisen myötä mallin ylisovittuminen, paikallinen optimi ja häviävä gradientti muodostuivat ongelmiksi, jotka johtivat konvoluutioverkkojen kehityksen laantumiseen. Kuitenkin piilokerroksien ja normalisoinnin hyödyntämisen avulla ongelmia saatiin minimoitua ja konvoluutioneuroverkkojen kehitys jatkui. [13]

Konvoluutioneuroverkkoja käytetään laajasti objektien ja alueiden tunnistamiseen ja erottamiseen kuvista. Kasvojen tunnistaminen on kehittynyt merkittävästi konvoluutioverkkojen kehityksen myötä. Nykyään konvoluutioneuroverkot ovat dominoiva tekniikka lähes kaikissa tunnistus- ja havainnointitehtävissä. [14] Konvoluutioneuroverkkojen toiminta perustuu kerroksiin, joilla kuva ”skannataan” läpi tietyn kokoisina paloina. Kerrokset muodostavat piirrekarttoja, joiden avulla kuvan sisältöä luokitellaan ja analysoidaan. [4] Piirrekartoille suoritettava suodatusopearaatio on matemaattisesti diskreetti konvoluutio, johon verkon nimikin viittaa [14]. Alla on yksinkertaistettu kuva konvoluutioverkon rakenteesta. Konvoluutio- ja poolingkerroksia esiintyy yleensä peräkkäin useampia.

Konvoluutioverkkojen ohella LSTM-verkoilla on myös saavutettu hyviä tuloksia etenkin ajallisiin riippuvuuksiin liittyvissä tehtävissä. LSTM-verkko on takaisinkytketyn neuroverkon yksi erityismuoto. Verkon nimi tulee englannin kielen sanoista Long Short Term Memory, joka viittaa verkon kykyyn oppia tunnistamaan ajallisia piirteitä. Konvoluutioverkon ja LSTM-verkon yhdistäminen on tuottanut hyviä



Kuva 2.1: Konvoluutioverkon rakenne

tuloksia videoihin liittyvissä tunnistustehtävissä. [15]

Tunnettuja konvoluutioneuroverkkoihin perustuvia kuvantunnistusverkkoja ovat VGG-, Resnet- ja Inception-verkkojen eri versiot [13]. Simonyan ja Zisserman kehittivät VGG-verkon, jonka tärkein innovaatio on 3 x 3 konvoluutiomatriisin käyttö tunnistuksen tehostamiseksi. Kehittäjät voittivat verkkoarkkitehtuurillaan ImageNet haasteen vuonna 2014. [16] Inception-verkosta on kehitetty kolme eri versiota V1, V2 ja V3 [13]. Inception V1 -arkkitehtuuri esitteli 1 x 1 konvoluutiomatriisin hyödyntämisen parametrien ja piirrekarttojen määrän minimointiin [17]. Inception V2 -arkkitehtuuri lisäsi aiempaan malliin normalisointia ja lisäsi verkon syvyyttä muuttamalla konvoluutiomatriisin kokoa. Verkkoarkkitehtuurin kolmas versio Inception V3 esitteli n x n matriisin korvaamisen 1 x n ja n x 1 matriiseilla. Kehittäjien mukaan arvon n kasvaessa, mallin resurssivaativuus pienenee huomattavasti. [18] Resnet-arkkitehtuuri pyrkii poistamaan verkon syventämisestä aiheutuvia ongelmia. Arkkitehtuurin tavoitteena on, että pelkästään syventämällä verkkoa, sen suoritusta voidaan parantaa. [19] Syväväärennösten tunnistusmenetelmät perustuvat vahvasti aiempiin hyviin tuloksiin, joita konvoluutioneuroverkoilla on saavutettu objektien ja etenkin kasvojen tunnistuksessa. Seuraavassa kappaleessa käsitellään tarkemmin syväväärennösten tunnistusmenetelmiä ja niiden toimintaperiaatteita.

3 Tunnistusmenetelmät

Tunnistusmenetelmiä on kehitetty paljon ja ne pohjautuvat laajasti erilaisiin visuaalisiin virheisiin ja teknologioihin. Syväväärennösvideoiden tunnistusmenetelmät voidaan erotella visuaalisiin tekijöihin perustuviin ja syväoppimiseen perustuviin menetelmiin [1] [2]. Visuaalisiin tekijöihin perustuva tunnistaminen voi olla avaruudellisiin tai ajallisiin virheisiin perustuvaa. Avaruudellisia virheitä ovat kasvojen ja taustan rajan epätarkkuus ja niiden keskinäiset poikkeavuudet esimerkiksi valotuksessa sekä luomiseen käytettyjen teknologioiden jättämät merkit kuten GAN-verkkojen jättämä sormenjälki. [4]

Ajalliset virheet liittyvät kohdehenkilön käytökseen, fysiologisiin signaaleihin, synkronisaatioon ja johdonmukaisuuteen. Kohdehenkilön käytöstä voidaan verrata aitoon videoon ja näiden erojen perusteella tunnistaa video syväväärennökseksi. Fysiologisia signaaleja on esimerkiksi ihon alla näkyvä pulssi, jonka virheet paljastavat väärennöksen. Synkronisaatiolla tarkoitetaan videon äänen ja kuvan yhteensopivuutta. Suun ja huulten liikeitä tarkkailemalla voidaan tunnistaa virheitä synkronisaatiossa. Johdonmukaisuuteen liittyvät virheet esiintyvät videon vilkkumisena tai värinänä. [4]

Mirsky ja Lee ovat koonneet laajan taulukon, joka erittelee monipuolisesti tunnistusmenetelmiä niiden ominaisuuksien perusteella [4]. Mubarak ja muut esittelevät artikkelinsa taulukossa 17 merkittävää syväoppivaa tunnistusmenetelmää, jotka perustuvat laajasti konvoluutioneuroverkkoihin, takaisinkytkettyihin neuroverkkoihin,

LSTM-verkkoihin ja näiden sovelluksiin sekä yhdistelmiin. [1] Verdoliva erittelee artikkelissaan 12 syväoppimiseen perustuvaa menetelmää, joista monet ovat samoja kuin Mubarakin ja muiden artikkelissa [2]. Nguyen ja muut ovat keränneet taulukkoonsa 15 nimekästä videoiden tunnistusmenetelmää [20]. Suurin osa näissä artikkeleissa esiintyvistä tunnistusmenetelmistä ovat konvoluutioverkkoihin perustuvia.

Kaikkia kehitettyjä tunnistusmenetelmiä on mahdotonta käsitellä tämän tutkielman puitteissa, mutta merkittävimpinä voidaan pitää niitä, joihin on viitattu monissa alan julkaisuissa. Tällöin menetelmän merkityksellisyydestä ovat samaa mieltä monet alan tutkijat. Alla olevassa taulukossa 3.1 on esitelty kaikki menetelmät, jotka esiintyivät vähintään kahdessa edellä mainitussa katsausartikkelissa. Artikkeleissa esiteltiin yhteensä 53 eri tunnistusmenetelmää, joista 13 esiintyi vähintään kahdessa lähteessä. Taulukkoon on lisäksi koottu menetelmien tämän hetkiset viittausmäärät GoogleScholarissa. Yhden menetelmän artikkelia ei löytynyt GoogleScholarista, joten viittausmäärä on arXivista.

Taulukon menetelmistä Sabirin ja muiden [21] sekä Chinthan ja muiden [22] tunnistusmenetelmät perustuvat takaisinkytkettyihin neuroverkkoihin. Cozzolinon ja muiden kehittämä ForensicTransfer pyrkii ratkaisemaan luvussa 4.3 esiteltävän yleistettävyyden ongelman muokkaamalla piirrekarttoja [23]. Fernandon ja muiden menetelmä pyrkii samaan tavoitteeseen, mutta hyödyntämällä hierarkkisia muistiverkkoja [24]. Dangin ja muiden tunnistusmenetelmä perustuu myös piirrekarttojen muokkaamiseen [25] kuten Cozzolinon ja muiden menetelmä.

Taulukon menetelmistä kolme esittää uuden teknologian soveltamista tunnistuksen välineenä. Amerinin ja muiden menetelmä hyödyntää optista vuota (engl. optical flow) [26], Nguyenin ja muiden menetelmä soveltaa kapseliverkkoja (engl. capsule networks) [27], ja Ciftcin ja muiden patentoitu menetelmä hyödyntää biologisia signaaleja [28]. Lin ja muiden kehittämä tunnistusmenetelmä luokittelee aineiston aitoihin ja väärennettyihin kaikkiin syvävääreennöksiin tehtävään häivytykseen

Kehittäjä	**Julkaisu vuosi**	**Julkaisun nimi**	**Viittaukset Google Scholar**
Afchar et al.	2018	MesoNet: A compact facial video forgery detection network	1326
Cozzolino et al.	2018	ForensicTransfer: Weakly-supervised domain adaptation for forgery detection	276
Güera and Delp	2018	Deepfake video detection using recurrent neural networks	1074
Li and Lyu	2018	Exposing deepfake videos by detecting face warping artifacts	992
Amerini et al.	2019	Deepfake Video Detection through Optical Flow based CNN	364
Fernando et al.	2019	Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks	30
Nguyen et al.	2019	Capsule-forensics: Using Capsule networks to detect forged images and videos	592
Sabir et al.	2019	Recurrent convolutional strategies for face manipulation detection in videos	513
Chintha et al.	2020	Recurrent convolutional structures for audio spoof and video deepfake detection	144
Ciftci et al.	2020	FakeCatcher: Detection of synthetic portrait videos using biological signals	391
Li et al.	2020	Face X-ray for more general face forgery detection	760
Mittal et al.	2020	Emotions don't lie: A deepfake detection method using audio-visual affective cues	arXiv: 252
Dang et al.	2020	On the detection of digital face manipulation	488

Taulukko 3.1: Katsausartikkeleissa esiintyvät menetelmät

perustuen [29]. Mittalin ja muiden menetelmä on ensimmäinen laatuaan, joka yhdistää visuaalisia ja auditiivisia tekijöitä sekä havaittuja tunteita syvävääreennösten tunnistukseen [30].

Taulukon menetelmistä kahta on käsitelty kaikissa artikkeleissa. Näihin kahteen menetelmään on myös viitattu taulukon menetelmistä eniten. Nämä menetelmät ovat Afcharin ja muiden kehittämä Mesonet [31] sekä Gueran ja Delpin kehittämä menetelmä, joka perustuu takaisinkytkettyyn neuroverkkoon [15]. Nämä menetelmät käsitellään tarkemmin kappaleissa 3.1 ja 3.2. Lisäksi esittelen tarkemmin Lin ja Lyun kehittämän affiiniseen kasvonmuokkaamiseen perustuvan menetelmän [32], koska tämä edustaa merkittävien tunnistusmenetelmien joukosta sellaista, joka lähestyy tunnistamista tietoaineiston näkökulmasta. Kuten tässä tutkielmassa on tuotu esiin kappaleessa 2.2, tietoaineistojen kehittäminen on avainasemassa tunnistusmenetelmien kehityksessä. Menetelmään on myös viitattu eniten Afcharin ja muiden sekä Gueran ja Delpin menetelmän jälkeen (kts. taulukko 3.1). Nämä kolme seuraavaksi tarkemmin esiteltävää menetelmää voidaan viittausmäärien perusteella nähdä erityisen merkityksellisinä syvävääreennösten tunnistusmenetelmien laajassa kentässä. Menetelmien keskinäinen erilaisuus mahdollistaa vertailun useasta näkökulmasta.

3.1 MesoNet

Vuonna 2018 julkaistun Mesonetin tavoitteena on tunnistaa syvävääreennösvideoita niiden mesoskooppisten ominaisuuksien perusteella. Mesoskooppiset ominaisuudet tarkoittavat mikroskooppisen ja korkeimman tasoon väliin sijoittuvia ominaisuuksia. Mikroskooppisella tasolla tarkastelu ei tuota hyvää tulosta, koska videoiden kompressoiminen on poistanut niistä yksityiskohtaisia ominaisuuksia, joihin tunnistaminen perustuisi. Korkeimmalla tasolla tarkastelu on tarpeettoman laajaa. Mallin tavoitteena on toteuttaa neuroverkko pienellä määrällä kerroksia mahdollisimman tehokkaasti. [31]

Mesonetistä on kehitetty kaksi erilaista neuroverkkoarkkitehtuuria Meso-4 ja MesoInception-4. Meso-4-neuroverkon arkkitehtuurissa on ensin neljä peräkkäistä konvoluutiokerrosta, joiden tarkoituksena on erotella tarkasteltavasta videosta ominaispiirteitä. Nämä konvoluutiokerrokset hyödyntävät ReLU-funktioita (engl. Rectified Linear Unit), joiden tehtävä on lisätä epälineaarisuutta ja näin parantaa opitun datan kompleksisuutta. Konvoluutiokerrokset sisältävät lisäksi normalisointia ja pooling-kerrokset. Näiden neljän konvoluutiokerroksen jälkeen Meso-4-arkkitehtuurissa on kaksi täysin kytkettyä kerrosta ja Sigmoid-kerros. Täysin kytketyt kerrokset hyödyntävät Dropout-teknologiaa tulosten generalisointiin ja luokitteluun. Sigmoid-kerros asettaa saadun tuloksen välille [0,1]. [31]

MesoInception-4-neuroverkkoarkkitehtuuri perustuu Meso-4-arkkitehtuuriin, mutta sen kaksi ensimmäistä konvoluutiokerrosta on korvattu tunnetusta Inception-verkosta [17] muokatuilla kerroksilla. Optimaalinen hyöty Inception-verkosta saavutettiin korvaamalla kaksi verkon kerrosta, verrattuna useampaan. Arkkitehtuurin tavoitteena on yhdistää eri laajuisilla konvoluutioilla kerätty data yhteen ja näin optimoida tunnistusmenetelmän tehokkuus. Konvoluutiokerroksissa käytetään 3 x 3 konvoluutiomatriisia Meso-4-arkkitehtuurin 5 x 5 matriisin sijaan, jotta saavutetaan parempi tarkkuus tunnistuksessa. Lisäksi mallissa hyödynnetään 1 x 1 matriiseja edelleen sujuvoittamaan tunnistusta. Molemmat arkkitehtuurit sisältävät noin 28,000 koulutettavaa parametria eli neuronien välistä painoa. [31]

Meso-4- ja MesoInception-4-konvoluutioarkkitehtuureja on testattu DeepFake- ja Face2Face-teknologioilla luotujen videoiden tunnistuksessa. DeepFake-teknologialla kehitetyistä videoista, joissa henkilön kasvot on siis vaihdettu, tunnistusmenetelmän kehittäjät ovat itse koostaneet tietoaineiston, jolla mallit on koulutettu ja testattu. Menetelmien toimivuutta Face2Face-teknologialla tuotettujen videoiden tunnistamiseen on testattu FaceForensics-tietoaineistolla. FaceForensics on olemassa oleva tietoaineisto, joka sisältää Face2Face-teknologialla tuotettuja videoita, joissa henki-

lön identiteetti säilyy saman, mutta ilme kopioidaan toisesta videosta [33]. Meso-4-arkkitehtuuri tunnisti DeepFake-väärennökset ja aidot videot oikein 89,1 prosentin tarkkuudella ja MesoInception-4 91,7 prosentin tarkkuudella. FaceForensics-tietoaineistolla tuloksiksi saatiin Meso-4-arkkitehtuurilla 94,6 ja MesoInception-4-arkkitehtuurilla 96,8 prosenttia. [31]

3.2 Conv-LSTM

Güera ja muut kehittivät tunnistusmenetelmän, joka hyödyntää konvoluutioneuroverkkoa ominaispiirteiden louhintaan ja edelleen takaisinkytketyn neuroverkon erityismuotoa LSTM-verkkoa videon ruutujen aikajatkuvuuden määrittelyyn. Menetelmä pyrkii tunnistamaan faceswap-teknologialla luodut syväväärennökset niiden kolmeen heikkouteen perustuen. Faceswap-teknologia edustaa edellisessä luvussa listatuista tuotantomenetelmistä ensimmäistä eli syväväärennösvideota, jossa lähdehenkilön kasvot on vaihdettu kohdehenkilön kasvoiksi. [15]

Tällä menetelmällä tuotettujen syväväärennösvideoiden ensimmäinen heikkous liittyy koulutukseen ja tietoaineistojen sisältöön. Tietoaineiston kasvokuvat on otettu eri olosuhteista, jolloin kuvien valot ja varjot sekä kontrasti eriävät toisistaan. Kun näitä olosuhteiden puolesta toisistaan eriäviä kuvia käytetään videoiden luomiseen, videon ruutujen välille syntyy epäjatkuvuutta. Toinen heikkous liittyy kasvojen leikkaamiseen videoista ja kuvista. Kasvojen ääriviivojen tarkka erottaminen on haasteellista, jolloin tuotetuissa väärennösvideoissa kasvojen ääriviivat voivat näkyä epäselvinä tai sumuisina. Kolmas heikkous on faceswap-videoiden tekoon käytettävien autoenkooderien aiheuttama värinä videon ruutujen välille. Värinä on seurausta autoenkooderien heikosta ajallisesta tietoisuudesta ja on ihmissilmälle lähes näkymätöntä. Conv-LSTM menetelmä pyrkii tunnistamaan väärennökset näihin virheisiin perustuen. [15]

Tunnistusmenetelmä rakentuu konvoloidusta LSTM-arkkitehtuurista. Mallissa

on ensin konvoluutioneuroverkko, joka on vastuussa videon yksittäisten ruutujen tai kuvien ominaisuuksien louhinnasta. Verkko perustuu Inception V3 -verkkoon [18], jonka ensimmäinen kerros on poistettu, jotta se soveltuu ominaisuuksien louhintaan sen alkuperäisen tavoitteen eli kuvien luokittelun sijaan. Konvoluutioverkon jälkeen mallissa oleva LSTM tarkastelee konvoluutioverkon erittelemien ruutujen ominaisuuksien aikajatkuvuutta, jonka perusteella malli määrittelee, onko kyseessä aito video vai syväväärennösvideo. LSTM hyödyntää Dropout-teknologiaa samaan tarkoitukseen kuin Mesonetin arkkitehtuurit. Lopuksi mallissa on yksi täysin kytketty kerros ja softmax-kerros, joka laskee todennäköisyyden videon aitoudesta. [15]

Mallia on testattu kehittäjien itse koostamalla tietoaineistolla, johon on kerätty syväväärennösvideoita monilta eri alustoilta ja ne on liitetty osaksi valmiin HOHA-tietoaineiston elokuvaklippejä. HOHA:n ja kerättyjen videoiden yhdistäminen lisää tietoaineiston yleistettävyyttä ja näin edelleen parantaa tunnistuksen laajuutta. Menetelmän koulutukseen, validointiin ja testaukseen on käytetty kolmea eri ruutusekvenssipituutta N=20, 40 ja 80, koska väärennetyt kasvot esiintyvät videoissa vain lyhyen ajan. Käsiteltävien ruutujen määrä vaikuttaa tunnistukseen prosentuaalisesti vain vähän. Kun tarkasteltavia ruutuja oli 20, tunnistus onnistuu 96,7 prosentin varmuudella. Molemmilla suuremmilla ruutujen määrillä tunnistus onnistuu yhtä todennäköisesti eli 97,1 prosentin varmuudella. [15]

3.3 Affiiniseen väärentämiseen perustuva

Kolmas tarkemmin käsiteltävä tunnistusmenetelmä pyrkii tunnistamaan videot niiden tuotantomenetelmästä riippumatta. Tunnistusmenetelmä perustuu syväväärennösvideoihin tehtävään affiiniseen väärentämiseen, joka varmistaa, että väärennetyssä videossa kasvot vastaavat mittasuhteiltaan videossa alun perin esiintyneiden kasvojen mittasuhteita. Menetelmän kehittäjien mukaa sama ongelma pätee syväväärennösvideoihin niiden tuotantomenetelmästä huolimatta. [32] Menetelmän ta-

voite on täten olla riippumaton väärennöksen tekotavasta toisin kuin monet muut tunnistusmenetelmät, mukaan lukien aiemmin esitellyt Mesonet ja Conv-LSTM.

Menetelmä käyttää hyödyksi neljää tunnettua konvoluutioneuroverkkoa, jotka esiteltiin jo kappaleessa 2.3; VGG16 [19], ResNet50, ResNet101 ja ResNet152 [16]. Menetelmä perustuu näiden verkkojen kouluttamiseen kehittäjien koostamalla tietoaineistolla. Tietoaineiston negatiiviset näytteet eli syväväärennösten tapauksessa väärennökset on korvattu aidoilla kuvilla, joille on suoritettu syväväärennösvideoille tyypillinen affiininen väärentäminen. Itse neuroverkkoja käytetään menetelmässä sellaisenaan. [32]

Affiininen väärentäminen toteutetaan aitoihin kuviin erottamalla aineistosta kasvot, jonka jälkeen niiden resoluutiota muokataan Gauss-sumennuksella ja kasvot liitetään takaisin alkuperäiseen aineistoon. Lisäksi kuvien valotusta, kontrastia ja terävyyttä muokataan. Samaa toistetaan tietylle määrälle kuvia, joista koostuu tietoaineiston negatiiviset näytteet. Tällaisten negatiivisten näytteiden luominen vaatii vähemmän resursseja kuin syväväärennösten luominen. Kehittäjien mukaan tällä tietoaineistolla koulutettu neuroverkko suoriutuu tunnistustehtävästä paremmalla tarkkuudella ja tehokkaammin kuin muilla tietoaineistoilla koulutettuna. [32]

Kehittäjien omalla tietoaineistolla koulutettua neljää neuroverkkoa on testattu kahdella valmiilla tietoaineistolla. UADFV-aineistolla tunnistusprosentit ovat seuraavat: VGG16 - 84,5, ResNet50 - 98,7, ResNet101 - 99,1 ja ResNet152 - 97,8 prosenttia. Neuroverkkojen toimintaa on testattu erikseen kuvilla ja videoilla. Molemmissa tapauksissa ResNet suoriutuu tunnistuksesta paremmalla tarkkuudella kuin VGG16. Toinen tietoaineisto, jolla neuroverkot on testattu on DeepfakeTIMIT. Kuten kappaleessa 2.2 on kuvattu, tietoaineisto on jaettu LQ (low quality) ja HQ (high quality) videoihin. Matalalaatuisilla videoilla testattaessa neuroverkot saavuttavat seuraavat tunnistusprosentit: VGG16 - 84,6, ResNet50 - 99,9, ResNet101 - 97,6, ResNet152 - 99,4 prosenttia. Korkealaatuisilla videoilla tunnistprosentit ovat mata-

lammat: VGG16 - 57,4, ResNet50 - 93,2, ResNet101 - 86,9, ResNet152 - 91,2 prosenttia. Myös DeepfakeTIMIT-tietoaineistolla ResNet-neuroverkot suoriutuvat tunnistuksesta paremmin kuin VGG16-verkko. [32] Menetelmien saavuttamat testaustulokset on esitetty selvemmin seuraavan kappaleen taulukossa 4.1.

4 Menetelmien arviointi ja tulevaisuus

Menetelmien kriittinen arviointi mahdollistaa niiden erojen ja samanlaisuuksien käsittelyä. Taulukkoon 4.1 on koottu kattavasti tarkemmin käsiteltävien tunnistusmenetelmien tiedot. Kvantitatiivinen vertailu kattaa menetelmien saavuttamien tulosten ohella niiden yhteyksien arviointia. Kvantitatiivisen vertailun lisäksi on oleellista arvioida menetelmien rakenteita ja niiden vaikutusta mallien toimintaan. Rakenteiden vertailu auttaa ymmärtämään menetelmien rajoitteita. Tunnistusmenetelmien tulevaisuuden haasteiden käsittely ja merkittävimpien haasteiden tunnistaminen mahdollistaa niiden ratkaisemisen tulevaisuudessa. Seuraavissa kappaleissa käsitellään näitä näkökulmia tarkemmin.

4.1 Kvantitatiivinen vertailu

Alla olevassa taulukossa on eritelty kolmen käsiteltävän menetelmän tietoja. Taulukossa on menetelmän kehittäjien ja menetelmän lisäksi eritelty sekä koulutukseen että testaukseen käytetyt tietoaineistot. Affiiniseen väärentämiseen perustuvalla tietoaineistolla koulutetut VGG16- ja ResNet-verkot on testattu muiden kehittämillä tietoaineistoilla. Mesonet ja Conv-LSTM on koulutettu ja testattu samalla tietoaineistolla. Menetelmien saavuttamia testaustuloksia verratessa on kiinnostava huomata, että korkeimmat tunnistusprosentit on saavutettu menetelmällä, jonka kou-

lutus ja testaus on toteutettu eri aineistoilla.

Taulukosta nähdään, että korkein tunnistusprosentti on saavutettu ResNet50-verkolla, joka on koulutettu affiiniseen kasvonmuokkaukseen perustuvalla aineistolla. Tämä tunnistusprosentti on saatu DeepfakeTIMIT-tietoaineiston heikkolaatuisilla videoilla. Tarkasteltaessa pelkästään DeepfakeTIMIT-tietoaineistolla saatuja tuloksia ResNet50 tunnistaa myös korkealaatuiset videot suurimmalla todennäköisyydellä. Heikoin tunnistusprosentti on VGG16-verkolla tunnistettaessa saman tietoaineiston korkealaatuisia videoita. Kun verrataan kaikkia tunnistusmenetelmiä toisiinsa, on huomattavissa, että pienimmän ja suurimman tunnistusprosentin välillä on jopa 42,5 prosenttiyksikön ero. Pienimmän ja toiseksi pienimmän tunnistusprosentin välillä on kuitenkin jo 27,1 prosenttiyksikön ero, joten pienintä tunnistusprosenttia voidaan pitää yksittäistapauksena. Kun pienin tunnistusprosentti jätetään huomioimatta, on pienimmän ja suurimman tunnistusprosentin ero 15,4 prosenttiyksikköä.

Kehittäjät	**Menetelmä**	**Koulutuksen tietoaineisto**	**Testauksen tietoaineisto**	**Testaus tulokset**	**Tunnistettavien videoiden tekomenetelmä**
Afchar et al.	Meso-4	Kehittäjien oma + FaceForensics	Kehittäjien oma	89,1 %	Kehittäjien oma = kasvojen korvaaminen toisilla FaceForensics = kasvojen ilmeiden siirtäminen
			FaceForensics	94,6 %	
	MesoInception-4		Kehittäjien oma	91,7 %	
			FaceForensics	98,6 %	
Güera, Delp	Conv-LSTM	HOHA + lisävideot	HOHA + lisävideot	N=20, 96,7 %	Kasvojen korvaaminen toisilla kasvoilla
				N=40, 97,1 %	
				N=80, 97,1 %	
Li, Lyu	Affiininen kasvonmuokkaus + VGG16	Affiinisella väärentämisellä luodut negatiiviset näytteet verkosta kerätyistä kuvista	UADFV	84,5 %	Kaikki menetelmät
			DeepfakeTIMIT	LQ: 84,6 %	
				HQ: 57,4 %	
	Affiininen kasvonmuokkaus + ResNet50		UADFV	97,4 %	
			DeepfakeTIMIT	LQ: 99,9 %	
				HQ: 93,2 %	
	Affiininen kasvonmuokkaus + ResNet101		UADFV	95,4 %	
			DeepfakeTIMIT	LQ: 97,6 %	
				HQ: 86,9 %	
	Affiininen kasvonmuokkaus + ResNet152		UADFV	93,8 %	
			DeepfakeTIMIT	LQ: 99,4 %	
				HQ: 91,2 %	

Taulukko 4.1: Menetelmien vertailutaulukko

Kuten aiemmin tutkielmassa on käynyt ilmi tietoaineiston valinta vaikuttaa saatuihin testaustuloksiin. Affiiniseen kasvonmuokkaukseen perustuvan tunnistuksen kehittäneet Li ja Lyu arvioivat artikkelissaan myös muiden tunnistusmenetelmien tehokkuutta. Vertailun perusteena he käyttävät niitä tietoaineistoja, joilla he testaa-

vat oman menetelmänsä toimintaa eli UADFV ja DeekFakeTIMIT. Tässä tutkielmassa tarkemmin käsitellyistä menetelmistä artikkelissa on arvioitu Mesonetin kahden arkkitehtuurin tunnistustehokkuutta. Molemmille Mesonetin arkkitehtuureille esitetään Lin ja Lyun artikkelissa heikommat tunnistusprosentit kuin Mesonetin esittelevässä artikkelissa. UADVF-tietoaineistolla Meso-4 saavuttaa tunnistusprosentin 84,3 ja MesoInception-4 tunnistusprosentin 82,1. DeepfakeTIMIT-aineiston heikkolaatuisilla videoilla testattaessa Meso-4 ja MesoInception-4 saavuttavat tunnistusprosentit 87,8 ja 80,4. Korkealaatuisilla videoilla prosentit ovat huomattavasti heikommat, 68,4 ja 62,7 prosenttia. [32]

Mesonetin kehittäjät saavuttivat omilla arkkitehtuureillaan keskimäärin 15,9 prosenttiyksikköä paremmat tulokset verrattaessa molempia arkkitehtuureja kaikilla testaukseen käytetyillä tietoaineistoilla. Tämä on seurausta siitä, että Mesonetin arkkitehtuurit on kehitetty ja koulutettu tunnistamaan tietyllä menetelmällä tehtyjä syvävääreennösvideoita, kun Li ja Lyu taas käyttävät testaukseen laajempaa tietoaineistoa. Tunnistusmenetelmiä arvioitaessa tutkijat voivat siis vaikuttaa saatuihin tuloksiin tietoaineiston valinnalla.

4.2 Rakenteiden ja resurssivaatimusten vertailu

Mesonet ja Conv-LSTM ovat molemmat konvoluutioneuroverkkoihin perustuvia tunnistusmenetelmiä. Mesonetin molemmissa verkkoarkkitehtuureissa on 4 konvoluutiokerrosta, mutta ne on järjestetty Meso-4-arkkitehtuurissa peräkkäin ja MesoIncpetion-4-arkkitehtuurissa rinnakkain [31]. Conv-LSTM-menetelmässä on konvoluutioneuroverkko ja LSTM-verkko peräkkäin [15]. Molemmissa menetelmissä hyödynnetään Dropout-teknologiaa koulutuksen yhteydessä ehkäisemään mallin ylisovittumista koulutusaineistoon.

Menetelmissä on käytetty lopussa eri aktivaatiofunktioita, mikä on kiinnostavaa, koska molemmissa on tavoitteena määrittää, onko arvioitava video aito vai

väärennys. Yleensä sigmoid-funktiota, jota on käytetty Mesonetin arkkitehtuureissa, käytetään arvioimaan binäärisiä todennäköisyyksiä[1]. Conv-LSTM-menetelmässä on käytetty aktivaatiofunktiona softmax-funktiota, joka soveltuu useamman kategorian luokittelutehtäviin[2]. Rakenteiden vertailu kattaa tarkoituksella vain kaksi ensimmäistä esitellyistä menetelmistä, koska affiinista väärentämistä hyödyntävä menetelmä perustuu olemassa olevien verkkojen koulutukseen eikä uuden tunnistusverkon rakentamiseen kuten Mesonet ja Conv-LSTM.

Tunnistusmenetelmiä on syytä arvioida myös niiden resurssivaatimusten perusteella. Tunnistusmenetelmän olennaisin resurssivaatimus on sitä suorittavan tietokoneen vaatima laskentateho tai vastaavasti tietyllä laskentateholla suoritukseen kuluva aika. Menetelmän ajallista suoritusta arvioitaessa on otettava huomioon sekä koulutukseen että testaukseen kuluva aika, jotta menetelmän kokonaisvaltaisesta raskaudesta saadaan todellinen kuva.

Mesonetin kehittäjät kuvaavat omalla menetelmällään olevan matala resurssivaativuus, mutta konkreettisia arvoja ei ole esitelty [31]. Conv-LSTM-menetelmällä tunnistetaan vain lyhyitä videoita, joten sen resurssivaativuuden voisi olettaa olevan matala, mutta kehittäjät eivät käsittele menetelmän resurssivaativuutta artikkelissaan [15]. Affiiniseen kasvonväärentämiseen perustuvan menetelmän kehittäjät arvioivat artikkelissaan, että heidän menetelmällään tietoaineiston kehittäminen on nopeaa ja säästää laskentaresursseja verrattuna syvävääreennösmallin käyttöön negatiivisen aineiston tuottamisessa [32]. Resurssivaatimuksille ei esitetä minkään menetelmän kohdalla konkreettisia arvoja.

Tunnistusmenetelmän resurssivaatimuksia arvioitaessa pitää ottaa huomioon myös menetelmän käyttötarkoitus. Esimerkiksi sosiaalisen median alustalla käytettävän tunnistusmenetelmän tulee olla kevyt ja nopea, jotta se tunnistaa väärennöksiä tehokkaasti eikä heikennä alustan käyttäjäkokemusta. Toisaalta merkittävän poliitikon

[1]https://deepai.org/machine-learning-glossary-and-terms/sigmoid-function

[2]https://deepai.org/machine-learning-glossary-and-terms/softmax-layer

kuvan tai videon aitouden määrittämiseen voidaan tarvittaessa käyttää paljonkin resursseja silloin, kun aitouden tunnistaminen on tärkeää. Resurssivaatimusten lisäksi menetelmien yleistettävyyteen ja käytettävyyteen kiinnitetään huomiota artikkeleissa minimaalisesti.

4.3 Menetelmien yleistettävyys ja käytettävyys

Yleisenä haasteena syväväärennösten tunnistusmenetelmissä on niiden yleistettävyys (engl. robustness). Menetelmät on usein kehitetty tunnistamaan juuri tietyllä tekotavalla tuotettuja syväväärennösvideoita. Yleistettävyys[3] viittaa menetelmien kykyyn tunnistaa täysin uusia videoita, jotka ei ole osa alkuperäistä koulutukseen käytettyä tietoaineistoa. Hyvän yleistettävyyden omaava menetelmä hallitsee lisäksi kohinaa sisältävän datan, sen jaottelun muutokset ja hyökkäykset paremmin.

Hussain ja muut tutkivat eri syväväärennöstunnistusmenetelmien yleistettävyyttä luomalla hyökkäyksiä, joissa syväväärennösvideoita muokattiin niin, että mallit epäonnistuivat niiden tunnistuksessa. [6] Tutkimus osoittaa, että nykyisiä tunnistusmenetelmiä on mahdollista huijata. Syväväärennösten luominen ja tunnistaminen voidaan nähdä toisiaan kehittävinä voimina. Jotta tunnistusmenetelmien laajempi käyttö tulisi käytännössä mahdolliseksi, tunnistusmenetelmien olisi hyvä tunnistaa eri menetelmillä tehtyjä syväväärennöksiä.

Yleistettävyyden ohella toinen tunnistusmenetelmien rajoite liittyy niiden käyttäjäystävällisyyteen. Menetelmillä ei ole graafisia käyttöliittymiä, joten niiden käyttö alan noviiseille ei ole tällä hetkellä mahdollista. Toisaalta yhdenkään tarkemmin esitellyn menetelmän kehittäjät eivät itse asettaneet tavoitteekseen kuluttajien mahdollisuutta käyttää menetelmiä, vaan menetelmät on kehitetty tutkimusnäkökulmasta. Kuluttajien ja organisaatioiden näkökulmasta olisi tärkeää myös levittää

[3]https://medium.com/@slavadubrov/understanding-machine-learning-robustness-why-it-matters-and-how-it-affects-your-models-5e2cb5838dab

tietoisuutta tunnistusmenetelmistä ja niiden toimintaperiaatteesta. Ymmärryksen avulla voidaan rakentaa luottamusta menetelmien toimintaan, joka todennäköisesti lisäisi menetelmien kysyntää myös alan ulkopuolella. Yleistettävyyden ja käytettävyyden parantaminen ovat tulevaisuuden haasteita, jotka rajoittavat tunnistusmenetelmien käytön lisääntymistä. Muita tulevaisuuden haasteita ja mahdollisuuksia käsitellään seuraavassa kappaleessa.

4.4 Tulevaisuuden haasteet ja mahdollisuudet

Menetelmien haasteet liittyvät uusien teknologioiden hyödyntämiseen ja syväväärennösten kehittymiseen, joka voidaan nähdä vastavoimana tunnistusmenetelmille. Menetelmien yksi merkittävä haaste on tulosten heikko ennustettavuus, joka on seurausta niin sanotusta blackbox-ongelmasta. Ongelman ytimessä on se, että mallien oppiminen ja päätösten perustelu on niin itsenäistä, että loppujen lopuksi kehittäjien on lähes mahdotonta tietää, mihin mallin antamat tulokset tarkalleen perustuvat. Tämän seurauksena ei voida varmasti ennustaa millaisen luokittelutuloksen malli antaa tietylle aineistolle. [2] [20] Mallien toiminnan ymmärtäminen yksityiskohtaisemmin helpottaisi niiden ongelmien määrittelyä ja korjaamista, mikä parantaisi tulosten ennustettavuutta.

Kehitykselle olennaista on uusien teknologioiden käyttöönotto. Tunnistusmenetelmissä voidaan mahdollisesti tulevaisuudessa hyödyntää peliteorian tekniikoita hyökkäysten torjumiseen ja videoiden tunnistuksessa voidaan hyödyntää ääniraitaa tunnistuksen apuna. Äänisyväväärennöksiin liittyy myös ääniohjattuihin IoT-laitteisiin kohdistuvat konkreettiset uhat. Vahvistusoppimisen yhdistäminen aktiiviseen oppimiseen voisi tehostaa tunnistusmenetelmien kykyä tunnistaa monilla eri menetelmillä luotuja syväväärennöksiä. [5]

Tunnistusmenetelmien käytettävyyttä voitaisiin parantaa sulauttamalla menetelmiä osaksi eri alustoja, joissa syväväärennöksiä esiintyy. Tämä mahdollistaisi vi-

deoiden tunnistuksen jo ennen niiden julkaisua. Alusta voisi estää videon lataamisen kokonaan tai lisätä sen yhteyteen tiedon, että video todennäköisesti on syväväärennös. [20] Syväväärennösten julkaisijoiden motiivit on myös tärkeä tiedostaa, jotta heidän toimintatapojaan voidaan ymmärtää ja jopa ennustaa. Rikollisuuden välineenä syväväärennösten käytön ennustetaan lisääntyvän etenkin rahastukseen liittyvissä rikoksissa. Nykyisten tunnistusmenetelmien heikkouksien tutkiminen auttaa määrittämään mallien tärkeimpiä kehityskohtia. [4]

5 Päätelmät ja yhteenveto

Syväväärennösten ja niiden tunnistusmenetelmien kenttä on laajentunut nopeasti syväväärennös-termin ensimmäisestä esiintymisestä vuonna 2017. Tällä hetkellä syväväärennöksiä voidaan tuottaa monilla tekoälyn sovelluksilla, joista eniten käytössä ovat GAN-verkot, enkoodaus-dekoodaus-verkot sekä diffuusiomallit. Tunnistusmenetelmät rakentuvat aiemmin kehitettyjen kuvantunnistusteknologioiden pohjalle. Tunnistusmenetelmien saavuttamiin tuloksiin vaikuttavat merkittävästi käytetyt tietoaineistot, joita voidaan pitää yhtenä kehitystä rajoittavana tekijänä.

Tutkielmani ensimmäinen tutkimuskysymys oli: *"Mihin syväväärennösvideoiden tunnistus voi perustua?"*. Laajemmasta näkökulmasta katsottuna menetelmät perustuvat laajasti eri teknologioihin, kuten konvoluutioneuroverkkoihin, takaisinkytkettyihin neuroverkkoihin ja LSTM-verkkoihin sekä näiden sovelluksiin ja yhdistelmiin. Etenkin konvoluutioverkkojen käyttö tunnistuksessa on osoittautunut tarkaksi ja tehokkaaksi teknologiaksi. Konvoluutioverkkojen hyvä soveltuvuus syväväärennösvideoiden tunnistukseen perustuu konvoluutioverkkojen hyviin tuloksiin objektien ja alueiden tunnistuksessa. Syväväärennösvideoiden tunnistuksessa hyödynnettävät visuaaliset tekijät voidaan jakaa avaruudellisiin ja ajallisiin virheisiin, joita havaitsemalla neuroverkot tunnistavat väärennöksen.

Toinen tutkimuskysymykseni oli *"Mitkä ovat merkittävimpien tunnistusmenetelmien toimintaperiaatteet ja miten ne eroavat toisistaan?"*. Merkittävinä tunnistusmenetelminä voidaan pitää niitä, joihin viitataan paljon ja keskeisimmissä läh-

teissä. Ensimmäinen keskeisimmistä menetelmistä on Mesonet. Mesonet on arkkitehtuuri, josta on kehitetty kaksi eri sovellusta eri tavoilla tuotetuille syvävääärennöksille. Meso-4 arkkitehtuuri rakentuu peräkkäisistä konvoluutiokerroksista ja MesoInception-4 rakentuu rinnakkaisista konvoluutiokerroksista. Conv-LSTM on toinen merkittävä tunnistusmenetelmä, joka hyödyntää konvoluutioverkkoa piirteiden keräämiseen ja LSTM-verkkoa videon ruutujen välisen aikajatkuvuuden analysointiin. Kolmas merkittävistä tunnistusmenetelmistä perustuu tietoaineiston luomiseen käyttämällä affiinista väärentämistä. Luodulla tietoaineistolla koulutetaan tunnettuja kuvantunnistusverkkoja ja tavoitteena on tunnistaa kaikilla eri menetelmillä tuotettuja syvävääärennöksiä.

Merkittävien tunnistusmenetelmien keskeisimmät erot liittyvät niiden rakenteisiin ja lähestymistapoihin. Mesonetin verkkoarkkitehtuurit ovat perinteisiä konvoluutioverkkoja, kun taas Conv-LSTM hyödyntää LSTM-verkkoa ja konvoluutioverkkoa. Affiiniseen väärentämiseen perustuva menetelmä rakentuu taas olemassa olevien verkkojen pohjalle. Mesonetin ja Conv-LSTM verkkojen tavoitteena on tunnistaa juuri tietyillä menetelmillä tuotettuja kuvia, kun taas kolmannen menetelmän tavoitteena on tunnistaa syvävääärennökset tuotantomenetelmästä huolimatta.

Kolmas tutkimuskysymykseni oli *”Miten menetelmiä tulee kehittää tulevaisuudessa?”*. Menetelmien keskeisimmät haasteet liittyvät yleistettävyyteen, ennustettavuuteen ja käytettävyyteen. Yleistettävyys viittaa tunnistusmenetelmien kykyyn tunnistaa täysin uusia syvävääärennöksiä, jotka eivät esiintyneet koulutusaineistossa. Heikko ennustettavuus on seurausta blackbox-ongelmasta, joka viittaa siihen, että syväoppivien mallien päätöksenteko nähdään niin sanottuna mustana laatikkona, jonka sisältö ei ole nähtävissä. Kehittäjillä ei siis ole tiedossa, mihin mallien päätöksenteko tarkalleen perustuu, joten luokittelun tuloksia ei voida ennustaa. Käytettävyyden ongelma viittaa mallien käytännön sovelluksien heikkouteen. Tällä hetkellä kuluttajien ja organisaatioiden käytössä ei ole tunnistusohjelmaa, jolla tutkimuksen

hyödyt saataisiin todelliseen käyttöön. Jotta syväväärennösten tunnistuksesta saadaan merkittävä vastavoima syväväärennöksille, tulee niiden näkyä eri alustoilla ja niistä on heräteltävä keskustelua myös tiedeyhteisön ulkopuolella.

Syväväärennökset rikollisten ja ääri-ideologisten järjestöjen välineenä aiheuttavat todellisia uhkia, jotka on otettava vakavasti ja joihin on varauduttava. Tulevaisuudessa syväväärennösten kehitys tulee kiihtymään edelleen, mikä lisää tunnistusmenetelmien tärkeyttä. Tunnistusmenetelmiä on kehitetty ja tutkittu paljon, mutta niiden todellinen hyöty saadaan käyttöön vasta, kun menetelmät ovat kuluttajien ja organisaatioiden käytössä. Tämän mahdollistamiseksi tulee jatkossa käyttää resursseja. Syväväärennösten ja niiden tunnistusmenetelmien kenttä tulee laajenemaan todennäköisesti jopa odottamattomilla tavoilla, joten tutkimusta tarvitaan edelleen lisää.

Lähdeluettelo

[1] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan ja S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats", *IEEE Access*, vol. 11, s. 144 497–144 529, 2023. DOI: 10.1109/ACCESS.2023.3344653.

[2] L. Verdoliva, "Media Forensics and DeepFakes: An Overview", *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, nro 5, s. 910–932, 2020. DOI: 10.1109/JSTSP.2020.3002101.

[3] P. Korshunov ja S. Marcel, "Deepfake detection: humans vs. machines", *arXiv preprint arXiv:2009.03155*, 2020.

[4] Y. Mirsky ja W. Lee, "The Creation and Detection of Deepfakes: A Survey", *ACM Computing Surveys (CSUR)*, vol. 54, nro 1, s. 1–41, 2021. DOI: 10.1145/3425780.

[5] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza ja H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward", *Applied intelligence*, vol. 53, nro 4, s. 3974–4026, 2023.

[6] S. Hussain, P. Neekhara, B. Dolhansky et al., "Exposing vulnerabilities of deepfake detection systems with robust attacks", *Digital Threats: Research and Practice (DTRAP)*, vol. 3, nro 3, s. 1–23, 2022. DOI: 10.1145/3464307.

[7] I. Goodfellow, Y. Bengio ja A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[8] K. Gurney, *An introduction to neural networks*. CRC press, 1997.

[9] P. Dhariwal ja A. Nichol, "Diffusion models beat gans on image synthesis", *Advances in neural information processing systems*, vol. 34, s. 8780–8794, 2021.

[10] A. Paullada, I. D. Raji, E. M. Bender, E. Denton ja A. Hanna, "Data and its (dis) contents: A survey of dataset development and use in machine learning research", *Patterns*, vol. 2, nro 11, 2021. DOI: 10.1016/j.patter.2021.100336.

[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies ja M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images", teoksessa *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, s. 1–11. DOI: 10.1109/ICCV.2019.00009.

[12] P. Korshunov ja S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection", *arXiv preprint arXiv:1812.08685*, 2018.

[13] W. Wang, Y. Yang, X. Wang, W. Wang ja J. Li, "Development of convolutional neural network and its application in image classification: a survey", *Optical Engineering*, vol. 58, nro 4, 2019. DOI: 10.1117/1.OE.58.4.040901.

[14] Y. LeCun, Y. Bengio ja G. Hinton, "Deep learning", *Nature*, vol. 521, nro 7553, s. 436–444, 2015. DOI: 10.1038/nature14539.

[15] D. Güera ja E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks", teoksessa *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, s. 1–6. DOI: 10.1109/AVSS.2018.8639163.

[16] K. Simonyan ja A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions", teoksessa *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, s. 1–9.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens ja Z. Wojna, "Rethinking the inception architecture for computer vision", teoksessa *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, s. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[19] K. He, X. Zhang, S. Ren ja J. Sun, "Deep residual learning for image recognition", teoksessa *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, s. 770–778. DOI: 10.1109/CVPR.2016.90.

[20] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey", *Computer Vision and Image Understanding*, vol. 223, s. 103 525, 2022. DOI: 10.1016/j.cviu.2022.103525.

[21] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi ja P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", *Interfaces (GUI)*, vol. 3, nro 1, s. 80–87, 2019.

[22] A. Chintha, B. Thai, S. J. Sohrawardi et al., "Recurrent convolutional structures for audio spoof and video deepfake detection", *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, nro 5, s. 1024–1037, 2020. DOI: 10.1109/JSTSP.2020.2999185.

[23] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner ja L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection", *arXiv preprint arXiv:1812.02510*, 2018.

[24] T. Fernando, C. Fookes, S. Denman ja S. Sridharan, "Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks", *arXiv preprint arXiv:1911.07844*, 2019.

[25] H. Dang, F. Liu, J. Stehouwer, X. Liu ja A. K. Jain, "On the detection of digital face manipulation", teoksessa *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 2020, s. 5781–5790. DOI: 10.1109/CVPR42600.2020.00582.

[26] I. Amerini, L. Galteri, R. Caldelli ja A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN", teoksessa *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, s. 1205–1207. DOI: 10.1109/ICCVW.2019.00152.

[27] H. H. Nguyen, J. Yamagishi ja I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos", teoksessa *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2019, s. 2307–2311. DOI: 10.1109/ICASSP.2019.8682602.

[28] U. A. Ciftci, I. Demir ja L. Yin, *Fakecatcher: detection of synthetic portrait videos using biological signals*, US Patent 11,687,778, kesäkuu 2023.

[29] L. Li, J. Bao, T. Zhang et al., "Face x-ray for more general face forgery detection", teoksessa *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, s. 5001–5010. DOI: 10.1109/CVPR42600.2020.00505.

[30] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera ja D. Manocha, "Emotions Don't Lie: A Deepfake Detection Method Using Affective Cues", *arXiv preprint arXiv:2003.06711v3*, 2020.

[31] D. Afchar, V. Nozick, J. Yamagishi ja I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network", teoksessa *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, s. 1–7. DOI: 10.1109/WIFS.2018.8630761.

[32] Y. Li ja S. Lyu, ”Exposing deepfake videos by detecting face warping artifacts”, *arXiv preprint arXiv:1811.00656*, 2018.

[33] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt ja M. Nießner, ”Face2Face: Real-Time Face Capture and Reenactment of RGB Videos”, teoksessa *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, s. 2387–2395. DOI: 10.1109/CVPR.2016.262.