Konvoluutioneuroverkkojen hyödyntäminen audiosyväväärennösten tunnistamisessa

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Joulukuu 2025
Heidi Nummela

TURUN YLIOPISTO
Tietotekniikan laitos

HEIDI NUMMELA: Konvoluutioneuroverkkojen hyödyntäminen audiosyväväärennösten tunnistamisessa

LuK-tutkielma, 24 s.
Tietojenkäsittelytiede
Joulukuu 2025

Audiosyväväärennökset eli keinotekoisesti muokatut tai luodut audiotallenteet ovat uhka sekä yksilöille että yhteiskunnalle. Menetelmät audiosyväväärennösten tunnistamiseksi ovat kehittyneet tekoälymenetelmien käyttöönoton myötä tarkemmiksi ja helppokäyttöisemmiksi. Toisaalta myös syväväärennösten luomiseen käytetyt teknologiat ovat kehittyneet, joten tunnistusmenetelmiä on kehitettävä jatkuvasti paremmiksi. Audiosyväväärennökset ovat kuitenkin viimeisimmässä syväväärennösten tunnistamista käsittelevässä akateemisessa tutkimuksessa jääneet videosyväväärennösten varjoon.

Tässä tutkielmassa tutkitaan konvoluutioneuroverkkojen käyttöä audiosyväväärennösten tunnistamisessa. Konvoluutioneuroverkot ovat yleisimpiä audiosyväväärennösten tunnistukseen käytettäviä algoritmeja johtuen niiden erityisen hyvästä kyvystä käsitellä audion kuvamuotoisia esityksiä. Lisäksi tutkielma tarkastelee audiosyväväärennösten tunnistamiseen käytettyjen menetelmien tarkkuutta. Aihetta käsittelevän tutkimuksen perusteella voidaan päätellä tunnistusmenetelmien olevan kohtuullisen tarkkoja, vaikkakin tunnistuksen tarkkuus vaihtelee melko paljon riippuen käytetystä algoritmista. Tarkkuuden vaihtelu voi johtua käytetyn algoritmin lisäksi monesta syystä, kuten tunnistusmallien koulutukseen käytetystä tietoaineistosta, aineiston esiprosessoinnista tai suorituksen arviointiin käytetystä menetelmästä.

Tulevassa tutkimuksessa tulisikin tarkemmin miettiä tunnistusalgoritmien arviointiin ja koulutukseen käytettävää tietoaineistoa. Koulutusaineiston tulisi olla laaja ja arviointiin käytetty tietoaineisto täytyisi ilmoittaa selvästi. Arviointiin käytettäviä mittareita tulisi myös harkita tarkkaan, jotta arvio olisi mahdollisimman tarkka. Lisäksi tulisi tutkia mitkä audion esitysmuodot ovat tarkkuudeltaan parhaimpia audiosyväväärennösten tunnistamiseen. Tunnistusmallien suorituskykyä koulutusaineiston ulkopuolella tulisi parantaa ja mallien käyttöönottoa tulisi helpottaa, jotta niistä olisi eniten hyötyä.

Asiasanat: syväväärennös, tunnistus, konvoluutioneuroverkot, syväoppiminen

Sisällys

Kuvat

Taulukot

1 Johdanto

Syväväärennös (engl. Deepfake) -termi on yhdistelmä sanoista syväoppiminen (engl. deep learning) ja väärennös (engl. fake) [1]. Syväväärennökset ovat tekoälyn tekniikoita, kuten generatiivisia kilpailevia verkostoja (engl. Generative Adversial Networks, GAN) ja autoenkoodaajia (engl. Autoencoder) hyödyntäen luotuja käsiteltyjä tai keinotekoisia audiomuotoisia tai visuaalisia mediasisältöjä [2]. Syväväärennökset luodaan hyödyntäen jotakin jo olemassa olevaa sisältöä [3]. Siten ne saadaan vaikuttamaan aidoilta ja luomaan vaikutelman, että syväväärennöksen kohde sanoisi tai tekisi jotain, jota hän ei tosiasiassa ole koskaan sanonut tai tehnyt [2]. Syväväärennöstekniikoilla voidaan esimerkiksi vaihtaa kuvasta tai videosta henkilön kasvot toisen henkilön kasvoihin sekä muuttaa alkuperäistä ääntä ja kasvojen liikkeitä videosta aidon oloisiksi [1]. Syväväärennösten luominen perustuu GAN-verkkojen ja neuroverkkojen (engl. Neural Networks) käyttöön [4]. Syväväärennöstekniikoita koulutetaan sekä luomaan syväväärennöksiä että myöskin tunnistamaan niitä, jotta ne voisivat luoda yhä parempia syväväärennöksiä [4].

Syväväärennösteknologioilla on lukuisia hyödyllisiä käyttötarkoituksia: syväväärennöksiä voidaan käyttää mykille puheeseen perustuvan kommunikointitavan kehittämiseen [5] tai reaaliaikaisen käännösjärjestelmän toimintaan [6]. Niitä voidaan myös soveltaa usealla eri tavalla opetuksessa, esimerkiksi kieliopinnoissa ääntämisen opettamiseen [7]. Lisäksi syväväärennösteknologialla on sovellusmahdollisuuksia terveydenhuollossa, jossa sillä voidaan tunnistaa puhemalleissa ajan myötä tapahtu-

via muutoksia ja siten tunnistaa neurologisia sairauksia ajoissa sekä seurata niiden kehitystä [7].

Syväväärennösteknologiaa voidaan kuitenkin käyttää myös pahantahtoisiin tarkoituksiin. Keinotekoisesti luotua ääntä voidaan käyttää hyväksi muun muassa rahahuijauksissa, tiedon kalastelussa, disinformaation levittämisessä ja toisen henkilön maineen vahingoittamisessa. Syväväärennösten käytöstä rahahuijauksissa on jo esimerkkitapauksia, joissa työntekijöille on soitettu, esiinnytty toimitusjohtajana ja pyydetty siirtämään nopeasti rahaa toiselle yritykselle. [7] Syväväärennösten avulla luodun mis- ja disinformaation leviäminen on suuri riski yhteiskunnassa, jossa tieto liikkuu ja leviää erilaisilla sosiaalisen median alustoilla eikä sen aitouden varmistamiselle ole vahvoja, helposti saatavilla olevia työkaluja. Lisäksi syväväärennösten muodossa leviävällä väärällä tiedolla voidaan vahingoittaa muun muassa valtioiden demokraattisia järjestelmiä, kansainvälisiä suhteita ja oikeusjärjestelmää [2].

Syväväärennösteen tunnistaminen pelkän visuaalisen tai kuuloon perustuvan tarkastelun perusteella on haasteellista. Perinteisesti syväväärennösten tunnistus on tapahtunut tunnistamalla epäjohdonmukaisuuksia suhteessa luonnolliseen mediaan tai analysoimalla kuvan kohinaa [1]. Tämä on kuitenkin käynyt liian tehottomaksi syväväärennösteknologioiden kehityksen myötä, jonka seurauksena on siirrytty käyttämään koneoppimiseen tai syväoppimiseen perustuvia tunnistusmenetelmiä [3]. Syväoppimisalgoritmit ovat näistä kahdesta menetelmästä tehokkaampia syväväärennösten tunnistamisessa [3] ja niihin syvennytään tarkemmin luvussa 3.

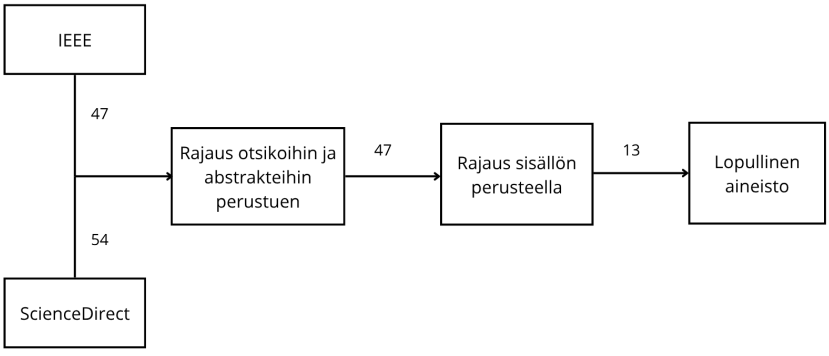
Syväväärennökset on ollut paljon huomiota osakseen saanut aihe viime vuosina, mutta tutkimus on keskittynyt suurimmaksi osaksi videomuotoisiin syväväärennöksiin. Tämän kirjallisuuskatsauksena toteutetun tutkielman tarkoituksena on keskittyä vähemmän huomiota saaneiden audiosyväväärennösten tunnistamiseen käymällä läpi viimeaikaista tieteellistä kirjallisuutta aiheeseen liittyen.

Tässä tutkielmasssa tarkastellaan konvoluutioneuroverkkojen (engl. Convolutional Neural Network, CNN) käyttöä audiosyväväärennösten tunnistamisessa, miten niitä käytetään, millaisia erilaisia CNN-malleja on käytössä ja kuinka tarkkoja niitä käyttävät menetelmät ovat. Tutkielmassa vastataan kahteen tutkimuskysymykseen, jotka ovat:

TK1: Miten konvoluutioneuroverkkoja käytetään audiosyväväärennösten tunnistamisessa?

TK2: Miten tarkkoja tunnistusmenetelmät ovat audiosyväväärennösten tunnistamisessa?

Tutkielma on toteutettu kirjallisuuskatsauksena. Aineistoa on haettu IEEE- ja Science Direct -tietokannoista. Hakulauseena on käytetty *("audio deepfake"OR "audio deep fake") AND (cnn OR "convolutional neural network") AND detection.* Hakutuloksia oli yhteensä 101, joista rajattiin otsikoiden ja tiivistelmien perusteella 47:ään aiheen kannalta olennaisimpaan. Tässä vaiheessa poistettiin toisiin aihealueisiin kuuluvat artikkelit ja artikkelit, joissa ei käsitelty mitään CNN-mallia. Lopullisessa rajauksessa rajattiin pois sisällön perusteella muuhun kuin vain audiosyväväärennöksiin kuuluva aineisto sekä ne, joiden suoritusta ei arvioitu tarkkuudella. Jäljelle jäi 13 artikkelia. Aineistojen valintaprosessi on esitetty kuvassa 1.1.



Kuva 1.1: Aineistojen rajausprosessi

Toisessa luvussa taustoitetaan aihetta käsittelemällä syväväärennösten tyyppejä sekä audiosyväväärennösten luomista ja tunnistamista. Kolmas luku käsittelee syväoppimismenetelmien käyttöä audiosyväväärennösten tunnistuksessa ja konvoluutioneuroverkkoja. Neljännessä luvussa tarkastellaan audiosyväväärennösten tunnistukseen käytettäviä konvoluutioneuroverkkomalleja ja tunnistusmenetelmien tarkkuutta aihetta koskevan tutkimuksen perusteella sekä esitetään keskeiset johtopäätökset. Tutkielman viidennessä luvussa vastataan tutkimuskysymyksiin ja esitetään kirjallisuuskatsauksen perusteella tehdyt tärkeimmät johtopäätökset tunnistusmalleista.

2 Audiosyväväärennökset

Ensimmäiseksi tässä luvussa esitellään syväväärennösten neljä päätyyppiä: kuva-, audio-, video- ja yhdistelmäsyväväärennökset. Lisäksi tarkastellaan kahdentyyppisten audiosyväväärennösten luontia. Ensin käsitellään tekstistäpuheeksi -syväväärennöksiä luovien integroitujen käsittelyputkien toimintaa, joka koostuu aineiston esiprosessoinnista, sen enkoodauksesta käsiteltävään muotoon ja dekoodauksesta takaisin sekä muuttamisesta takaisin ääniaaltomuotoon. Tämän jälkeen tarkastellaan äänenvaihtotekniikkaa, jossa ensin koulutetaan malli erottamaan ja tunnistamaan äänten ominaisuuksia, sen jälkeen käytetään koulutettua mallia muokkaamaan ääntä ja lopulta luodaan haluttu puhe. Luvun lopussa taustoitetaan audiosyväväärennösten tunnistamisen tekniikoita.

2.1 Syväväärennösten tyypit

Syväväärennöksistä voidaan tunnistaa neljää erilaista tyyppiä: kuvasyväväärennös, audiosyväväärennös, videosyväväärennös ja yhdistelmäsyväväärennös. Kuvasyväväärennöksissä alkuperäisen kuvan henkilön ulkonäköä muokataan toista henkilöä vastaavaksi. [1]

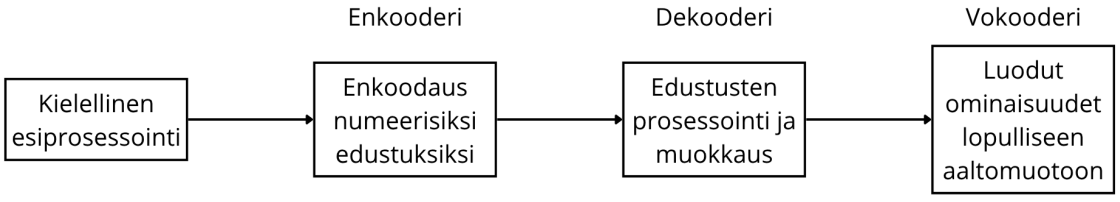
Audiosyväväärennöksiä on kahdenlaisia. Äänenvaihto (engl. Voice Swapping) -syväväärennöksessä muunnetaan yhden henkilön ääni kuulostamaan toisen henkilön ääneltä. Tekstistäpuheeksi (engl. Text to Speech, TTS) -syväväärennöksessä puolestaan muutetaan kirjoitettua tekstiä tietyllä äänellä puhutuksi. [1]

Myös videosyväväärennökset voidaan jakaa kahteen tyyppiin. Kasvojenvaihto (engl. Face-Swapping) -syväväärennöksessä vaihdetaan yhden henkilön kasvot näyttämään toisen kasvoilta. Kasvojenmuodonmuutos (engl. Face-morphing) -syväväärennöksessä tehdään huomaamattomampi siirros yhdestä henkilöstä toiseen muokaten ensimmäisen henkilön kasvoja vähitellen toisen kasvoja muistuttaviksi. [1]

Neljäs syväväärennösten muoto, yhdistelmäsyväväärennös, on myös luultavasti näistä uhkaavin, sillä siinä on yhdistettynä sekä audio- että videosyväväärennös ja seurauksena syväväärennöksen kokonaisuus vaikuttaa entistä aidommalta [1].

2.2 Audiosyväväärennösten luominen

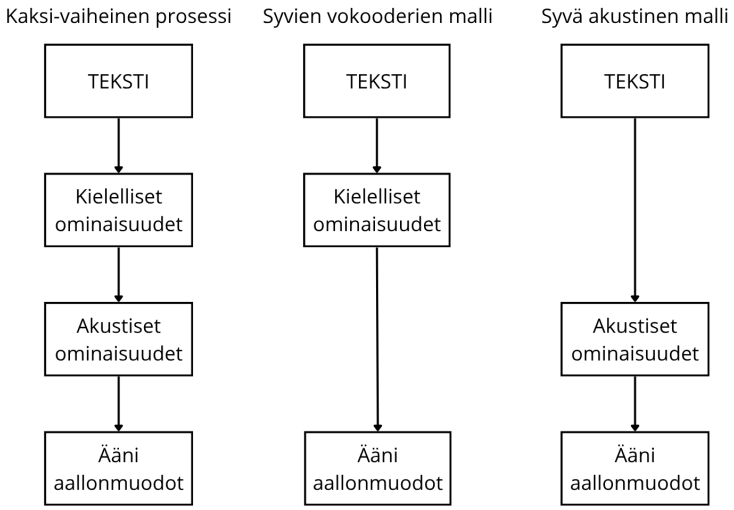
Audiosyväväärennöksiä on siis kahdenlaisia: TTS-syväväärennökset ja äänenmuuntamissyväväärennökset. TTS-tekniikan toiminta alkaa esiprosessoinnilla, jossa käsitellään syötteenä annettua tekstiä ja siitä erotetaan prosodiset ja foneettiset ominaisuudet seuraavaa vaihetta varten. Esiprosessoinnista saatu data muutetaan foneettisia yksityiskohtia ja kielellisiä piirteitä sisältäviksi numeerisiksi edustuksiksi enkooderissa. Enkoodauksen jälkeen dekooderi käsittelee luotuja edustuksia ja muokkaa ne vaihteleviksi akustisiksi ominaisuuksiksi tai spektrogrammeiksi eli äänen aikataajuusesityksiksi. Lopulta vokooderi muokkaa luodut ominaisuudet lopulliseen ääniaaltomuotoon ja luo siten keinotekoista puhetta. [7] TTS-tekniikan toimintaa on kuvattu kuvassa 2.1.



Kuva 2.1: Audiosyväväärennösten tunnistus enkooderi-dekooderi -arkkitehtuurilla

Syvät generatiiviset mallit ovat entisestään parantaneet mahdollisuuksia luoda yhä aidommalta kuulostavaa keinotekoista puhetta. Nämä mallit käyttävät yleensä variationaalisia autoenkoodereita (engl. Variational Autoencoder, VAE), GAN-verkkoja tai autoregressiivisiä malleja. Syvät generatiiviset mallit nopeuttavat TTS-järjestelmien koulutusta jättäen joko kielellisten ominaisuuksien käsittelyvaiheen tai akustisten ominaisuuksien käsittelyvaiheen vokoodereista pois. Syvät vokooderit luovat aaltomuotoja suoraan kielellisistä ominaisuuksista, jolloin akustisten ominaisuuksien käsittely jää pois. Syviä vokoodereita yleisemmin käytetään kuitenkin syvää akustista mallia, jossa kielellisten ominaisuuksien käsittely jää pois ja aaltomuotoja luodaan suoraan akustisista ominaisuuksista. Syviä vokoodereita käyttäviä sovelluksia ovat esimerkiksi WaveNet ja DeepVoice2, kun taas syvää akustista mallia käyttäviä ovat DeepVoice3 ja FastSpeech2. [7] Perinteistä, syvää vokooderi- ja syvää akustista mallia on havainnollistettu kuvassa 2.2.

Autoregressiiviset mallit tuottavat korkealaatuista ja luonnolliselta kuulostavaa puhetta, mutta ne käyttävät toimiakseen suuria tietoaineistoja, niillä on suuret laskennalliset kustannukset ja käytön myötä niihin kerääntyy virheitä. Transforme-



Kuva 2.2: Kolme mallia tekstistäpuheeksi -audiosyvävääärennösten luomiseen

riin eli tarkkaavaisuusmekanismeihin perustuvaan verkkoarkkitehtuuriin pohjautuvat mallit pystyvät mallintamaan pitkän kantaman riippuvaisuuksia ja luomaan korkealaatuista keinotekoista puhetta, mutta ne vaativat toimiakseen laajoja tietoaineistoja tarkoista puheen äänityksistä. VAE-pohjaiset mallit ovat kehittyneitä luonnollisen ja korkealaatuisen puheen luomisessa. [7] GAN-pohjaiset mallit kouluttavat itseään tunnistamaan keinotekoista sisältöä itse luomillaan syvävääreennöksillä [8]. Siten ne parantavat jatkuvasti luomiensa syvävääreennösten tarkkuutta ja luovat siten hyvin luonnolliselta vaikuttavia lopputuloksia [8]. Audiodiffuusiomallit luovat ensin akustisella mallilla äänen akustiset ominaisuudet ja sen jälkeen vokooderilla syntesisoivat aallonmuodon [7]. Ne parantavat kohinasignaalia iteratiivisesti mallintamalla ja toistamalla audiomalleja neuroverkoilla [7].

Äänen muuntamisen tekniikoissa tarkoituksena on pyrkiä muuntamaan henkilön ääni kuulostamaan toisen henkilön ääneltä siten, että audio säilyttää silti alkuperäisen kielellisen sisällön. Rinnakkaisessa äänenmuuntamisessa lähde- ja kohdepuhujan tulee puhua samaa asiaa, mikä rajoittaa sen käyttöä. Ei-rinnakkaisessa äänen muutoksessa ei ole tällaista rajoitetta. Rinnakkaisen äänenmuuntamisen koulutusvaiheessa erotetaan lähde- ja kohdepuhujan äänitteistä akustiset ominaisuudet, jonka jälkeen ne sovitetaan keskenään optimaaliseksi yhteensopivuudeltaan ja ajoitukseltaan. Sen jälkeen luovaa mallia koulutetaan sovitetuilla ominaisuuspareilla, jotta malli oppisi kartoittamaan lähdeominaisuuksista kohdeominaisuuksia. Muutosvaiheessa sovelletaan ensin luovaa mallia uuteen lähdepuheeseen ja muutetaan sen äänen ominaisuuksia kohdepuhujan äänen ominaisuuksiin. Tämän jälkeen muutetut ominaisuudet vaihdetaan takaisin puheaaltomuotoon ja lopulliseksi puheeksi käyttäen vokooderia tai neuraalista vokooderia. Ei-rinnakkaisen äänen muutoksen prosessi on monimutkaisempi ja luovan mallin koulutus vaatii enemmän työtä. [7] Esimerkiksi Hsu ym. [9] ehdottavat luovaa mallia, jossa enkooderi oppii puhujas-

ta itsenäiset foneettiset edustukset ja dekooderi oppii luomaan puhetta halutulle puhujalle.

2.3 Audiosyväväärennösten tunnistaminen

Tunnistusteknologiat voidaan jakaa yksimuotoisiin ja monimuotoisiin teknologioihin sen perusteella, käytetäänkö tunnistuksessa yhtä vai useampaa syötemuotoa. Jos siis käytetään vain esimerkiksi kuvaa tunnistukseen, on tunnistusteknologia yksimuotoinen, kun taas videota ja audiota yhdessä tunnistuksessa käytettäessä tunnistusteknologia on monimuotoinen. Hyvää yksimuotoisissa tunnistusteknologioissa on niiden yleistettävyys, niiden pienempi laskennallinen monimutkaisuus ja helposti saatavilla olevat tietoaineistot. Ne eivät kuitenkaan ole enää yhtä tehokkaita tunnistamaan syväväärennöksiä niiden luontitapojen kehittyessä kuin monimuotoiset tunnistusteknologiat ovat. Yksimuotoiset tunnistusteknologiat jaetaan edelleen perinteisiin ja edistyneisiin tunnistustekniikoihin. Monimuotoiset teknologiat voidaan jakaa edelleen syväoppimiseen perustuviin ja yhdistelmäfuusio-tekniikoihin. [10]

Tässä tutkielmassa keskitytään yksimuotoisiin edistyneisiin tunnistustekniikoihin, jollaisia ovat syväoppimiseen perustuvat algoritmit ja hybridi-algoritmit. Yksimuotoiset edistyneet tekniikat käyttävät syväoppimisen tekniikoita erottamaan piirteitä syötteestä sekä prosessoimaan ja luokittelemaan erotettuja piirteitä. Syväoppimiseen perustuvat algoritmit prosessoivat erotettuja audion syviä piirteitä syväoppimismalleilla. Hybridi-algoritmit sen sijaan perustuvat useiden samanaikaisten koneoppimis- tai syväoppimismallien käyttöön audion piirteiden erottamisessa, prosessoimisessa ja luokittelemisessa. [10] Tämä tutkielma keskittyy erityisesti syväoppimiseen perustuviin malleihin sillä perusteella, että hybridi-algoritmien taustalla on yleensä syväoppimiseen perustuvia algoritmeja ja siten ymmärtämällä syväoppimiseen perustuvia algoritmeja ymmärtää pintapuolisesti myös hybridi-algoritmeja.

Syväoppimiseen perustuvia tunnistusalgoritmeja käsitellään tarkemmin kappaleessa 4.

3 Syväoppimismenetelmät audiosyväväärennösten tunnistuksessa

Syvävääärennöstekniikat jättävät luomiinsa audiosignaaleihin jälkiä, joiden avulla keinotekoisesti luotu audio voidaan erottaa aidosta. Syväoppimisen tekniikoita voidaan käyttää näiden jälkien tunnistamiseen ja siten keinotekoisen puheen tunnistamiseen. Etuna syväoppimisen käytöllä tähän tehtävään on sen tehokkuus ja luotettavuus. [11] Yleisimmin syvävääärennösten tunnistukseen käytetyt syväoppimisalgoritmit ovat CNN, Pitkän aikavälin muisti (engl. Long Term Short Term Memory, LSTM) ja siirto-oppiminen (engl. Transfer Learning) [3]. Tässä tutkielmassa keskitytään CNN:ään, jota käsitellään tämän luvun lopussa. Ennen sitä käsitellään audion esitysmuotoja ja tunnistusmallien perusteita.

3.1 Audion esitysmuodot

Audio täytyy muuntaa toiseen muotoon ennen tunnistusmallille syöttöä, jotta aito ja synteettinen audio voidaan erottaa toisistaan [12]. Epäjohdonmukaisuudet missä tahansa audion esityksessä voi olla merkki synteettisestä audion luomisesta tai muutoksesta [13].

Mel-spektrogrammi (engl. Mel-spektrogrammi) on yksi tapa esittää audiota. Mel-spektrogrammit kuvaavat audiosignaaleja amplituudin, tiheyden ja ajan suhteen. Ne tuotetaan käyttämällä lyhytaikaista Fourier-muunnosta (engl. short term Fourier transform) aika-alueen audiosignaaliin ja muuttamalla tiheysarvot Hertzistä Meliin. [12]

Toinen tapa esittää audiota on sen spektrin avulla. Mel-taajuus kepstraalikertoimet (engl. Mel-Frequency Cepstral Coefficients, MFCC) esittävät audion spektriin liittyviä ominaisuuksia [13]. Lineaariset taajuuskepstraalikertoimet (engl. Linear Frequency Cepstral Coefficients, LFCC) sen sijaan korostavat suoraviivaisia spektriin liittyviä resoluutioita [14]. Spektrin keskipiste ja kaistanleveys ovat myös spektriin liittyviä esityksiä [13]. Spektrin keskipiste edustaa taajuusspektrin keskipistettä ja kaistanleveys keskipisteen ympärillä olevien taajuuksien määrää [13].

Audiota voi esittää myös muilla sen taajuuteen liittyvillä tavoilla. Vakio Q-muunnos (engl. Constant Q Transform, CQT) ja ikkunoitu fourier-muutos (engl. Short-Time Fourier Transform, STFT) esittävät audion taajuutta [15]. Lineaariset suodattimet (engl. linear filter, LF) sen sijaan esittävät vain tiettyjä eristettyjä taajuuskaistoja [16]. Nollakohtien ylitysten tahti (engl. Zero Crossing Rate, ZCR) esittää taajuutta, jolla audiosignaali vaihtaa suuntaansa x-akselilla positiivisesta negatiiviseen tai negativiisesta positiiviseen [13].

Spektrin ja taajuuden lisäksi on muitakin tapoja esittää audiota, kuten äänen sävyjen mukaan. Chromagrammi esittää eri äänensävyjen tai äänensävyluokkien esiintyvyyttä läpi ajan [13]. Chroma-STFT sen sijaan esittää audion harmonista sisältöä energian jakautumisella sävelluokkien välillä [14]. Gammatone-suodatin (GAM) keskittyy edustamaan ihmisen kuulemia audion muutoksia [16], kun taas laajennetut paikalliset kolmiarvoiset mallit (engl. Extended Local Ternary Patterns, ELTP) esittävät audion rakenteellisia ominaisuuksia [17].

3.2 Tunnistusmallien perusteet

Pohjimmiltaan syväväärennösten tunnistusmallit perustuvat tietoaineistoihin, esiprosessointiin, ominaisuuksien poimintaan sekä luokittelualgoritmiin. Tunnistusmalli koulutetaan tietoaineistolla, joka koostuu aidoista sekä väärennetyistä audiotallenteista. Tallenteet luokitellaan joko aidoiksi tai väärennetyiksi, jotta tunnistusmalli oppii erottamaan niiden välillä. [18], [19] Mitä laajempi ja monimuotoisempi tietoaineisto on, sen luotettavammaksi tunnistusmallia voidaan kehittää [18]. Lisäksi laajalla ja monimuotoisella tietoaineistolla koulutettu malli toimii paremmin koulutusaineiston ulkopuolella [18]. Esiprosessointivaiheessa tallenteista poistetaan epäolennaiset taustaäänet ja niitä voidaan laajentaa mallin yleistettävyyden parantamiseksi luomalla olemassa olevasta aineistosta uutta aineistoa muokkaamalla sitä hieman [19]. Muokkaamisen jälkeen raaka audio muutetaan audion esityksiksi, esimerkiksi mel-spektrogrammeiksi, ja ne muutetaan yhdenmukaisiksi pituudeltaan [19]. Esiprosessoinnin valmistuttua tietoaineisto syötetään mallille, joka opettelee luokittelemaan aineistoa [19].

3.3 Konvoluutioneuroverkot

Konvoluutioneuroverkot eli CNN:t ovat useista kerroksista koostuvia [3] täysin yhdistettyjä eteenpäin syöttäviä neuroverkkoja, joita käytetään oppimaan kuvamuotoisesta syötteestä ominaisuuksia ja luokittelemaan tietoa [20]. Jokaisella CNN:n kerroksella on oma tehtävänsä [3]. CNN:n syötekerros vastaanottaa syötteenä annetun kuvan [3]. Syötekerroksen jälkeinen konvoluutiokerros asettaa syötteeseen suodattimia ja erottaa siten kuvasta olennaisia kaavoja, jotka kuvataan ominaisuuskarttoina [3]. Sen jälkeen aktivaatiofunktiot luovat malliin epälineaarisuutta eli epäsuoraviivaisuutta mukauttamalla konvoluutiotasojen jokaisen neuronin tuotosta [3]. Epälineaarisuus mallissa on tärkeää, jotta malli voi oppia paremmin tunnistamaan syvä-

väärennöksiä ja käyttämään oppimaansa paremmin koulutusaineiston ulkopuolella [21]. Aktivaatiofunktioiden jälkeen yhdistämiskerros pienentää ominaisuuskarttojen kokoa valitsemalla arvojen joukosta yhden arvon [3]. Valittu arvo on maksimipoolauskerroksessa suurin arvo [22] ja keskiarvoituskerroksessa arvojen keskiarvo [12]. Lopulta täysin yhdistyneet kerrokset ja viimeiset kerrokset prosessoivat aikaisemmilta kerroksilta saatua tietoa ja tekevät CNN:n tehtävän mukaisen ennusteen annetusta syötteestä [3]. Esimerkkejä CNN-arkkitehtuureista ovat GoogLeNet, VGG-16 ja ResNet[4].

4 Tunnistusmallien tehokkuus

Tässä luvussa tarkastellaan audiosyväväärennösten tunnistamiseen käytettyjä ratkaisuja sekä tunnistusmallien tarkkuutta. Yksittäisten mallien tarkastelun jälkeen käsitellään konvoluutioneuroverkkojen käyttöä tunnistuksessa. Lopulta vertaillaan mallien tarkkuuksia ja esitetään katsauksen perusteella tehdyt johtopäätökset tunnistusmalleista.

4.1 Tutkimuskatsaus

Ahmadi ym. [17] ehdottavat artikkelissaan audiosyväväärennösten tunnistamiseen kaksisyötteistä CNN-mallia. Mallin perustana on VGG-16 eli Visuaalisen geometriaryhmän verkoston 16-kerroksinen malli [23]. Tunnistusmallin koulutukseen Ahmadi ym. käyttivät tietoaineistona Fake-or-Real- ja ASVspoof2019-aineistoja, jotka muutetaan mel-taajuus kepstraalikertoimiksi ja laajennetuiksi paikallisiksi kolmenkomponenttikuvioiksi. Näiden edustusten ominaisuudet yhdistetään ja VGG-16 erottaa niistä väärennetyt ominaisuudet. Ahmadi ym. saivat tunnistusmallin tarkkuudeksi 94,21 % hämmennysmatriisilla arvioiden käyttäen FoR-original- ja ASVspoof2019 A07 -tietoaineistoja arviointiin.

Myös Valente ym. [12] esittävät artikkelissaan VGG-verkkoihin perustuvaa CNN:ää. Heidän tunnistusmallinsa käyttää syötteenään mel-spektrogrammeja. Tunnistusmallin esiprosessointivaiheessa audiotallenteet muokataan samanpituisiksi lyhentämällä niitä tai poistamalla liian lyhyet tallenteet. Konvoluutiolohkoissa ensim-

mäisenä konvoluutiokerrokset käyttävät syötteeseen suodattimia, joilla poimitaan ominaisuuksia siitä. Sen jälkeen otoksen normalisointikerros varmistaa edustusten tasa-arvoisen aseman. Keskiarvoitus- ja pudotuskerrokset pienentävät edustuksen kokoa. Konvoluutiolohkon lopussa on toinen konvoluutiokerros ja toinen otoksen normalisointikerros. Konvoluutiosuodattimien määrä kasvaa lohkon loppua kohti, jotta tietoa saadaan enemmän. Viimeisten konvoluutiokerrosten jälkeen litistyskerros muuttaa datan vektorimuotoon monikerroksista perseptronia (engl. Multi Layer Perceptron) varten, joka luokittelee syötteen perusteella audion aidoksi tai väärennetyksi. Valente ym. mallin tarkkuus on 96,75 prosenttia keskiarvoitettuna kolmen tietoaineiston arvioinnin tuloksista.

Akmal ym. [24] ehdottavat artikkelissaan audiosyväväärennösten tunnistukseen Xception-mallia perustuen sen tehokkuuteen kuvan luokittelutehtävissä. Akmal ym. ovat vaihtaneet tunnistusmallinsa Xception-mallin viimeiset kerrokset binääriseen luokitteluun sopiviin kerroksiin. He alustavat tunnistusmallinsa esikoulutetuilla ImageNet-tietokannan painoarvoilla. Syötteenä heidän tunnistusmallinsa käyttää mel-spektrogrammeja, joista tunnistusmallin syvyyssuunnassa erotettavat konvoluutiot (engl. depthwise separable convolutions) erottavat audion avaruudelliset ja spektraaliset ominaisuudet. Sen jälkeen ominaisuuksien ulottuvuuksia pienennetään ylisovittamisen riskin vähentämiseksi ja ne muutetaan ominaisuusvektoreiksi. Täysin yhdistyneet kerrokset oppivat ominaisuusvektoreista ominaisuuksien välisiä vuorovaikutuksia, joiden jälkeen pudotuskerrokset vähentävät neuroverkon neuroneja ylisovituksen vähentämiseksi. Lopulta tiheä kerros arvioi onko syöte aito vai väärennetty ja muuttaa arvion binäärimuotoon. Akmal ym. saivat mallin tarkkuudeksi hämmennysmatriisi-analyysillä arvioiden 96 prosenttia.

Artikkelissaan Anagha ym. [22] esittelevät mel-spektrogrammeja syötteenään käyttävää CNN-mallia. He valmistelevat aineistoa luokittelemalla audiotallenteet binäärisesti aitojen ja väärennettyjen ryhmiin. Sen jälkeen audiotiedostoista erotetaan

mel-spektrogrammit, jotka muutetaan kiinteään pituuteen ja nimikoidaan aidoiksi tai väärennetyiksi. Mallin syötekerros vastaanottaa mel-spektrogrammit. Tämän jälkeen mallin kaksi konvoluutiokerrosta, joista toinen on 32-suodattiminen ja toinen 64-suodattiminen, käyttävät aineistoon suodattimia tunnistamaan paikallisia kaavoja ja ominaisuuksia. Kummankin konvoluutiokerroksen jälkeen maksimikoontikerrokset pienentävät aineiston kokoa. Konvoluutiokerroksia seuraa litistyskerros, joka muuttaa syötteen vektorimuotoon. Tiheä kerros yhdistää erotetut ominaisuudet ja siihen kuuluva ReLu (engl. Rectified Linear unit) -aktivaatio aiheuttaa malliin epälineaarisuutta. Sen jälkeen pudotuskerros tiivistää aineistoa. Mallissa on viimeisenä tiheä kerros, joka muuttaa todennäköisyysvektorit todennäköisyyksiksi ja luokittelee aineiston aidoksi tai vääräksi. Anagha ym. arvioivat tunnistusmalliaan hämmennysmatriisilla ja saivat tarkkuudeksi 85 prosenttia. Arviointiin käytettyä tietoaineistoa he eivät ole maininneet.

Krishnan ym. [14] esittelevät artikkelissaan Monen ominaisuuden audio-autenttisuusverkosto -mallin (engl. Multi-Feature Audio Authenticity Network, MFAAN). He käyttävät mallissa useita CNN-polkuja käsittelemään monia audioedustuksen malleja. Eri edustuksien käytöllä on hyötynä se, että ne huomioivat äänen eri ominaisuuksia paremmin ja yhdistettynä tunnistaminen voi tulla paremmaksi. Krishan ym. käyttävät tunnistusmallissaan audioedustuksina MFCC:tä, LFCC:tä ja ChromaSTFT:tä. Käsittelyn jälkeen ominaisuudet yhdistetään ketjutuksella ja syötetään tiheiden kerrosten läpi. Tiheät kerrokset erottavat ominaisuuksista kaavoja. Mallin lopussa on päätöksenteko-moduuli, joka koostuu ketjusta tiheitä kerroksia. Päätöksenteko-moduuli päättyy audion binäärisen luokittelun tulos -kerrokseen, joka tekee päätöksen audion aitoudesta. Krishnan ym. saivat mallin tarkkuudeksi In-the-Wild-tietoaineistolla arvioituna 99,21 prosenttia ja Fake-or-Real-tietoaineistolla arvioituna 94,47 prosenttia.

Dua ym. [25] esittävät artikkelissaan GFCC ja ResNet50 -tunnistusmallin, jossa audion ominaisuudet erotetaan käyttämällä GFCC-ominaisuuksia. Koulutusaineiston monimuotoisuuden lisäämiseksi Dua ym. loivat alkuperäisestä tietoaineistosta uutta aineistoa mukauttamalla puheen nopeutta sekä muuttamalla tai peittämällä audiosignaalin edustusta. Erotetut ominaisuudet syötetään esikoulutettuun ResNet50-malliin, joka luokittelee syötteen aidoksi tai väärennetyksi. Dua ym. korvasivat tehtävää varten ResNet-50-mallista sen viimeiset kerrokset uusilla luokitteluun sopivilla ja uudelleenkouluttivat ne uudella tietoaineistolla. Tunnistusmallin tarkkuudeksi he arvioivat ASVspoof2021-tietoaineistoa käyttäen 98 prosenttia.

Mahum ym. [26] ehdottavat artikkelissaan Ensemble Deep Learning Detector -nimistä tunnistusmallia, joka on keskittynyt tunnistamaan erityisesti TTS-syväväärennöksiä. He käyttävät audiotallenteiden edustamiseen mallissa mel-spektrogrammeja. Tunnistusmalli perustuu YAMNet:iin ja se on koulutettu ASVspoof2019-tietoaineistolla. YAMNet:in rinnalla Mahum ym. ovat hyödyntäneet mallissaan siirto-oppimista käyttämällä esikoulutettuja InceptionNetV2- ja ResNet50-malleja tunnistukseen. Malli tekee päätöksen äänen aitoudesta enemmistöperiaatteella kolmen mallin päätöksistä. Tällainen siirto-oppimista hyödyntävä malli saavuttaa heidän arvioimanaan ASVspoof2019 aineistoilla 99,70 prosentin tarkkuuden.

Lisäksi audiosyväväärennösten tunnistukseen on käytössä hybridialgoritmeja, joissa on yhdistelty eri syväoppimismalleja. Hybridimalleja on käsitelty esimerkiksi seuraavissa tutkimuksissa [12], [13], [16], [18], [19], [27], [28]. Taulukossa 4.1 vertaillaan sekä CNN-pohjaisia tunnistusmalleja että hybridialgoritmeja hyödyntäviä tunnistusmalleja sekä tunnistusmallien tarkkuuksia.

Taulukko 4.1: Tarkastellut aineistot

Tutkimus	**Ominaisuuksien edustus**	**Ominaisuuksien erotus**	**Tarkkuus**	**Arvioimiseen käytetty tietoaineisto**
Anagha ym., 2023 [22]	Mel-spektrogrammi	CNN	85,00 %	-
Ahmadi ym., 2024 [17]	MFCC & ELTP	VGG-16	94,21 %	for-original & ASVspoof2019 A07
Akmal ym., 2025 [24]	Mel-spektrogrammi	Xception	96,00 %	ASVspoof2021
Valente ym., 2024 [12]	Mel-spektrogrammi	CNN	96,75 %	Fake-or-Real, ASVspoof2019 & WaveFake
Krishnan ym., 2023 [14]	MFCC, LFCC & Chroma-STFT	CNN	94,47 % 99,21 %	Fake-or-Real In-the-Wild
Dua ym., 2023 [25]	GFCC-spektrogrammi	ResNet50	98,00 %	ASVspoof2021 DF
Mahum ym., 2023 [26]	Mel-spektrogrammi	YAMNet (+ResNet50 & InceptionNetV2)	99,70 %	ASVspoof2019 LA & ASVspoof2019 PA
Haluška ym., 2024 [13]	Spektrogrammi, MFCC, Spektrin keskipiste ja -kaistanleveys ZCR & Chromagrammi	CNN & RNN	79,72 %	-
Divya ym., 2024 [28]	Mel-spektrogrammi, MFCC, Spektrogrammi & Chromagrammi	CNN, RNN & LSTM	90,00 %	ADD2022
Pham ym., 2024 [16]	CQT, STFT&LF, STFT&GAM	CNN, ConvNeXt-Tiny & Whisper	90,00 %	ASVspoof2019 LA Evaluation
Chakravarty ym., 2024 [27]	Mel-spektrogrammi	ResNet-50, SVM, RF, KNN & NB	97,33 %	ASVspoof2019 LA, ASVspoof2019 PA & VIHL
Ali ym., 2024 [18]	Chroma, MFCC & ZCR	CNN, ConvLSTM, LSTM & RNN	98,00 %	Fake-or-Real
Gaikawad ym., 2025 [19]	Mel-spektrogrammi & LFCC	CNN & Transformer	98,07 %	ASVspoof2021, In-the-Wild & uusi kattava tietoaineisto

4.2 Konvoluutioneuroverkkomallit audiosyväväärennösten tunnistuksessa

Konvoluutioneuroverkkoihin perustuvat audiosyväväärennösten tunnistusmallit käyttävät audion esityksiin erilaisia suodattimia, joilla ne poimivat audion ominaisuuksia niistä [26]. CNN:ää käytetään siis erottamaan kuvamuotoisista audion esityksistä audion ominaisuudet. CNN-malli oppii erottamaan oleelliset ominaisuudet, kun sitä koulutetaan tarpeeksi tietoaineistoilla, joissa on sekä aidoiksi että väärennöksiksi luokiteltuja audiotallenteita. Lisäksi sitä voidaan käyttää tallenteiden luokittelemisessa aidoiksi tai väärennettyiksi. Artikkeleissa käytettyjä nimettyjä CNN-malleja ovat Resnet50, VGG-16, Xception ja InceptionNet V2.

4.3 Mallien tarkkuus

Mallien tarkkuuden vertailuun on käytetty mittaria, joka edustaa osuutta oikein luokitelluista audionäytteitä kaikista audionäytteistä. Tarkkuus on valittu mittariksi siksi, että se oli yleisin saatavilla oleva arviointimittari joukosta harkittuja aineistoja ja siten parhaiten vertailuun käytettävissä oleva mittari.

Audiosyväväärennösten tunnistamisen tarkkuus on nykyisen tutkimuksen valossa hyvällä tasolla. Kappaleessa 4.1 tarkastellut CNN perustaiset tunnistusmallit ovat tarkkuudeltaan keskiarvoisesti noin 95 prosenttia. Tunnistusmallien välillä on kuitenkin merkittäviä eroja, sillä tarkkuudeltaan heikoin on vain 85 prosenttia (Anagha ym. [22]) ja vahvin 99,70 prosenttia (Mahum ym. [26]). Tutkimuksissa esitetyistä malleista johdonmukaisesti tarkimpia olivat Mahum ym. [26] ja Dua ym. [25] esittämät CNN-mallit ja niissä molemmissa on molemmissa hyödynnetään siirto-oppimista.

Myös taulukon 4.1 hybridi-mallien tarkkuudet vaihtelevat suuresti. Tarkkuudeltaan heikoin on vain 79,72 prosenttia (Haluška ym. [13]), eli huomattavasti epätar-

kempi kuin heikoin CNN-malli (Anagha ym. [22]). Myöskään tarkin hybridi-malli (Gaikawad ym. [19]) ei yllä kahden tarkimman CNN-mallin tasolle, vaan jää reilulla prosenttiyksiköllä niistä vajaaksi. Tarkkuuden keskiarvoltaankin taulukon 4.1 hybridi-mallit jäävät 3 prosenttiyksikköä CNN-mallien keskiarvoa matalammaksi.

On kuitenkin huomioitava, että mallien tehokkuuden arviointiin on käytetty keskenään eri tietoaineistoja, joka saattaa aiheuttaa vaihtelevuutta tuloksissa. Tämä näkyy esimerkiksi Krishnan ym. [14] mallin vertailussa, jossa In-the-Wild -tietoaineistolla arvioituna tarkkuus on 99,21 prosenttia, mutta Fake-or-Real-tietoaineistolla arvioituna se on vain 94,47 prosenttia. Kaikissa tutkimuksissa ei myöskään ole mainittuna selkeästi mitä tietoaineistoa arviointiin on käytetty.

4.4 Johtopäätokset

Tulevaisuudessa olisi hyvä määritellä yksi mallien tehokkuuden arviointiin käytettävä mittari. Esimerkiksi hämmennysmatriisia käytetään yleisesti suorituksen arviointiin. Hämmennysmatriisin avulla voidaan laskea tarkkuus, täsmällisyys, herkkyys (engl. recall) sekä F1-arvo [20]. F1-arvo on täsmällisyyden ja herkkyyden harmoninen keskiarvo [29]. Nämä lasketaan käyttämällä aitoja positiiveja, aitoja negatiiveja, vääriä positiiveja ja vääriä negatiiveja [20]. Lisäksi tulisi olla käytössä yksi laaja ja monipuolinen arviointiin käytettävä tietoaineisto, joilla malleja voidaan arvioida, jotta arviointi ja vertailu olisi helpompaa sekä täsmällisempää.

Menetelmien suhteen huomataan, että useamman menetelmän hyödyntäminen tunnistuksessa ei suoraan tarkoita parempaa tarkkuutta. Mahum ym. [26] tunnistusmallissa on käytössä useampi menetelmä ja se on tarkkuudeltaan paras tunnistusmalli. Tunnistusmallin suorituskyvyn paremmuus voi kuitenkin johtua myös siirto-oppimisen paremmuudesta, sillä toisiksi tarkin Dua ym. [25] esittelemä CNN-malli hyödyntää myös siirto-oppimista.

Myöskään audion ominaisuuksien esitysmuodon vaikutuksesta syvävääreennösten tunnistuksen tarkkuuteen ei voida tehdä suoria päätelmiä, vaikka tarkastelluista malleista monet tarkimmat hyödyntävätkin useampaa eri esitysmuotoa. Yleisimmin käytetty esitysmuoto on mel-spektrogrammi, jota on käytetty sekä heikommin suoriutuvissa että vahvemmin suoriutuvissa malleissa. Malleja valitessa tulisikin katsoa myös niiden suorituksen mittareita, ei vain sitä, mitä menetelmiä niissä on käytetty. Lisäksi yksittäisten mallien tarkastelun sijaan tulisi tutkia koulutusaineiston esiprosessoinnin vaikutusta tarkkuuteen ja yleistettävyyteen.

Konvoluutioneuroverkkoja käytetään siis audiosyvävääreennösten tunnistuksessa kouluttamaan malleja erottamaan ominaisuuksia audion edustuksista ja luokittelemaan audiotallenteita niiden aitouden mukaan. Käytetyimpiä CNN-malleja ovat muun muassa VGG-16, Xception ja ResNet. Lisäksi CNN-pohjaisia malleja, kuten ResNet50- ja InceptionNetV2-malleja, käytetään tunnistusmallien siirto-oppimisessa, jossa valmista mallia käytetään hieman muokattuna uuteen tehtävään.

5 Yhteenveto

Audiosyväväärennökset ovat tekoälyn tekniikoilla käsiteltyjä tai luotuja audioita. Ne ovat suuri riski yksilöille ja yhteiskunnalle, sillä ne voivat levittää väärää tietoa ja niitä voidaan käyttää huijaustarkoituksiin. Audiosyväväärennösten tunnistamiseen tekoälyä käyttäen on monia mahdollisia tekniikoita.

Tässä tutkielmassa tarkasteltiin konvoluutioneuroverkkojen eli CNN:ien käyttöä osana audiosyväväärennösten tunnistamista sekä viimeaikaisessa tutkimuksessa esiteltyjen tutkimusmenetelmien tarkkuutta. Tutkielman toinen luku käsitteli audiosyväväärennöksiä sekä niiden luomista ja tunnistamista. Kolmannessa luvussa syvennyttiin audion esitysmuotoihin, syväväärennösten tunnistusmalleihin ja CNN:iin. Neljännessä luvussa käsiteltiin seitsemän tutkimuksen avulla CNN-pohjaisia tunnistusmalleja ja CNN:ien käyttöä niissä. Tämän jälkeen käsiteltiin näiden tunnistusmallien tarkkuutta ja vertailtiin niitä kuuden hybridi-mallin tarkkuuksiin. Luvun lopussa esitettiin johtopäätökset CNN-mallien nykytilanteesta ja niihin kohdistuvasta tutkimuksesta.

TK1. CNN-algoritmeja käytetään yleisimmin kahteen tarkoitukseen tunnistusmenetelmissä. Ensinnäkin CNN:iä käytetään erottamaan audiosta tunnistukselle olennaisia ominaisuuksia. Toiseksi niitä käytetään luokittelemaan audiot erotettujen audion ominaisuuksien perusteella aidoiksi tai väärennetyiksi. CNN:t voivat toimia yksin tai hybridi-algoritmina yhdessä toisten syväoppimisalgoritmien kanssa. Mallit voivat olla erityisesti tehtävää varten luotuja tai siirto-oppimismalleja, joissa val-

mista mallia muokataan tehtävään sopivaksi. Esimerkkejä yleisesti syvävääreennösten tunnistuksessa käytetyistä CNN-malleista ovat VGG-16, ResNet ja Xception. Siirto-oppimiseen käytettyjä malleja ovat erityisesti ResNet-50 ja InceptionNetV2.

TK2. Tunnistusmallien tarkkuuteen vaikuttaa käytetty audion esitysmuoto, tunnistusmallin koulutukseen käytetty tietoaineisto ja sen esiprosessointi sekä tunnistusmallissa käytetty algoritmi tai algoritmit. Tunnistusmallien tarkkuuden arviointi ja vertailu on tällä hetkellä epätarkkaa, sillä eri tutkimuksissa käytetään arviointiin eri tietoaineistoja ja eri arviointimittareita. Tässä tutkielmassa arvioitujen CNN-perustaisten tunnistusmallien keskiarvo on 95 prosenttia, joten tunnistusmallien tarkkuus on korkealla tasolla.

Hyvästä tarkkuudesta huolimatta tunnistusmallien tarkkuutta tulisi kehittää edelleen, jotta tunnistusmalleilla saataisiin johdonmukaisempia ja tarkempia tuloksia. Tulevassa tutkimuksessa tulisi esimerkiksi arvioida tunnistusmalleja yhdenmukaisin menetelmin sekä tietoaineiston että arviointimittojen suhteen. Tunnistusmallien koulutuksessa tulisi muistaa käytetyn tietoaineiston laajuuden ja monipuolisuuden tärkeys sekä myöskin ylisovituksen riskin vähentäminen. Tulisi myös tutkia tarkemmin, miten eri audion esitysmuotoa käyttävät tunnistusmallit eroavat tarkkuudeltaan. Lisäksi tunnistusmenetelmien yleistettävyyttä koulutusaineiston ulkopuoliseen käyttöön tulisi parantaa.

Lähdeluettelo

[1] A. Chadha, V. Kumar, S. Kashyap ja M. Gupta, "Deepfake: An Overview", *Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020*, s. 557–566, lokakuu 2020, Delhi, Intia, ISSN: 23673389. DOI: 10.1007/978-981-16-0733-2_39.

[2] M. v. Huijstee et al., "Tackling deepfakes in European policy", European Parliament, s. 1–4, 2021, Viitattu: 2025-11-05. url: https://op.europa.eu/en/publication-detail/-/publication/bd90819e-9b6c-11ec-83e1-01aa75ed71a1/language-en.

[3] D. Garg ja R. Gill, "Deepfake Generation and Detection - An Exploratory Study", *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, UPCON 2023*, s. 888–893, joulukuu 2023, Gautam Buddha Nagar, Intia. DOI: 10.1109/UPCON59197.2023.10434896.

[4] M. McUba, A. Singh, R. A. Ikuesan ja H. Venter, "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation", *Procedia Computer Science*, vol. 219, s. 211–219, tammikuu 2023, ISSN: 1877-0509. DOI: 10.1016/J.PROCS.2023.01.283.

[5] C. M. Tsai, Y. H. Tseng ja S. J. Ruan, "Deep Learning-Based Voice Production System for Deaf and Mute Individuals Combining Lip-reading and Text-to-Speech Technologies", *2024 International Automatic Control Conference,*

CACS 2024, s. 149–153, marraskuu 2024, Taoyuan, Taiwan. DOI: 10.1109/CACS63404.2024.10773205.

[6] K. Sudoh, T. Kano, S. Novitasari, T. Yanagita, S. Sakti ja S. Nakamura, "Simultaneous Speech-to-Speech Translation System with Neural Incremental ASR, MT, and TTS", s. 1–6, marraskuu 2020. DOI: 10.48550/arXiv.2011.04845.

[7] G. Bendiab, K. Zeltni, M. Bader-El-Den ja S. Shiaeles, "Audio-deepfake: Generation Methods, Legitimate Applications and the Potential for Misuse", *Proceedings of the 2025 IEEE International Conference on Cyber Security and Resilience, CSR 2025*, s. 50–56, elokuu 2025, Kreeta, Kreikka. DOI: 10.1109/CSR64739.2025.11130020.

[8] R. Chesney ja D. Citron, "Deepfakes and the new disinformation war", *Foreign Affairs*, tammikuu 2019, viitattu: 14. lokakuuta 2025. url: https://www.scopus.com/pages/publications/85074277064?inward.

[9] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao ja H. M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder", *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, s. 1–6, joulukuu 2016, Jeju, Korea. DOI: 10.1109/APSIPA.2016.7820786.

[10] A. Kumar, D. Singh, R. Jain, D. K. Jain, C. Gan ja X. Zhao, "Advances in DeepFake detection algorithms: Exploring fusion techniques in single and multi-modal approach", *Information Fusion*, vol. 118, nro 102993, kesäkuu 2025, ISSN: 1566-2535. DOI: 10.1016/J.INFFUS.2025.102993.

[11] S. S. Baraheem ja T. V. Nguyen, "AI vs. AI: Can AI Detect AI-Generated Images?", *Journal of Imaging*, vol. 9, nro 10, lokakuu 2023, ISSN: 2313433X. DOI: 10.3390/JIMAGING9100199.

[12] L. P. Valente, M. M. D. Souza ja A. M. Rocha, ”Speech Audio Deepfake Detection via Convolutional Neural Networks”, *IEEE Conference on Evolving and Adaptive Intelligent Systems*, s. 1–6, toukokuu 2024, Madrid, Espanja, ISSN: 24734691. DOI: 10.1109/EAIS58494.2024.10569111.

[13] R. Haluška, E. Kupcová, M. Pleva ja F. Hric, ”Automatic Speech Liveness Detection”, *ICETA 2024 - 22nd Year of International Conference on Emerging eLearning Technologies and Applications, Proceedings*, s. 168–173, joulukuu 2024, Stary Smokovec, Slovakia. DOI: 10.1109/ICETA63795.2024.10850831.

[14] K. S. Krishnan ja K. S. Krishnan, ”MFAAN: Unveiling Audio Deepfakes with a Multi-Feature Authenticity Network”, *2023 9th International Conference on Signal Processing and Communication, ICSC 2023*, s. 150–155, joulukuu 2023, Noida, Intia. DOI: 10.1109/ICSC60394.2023.10441405.

[15] S. D. H. Permana ja T. K. Rahman, ”Improved Feature Extraction for Sound Recognition Using Combined Constant-Q Transform (CQT) and Mel Spectrogram for CNN Input”, *Proceedings: ICMERALDA 2023 - International Conference on Modeling and E-Information Research, Artificial Learning and Digital Applications*, s. 185–190, marraskuu 2023, Karawang, Indonesia. DOI: 10.1109/ICMERALDA60125.2023.10458162.

[16] L. Pham, P. Lam, T. Nguyen, H. Nguyen ja A. Schindler, ”Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models”, *IEEE 5th International Symposium on the Internet of Sounds, IS2 2024*, s. 1–5, lokakuu 2024, Erlangen, Saksa. DOI: 10.1109/IS262782.2024.10704095.

[17] C. Ahmadi, S. H. Wang, S. P. Chiu ja J. L. Chen, ”Dual Acoustic Feature Fusion for Enhanced Audio Deepfake Detection Using VGG-16 Architecture: Mitigating Speech Tampering with MFCC and ELTP”, *Proceedings - 2024*

RIVF International Conference on Computing and Communication Technologies, RIVF 2024, s. 216–220, joulukuu 2024, Danang, Vietnam. DOI: 10.1109/RIVF64335.2024.11009070.

[18] G. Ali, J. Rashid, M. R. U. Hussnain, M. U. Tariq, A. Ghani ja D. Kwak, "Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection", *IEEE Access*, nro 12, s. 149940 –149 959, syyskuu 2024, ISSN: 21693536. DOI: 10.1109/ACCESS.2024.3457866.

[19] M. Gaikawad ja S. Ghosh, "A Robust and Lightweight CNN-Transformer Model for Audio Deepfake Detection in Indian Languages", *Proceedings of IEEE International Conference on Signal Processing,Computing and Control*, s. 382–387, maaliskuu 2025, Solan, Intia, ISSN: 26438615. DOI: 10.1109/ISPCC66872.2025.11039572.

[20] H. H. Aghdam ja E. J. Heravi, "Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification", *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*, s. 1–282, tammikuu 2017. DOI: 10.1007/978-3-319-57550-6.

[21] geeksforgeeks, *Activation functions in Neural Networks - GeeksforGeeks*, Viitattu: 03. marraskuuta 2025. url: https://www.geeksforgeeks.org/machine-learning/activation-functions-neural-networks/.

[22] R. Anagha, A. Arya, V. H. Narayan, S. Abhishek ja T. Anjali, "Audio Deepfake Detection Using Deep Learning", *Proceedings of the 2023 12th International Conference on System Modeling and Advancement in Research Trends, SMART 2023*, s. 176–181, joulukuu 2023, Moradabad, Intia. DOI: 10.1109/SMART59791.2023.10428163.

[23] K. Simonyan ja A. Zisserman, *Visual Geometry Group - University of Oxford*, Viitattu: 31. lokakuuta 2025. url: https://www.robots.ox.ac.uk/~vgg/research/very_deep/.

[24] D. Akmal ja V. Suryani, "Audio Deepfake Detection Using Xception Model", *2025 International Conference on Information and Communication Technology (ICoICT)*, s. 1–6, heinäkuu 2025, Bandung, Indonesia. DOI: 10.1109/ICOICT66265.2025.11193020.

[25] M. Dua, S. Meena, Neelam, Amisha ja N. Chakravarty, "Audio Deepfake Detection Using Data Augmented Graph Frequency Cepstral Coefficients", *2023 International Conference on System, Computation, Automation and Networking, ICSCAN 2023*, s. 1–6, marraskuu 2023, Puducherry, Intia. DOI: 10.1109/ICSCAN58655.2023.10395679.

[26] R. Mahum, A. Irtaza ja A. Javed, "EDL-Det: A Robust TTS Synthesis Detector Using VGG19-Based YAMNet and Ensemble Learning Block", *IEEE Access*, vol. 11, s. 134701–134 716, marraskuu 2023, ISSN: 21693536. DOI: 10.1109/ACCESS.2023.3332561.

[27] N. Chakravarty ja M. Dua, "Audio Spoof Detection using Deep Residual Networks based Feature Extraction: Unveiling Synthetic, Replay and Mimicry Threats", *2024 International Conference on Recent Innovation in Smart and Sustainable Technology, ICRISST 2024*, s. 1–5, maaliskuu 2024, Bengaluru, Intia. DOI: 10.1109/ICRISST59181.2024.10922028.

[28] K. Divya, G. Chhabra, Pallavi, S. A. Tiwaskar, S. Hemelatha ja V. Saraswat, "Application of Machine Learning for the Detection of Audio Deep Fake", *2024 Global Conference on Communications and Information Technologies, GCCIT 2024*, lokakuu 2024, Bangalore, Intia. DOI: 10.1109/GCCIT63234.2024.10862652.

[29] geeksforgeeks, *F1 Score in Machine Learning - GeeksforGeeks*, Viitattu: 12. marraskuuta 2025. url: https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/.