Leevi Härmä

TEKOÄLYN HYÖDYNTÄMINEN URHEILUTULOSTEN ENNUSTAMISESSA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2023

TIIVISTELMÄ

Härmä, Leevi
Tekoälyn hyödyntäminen urheilutulosten ennustamisessa
Jyväskylä: Jyväskylän yliopisto, 2023, 25 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja: Riekkinen, Janne

Urheiluotteluihin liittyvän tiedon määrä on kasvanut nopeaa vauhtia teknologian avulla. Tämä on mahdollistanut suurien tietomäärien käsittelyyn sopivien tekoälymallien soveltamisen urheiluotteluiden tulosten ennustamiseen. Alalla käytetään pääasiallisesti perinteisiä tilastotieteisiin perustuvia keinoja, mutta tekoälyn hyödyntämismahdollisuuksista tehdään aktiivista tutkimusta. Tässä tutkielmassa tarkasteltiin tekoälyn hyödyntämismahdollisuuksia urheilutulosten ennustamiseksi. Tutkielmassa rajattiin lähempään tarkasteluun kolme tekoälymallia. Nämä tekoälymallit olivat perinteiset neuroverkot, satunnaismetsät ja konvoluutioneuroverkot. Tutkielma toteutettiin kirjallisuuskatsauksena hyödyntäen alalla aikaisemmin tehtyjä tapaustutkimuksia, joissa sovellettiin erilaisia tekoälymalleja urheilutulosten ennustamiseksi. Tutkielman tarkoituksena oli selvittää, kuinka hyvin tekoälymallit sopivat urheilutulosten ennustamiseen. Asiaa arvioitiin tarkastelemalla tekoälymallien toimintaperiaatteita sekä tapaustutkimuksissa saavutettuja ennustustarkkuuksia. Tutkielman perusteella tekoälymallit soveltuvat erittäin hyvin urheiluotteluiden ennustamiseen, koska varsinkin joukkuelajeissa huomioonotettavia muuttujia on hyvin paljon. Tekoälymallien ennustustarkkuuksissa ei havaittu merkittäviä eroja, mutta opetukseen vaadittavien resurssien määrissä oli eroja. Konvoluutioneuroverkot pystyivät itse suorittamaan syötemuuttujien valintaa. Tämä on merkittävää, koska tutkielman perusteella syötemuuttujien valinta on yksi tärkeimmistä asioista tarkkojen ennustuksien saavuttamiseksi. Tutkielman perusteella tekoälymallien tuomat hyödyt keskittyvätkin toimintojen automatisointiin, mikä voi alentaa käyttäjiltä vaadittuja tietotaitoja urheilutulosten ennustamiseksi.

Asiasanat: tekoäly, neuroverkot, satunnaismetsä, konvoluutioneuroverkot, urheilutulosten ennustaminen

ABSTRACT

Härmä, Leevi
Utilizing artificial intelligence for predicting sports outcomes
Jyväskylä: University of Jyväskylä, 2023, 25 pp.
Information Systems, Bachelor's Thesis
Supervisor: Riekkinen, Janne

The amount of information related to sports matches has rapidly increased with technology. This has enabled the application of artificial intelligence models suitable for processing large amounts of data to predict sports match results. While the field predominantly employs traditional statistical methods, active research is being conducted to explore the possibilities of utilizing artificial intelligence. This thesis examined the potential of utilizing artificial intelligence to predict sports outcomes. Three artificial intelligence models were chosen for closer examination. These models were traditional neural networks, random forests, and convolutional neural networks. The thesis was conducted as a literature review, utilizing previous case studies in the field where different artificial intelligence models were applied for predicting sports results. The purpose of the thesis was to determine how well artificial intelligence models are suited for predicting sports outcomes. This was evaluated by examining the principles of the artificial intelligence models and the prediction accuracies achieved in the case studies. Based on the thesis, artificial intelligence models are highly suitable for predicting sports match outcomes, particularly in team sports where there are numerous variables to consider. Noteworthy differences in prediction accuracies among the artificial intelligence models were not observed, although there were some differences in the resources required for training. Convolutional neural networks were capable of performing feature selection on their own, which is significant because based on the thesis the selection of input variables is one of the key factors in achieving accurate predictions. Overall, the benefits brought by artificial intelligence models focus on automating processes, which can reduce the expertise required from users for predicting sports outcomes.

Keywords: artificial intelligence, neural networks, random forest, convolutional neural networks, sports prediction

KUVIOT

SISÄLLYS

1 JOHDANTO

Internetin kasvun myötä urheiluotteluista saatavasta datasta on tullut yhä julkisempaa ja sen määrä sekä laajuus on kasvanut erittäin paljon. Tämän kehityksen myötä urheilulajien analysoiminen datalähtöisten keinojen avulla on tullut mahdolliseksi käytännössä jokaiselle tietokoneen omistavalle henkilölle. Näiden keinojen avulla voidaan analysoida ja laskea useita urheiluun liittyviä tekijöitä, kuten pelaajan todellisen potentiaalin havaitseminen tai yksittäisen urheilutapahtuman eri lopputuloksien tapahtumatodennäköisyyksien selvittäminen (Huang & Li, 2021).

Yleisin keino urheilutulosten ennustamiseksi on perustaa ennustukset subjektiivisiin mielipiteisiin ja tuntemuksiin (Leung & Joseph, 2014). Tällaista toimintaa harrastavat useat katsojat ja sen hyötynä on pääasiallisesti viihdearvon nostattaminen. Kun halutaan tavoitella tarkempia ennustuksia, niin päättely tulisi perustaa ennustusta tukeviin tekijöihin. Tilastot toimivat usein erinomaisesti ennustusten tukena, kunhan niitä hyödynnetään järkevästi. Wilkens (2021) huomauttaakin, että perinteiset tilastolähtöiset lähestymistavat ovat edelleen eniten käytettyjä urheilutulosten ennustamiseksi, kun tavoitteena on mahdollisimman tarkkojen ennustusten luonti. Tekoälyn hyödyntäminen on kuitenkin alkanut yleistymään ja sen mahdollisista hyödyntämiskeinoista tehdään aktiivisesti tutkimusta (Kollár, 2021).

Tutkielmassa esitellään erilaisia tekoälyyn perustuvia tekniikoita, joiden avulla urheilutuloksia ennustavia malleja pystytään luomaan. Tekoäly ymmärretään tutkielmassa Haenlein ja Kaplanin (2019) määritelmän mukaisesti, eli kykynä tulkita ulkopuolista tietoa ja sen avulla oppia ratkaisemaan ennalta määriteltyjä tehtäviä. Täten tutkielmassa esiteltävät tekoälymallit voidaan määritellä systeemeiksi, jotka toteuttavat edellä mainittua määritelmää. Näiden tekoälymallien peruslähtökohtana toimii suuren aineistomäärän käsitteleminen ja siitä järkevien johtopäätösten tekeminen. Tästä syystä tekoäly saattaa olla hyvä keino urheilutulosten ennustamiseksi, sillä urheiluotteluiden lopputulemaan vaikuttaa hyvin monet erilaiset pienet tekijät, joiden käsitteleminen kattavasti on käytännössä mahdotonta ihmiselle (Fialho ym., 2019).

Näiden tekoälymallien onnistunut toteuttaminen ei kuitenkaan ole yksinkertaista ja erilaisia mahdollisuuksia käytännön toteutukseen on useampia. Tämän tutkielman yhteydessä ei pystytä käsittelemään näitä kaikkia vaan tarkasteluun rajataan kolme yleistä tekoälymallia. Nämä ovat keinotekoiset neuroverkot, satunnaismetsät sekä konvoluutioneuroverkot. Tutkielman tavoitteena on esitellä näiden tekoälymallien toimintaperiaatteet sekä selvittää, miten niitä on kirjallisuudessa sovellettu urheilutulosten ennustamiseksi. Tutkielmasta on hyötyä kaikille, ketkä haluavat tutustua edellä mainittujen tekoälymallien hyödyntämismahdollisuuksiin urheilutulosten ennustamisessa. Tutkielma antaa hyvän yleiskatsauksen kirjallisuuteen, minkä avulla lukija pystyy itse valikoimaan omaan tehtävään parhaiten soveltuvat toteutukset. Varsinaisiksi tutkimuskysymyksiksi muodostui:

- Miten tekoälyä on hyödynnetty urheilutulosten ennustamisessa?
- Kuinka tarkasti tekoäly pystyy ennustamaan urheilutuloksia?

Tutkielma toteutettiin kirjallisuuskatsauksena hyödyntäen aiheesta tehtyjä tieteellisiä tapaustutkimuksia. Tämän myötä lähdeaineistona käytettiin pääasiallisesti vertaisarvioituja artikkeleita. Näiden artikkelien löytämiseksi hyödynnettiin Google Scholar- ja JYKDOK- hakupalveluita. Hakusanoina käytettiin enimmäkseen englanninkielisiä termejä, joihin sisältyi esimerkiksi seuraavat termit: artificial intelligence, sports prediction, neural networks ja AI models. Tutkielman aihe on hyvin ajankohtainen, mikä näkyy myös alan kirjallisuudessa tuoreiden tutkimusten määrässä. Tämän ansiosta tapaustutkimuksen sisältävät artikkelit ovat hyvin uusia, mutta tutkielmassa hyödynnettiin myös vanhempaa lähdekirjallisuutta tekoälymallien toimintaperiaatteista puhuttaessa.

Tutkielma koostuu johdannosta, kolmesta sisältöluvusta ja yhteenvetoluvusta. Ensimmäisessä sisältöluvussa rajataan ja käsitellään tutkimukseen tarkemmin mukaan otettavat tekoälymallit. Nämä ovat keinotekoiset neuroverkot, konvoluutioneuroverkot sekä satunnaismetsät. Toisessa sisältöluvussa määritellään erilaiset urheilutulokset ja tarkastellaan yleisimpiä lähestymistapoja urheilutulosten ennustamiseksi. Kolmannessa sisältöluvussa käydään aluksi läpi eri tutkimuksien välisten tulosten vertailuun vaikuttavia tekijöitä. Tämän jälkeen tarkastellaan oikeanlaisten syötemuuttujien vaikutusta tekoälymallien toimintaan. Viimeisenä asiana kolmannessa sisältöluvussa on eri tutkimuksien saavuttamien tarkkuuksien vertaileminen.

2 TEKOÄLYN MENETELMÄT

Tekoälyä hyödynnetään nykyään hyvin monilla eri yhteiskunnan osa-alueilla ja sen mahdollisista uusista sovelluskeinoista tehdään jatkuvasti tutkimustyötä. Yleinen määritelmä tekoälylle on Haenlein ja Kaplanin (2019) mukaan kyky tulkita ulkopuolista tietoa ja sen avulla oppia ratkaisemaan ennalta määriteltyjä tehtäviä. Tekoälymallit voidaan siten ajatella algoritmien varaan rakennettuina systeemeinä, jotka toteuttavat edellä mainitun määritelmän. Tämän tutkielman tarkoituksena ei ole syventyä tekoälyn historiaan tai luonteeseen yleisellä tasolla, vaan tarkastellaan tekoälymallien soveltamista tietyssä rajatussa kontekstissa. Tässä luvussa ensimmäiseksi rajataan tarkempaan tarkasteluun tulevat tekoälymallit. Tämän jälkeen syvennytään tarkemmin näihin käyden läpi niiden perusteet ja toiminnallisuudet.

2.1 Menetelmien rajaus

Tekoälymalleja on aktiivisesti käytössä useita erilaisia eri tarkoituksissa. Nämä mallit eroavat toiminnallisuuksiltaan toisistaan ja tästä syystä tietyt mallit sopivat tietynlaisiin tehtäviin paremmin kuin toiset. Näille kaikille erilaisille malleille yhteistä on kuitenkin se, että ne toimivat kuten mustat laatikot. Eli ne piilottavat osan toiminnastaan sisälleen siten, että käyttäjän saattaa olla vaikeaa päästä tarkastelemaan sitä (Setzu ym., 2021).

Urheilutulosten ennustamiseen on sovellettu monia erilaisia tekoälyn malleja, joten tässä tutkielmassa ei pystytä näitä kaikkia tarkastelemaan. Tästä syystä rajataan lähempään tarkasteluun kolme mallia. Nämä mallit ovat keinotekoiset neuroverkot, satunnaismetsät ja konvoluutioneuroverkot. Näistä kahden ensimmäisen valikoitumisen syynä on niiden suuri yleisyys alan kirjallisuudessa. Nämä mallit myös soveltuvat urheilutulosten ennustamiseen erityisen hyvin ja niiden avulla saadut tulokset ovat usein tutkimuksissa parhaimpia. Tutkielman kolmanneksi malliksi valitaan konvoluutioneuroverkot. Niistä ei vielä ole tarjolla

kovinkaan paljon kirjallisuutta varsinkaan urheilun yhteydessä, mutta ne ovat monella alalla alkaneet saamaan suosiota yli perinteisempien tekoälymallien.

Näille kaikille tekoälymalleille yhteistä on kuitenkin niiden opetustavat. Oikeanlaisen opetustavan valitseminen on erittäin tärkeässä roolissa tarkkojen tulosten saavuttamisessa. Tätä varten onkin kehitetty useita erilaisia tapoja toteuttaa opetus, mutta ne voidaan käytännössä jakaa kahteen kategoriaan eli ohjattuun sekä ohjaamattomaan opetukseen. Agatonovic-Kustrin ja Beresford (2000) mukaan ohjatussa opetuksessa on tiedossa halutut lopputulosteet, kun taas ohjaamattomassa opetuksessa ei anneta haluttuja lopputulosteita, jotta tekoälymalli itse löytäisi annetusta datasta oleelliset piirteet ja onnistuisi täten tuottamaan järkeviä tulosteita. Näistä ohjattu opetus sopii huomattavasti paremmin sellaisille tekoälymalleille, joiden tavoitteena on luoda realistiset tapahtumatodennäköisyydet luokitelluille vaihtoehdoille. Ohjaamaton opetus sen sijaan on parempi laajojen ja monimutkaisten datojen jäsentelyssä (Agatonovic-Kustrin & Beresford, 2000).

2.2 Keinotekoiset neuroverkot

Keinotekoiset neuroverkot (engl. *artificial neural networks*) ovat viimevuosina nousseet yhdeksi tehokkaimmista vaihtoehdoista monien erilaisten ongelmien ratkaisemiseksi. Tämän kehityksen taustalla on yleisesti parantunut ymmärrys neuroverkkojen suunnittelusta ja algoritmeista (Vasilev, Slater, Spacagna, Roelants & Zocca, 2019).

Kollár (2021) toteaa keinotekoisten neuroverkkojen toiminnan perustuvan yksittäisiin neuroneihin ja näiden välisiin linkkeihin. Hän kuvaa, miten yksittäinen neuroni vastaanottaa syötteitä useista eri neuroneista ja toteuttaa näiden syötteiden perusteella oman toimintansa. Tämän toiminnan tuloksena syntyy neuronin tuottama tuloste, joka lähetetään eteenpäin linkkien kautta seuraaville neuroneille. Nämä linkit eivät kuitenkaan toimi pelkästään tiedon kuljettajina, vaan ne osallistuvat myös tiedon käsittelyyn. Tämä tapahtuu linkeille määriteltyjen vahvuuslukemien avulla (Kollár, 2021).

Vasilev ym. (2019) kuvaavat tarkemmin tiedon käsittelyä neuroneiden välillä. He avaavat linkkien vahvuuslukujen ja neuroneiden sisältämien aktivointifunktioiden yhteyttä. Ensin neuronin vastaanottamat arvot summataan yhteen ottaen huomioon linkkien vahvuusluvut. Tämän jälkeen lasketaan neuronille määritellyn aktivointifunktion arvo lasketun summan avulla ja saatu arvo toimii neuronin tulosteena. Aktivointifunktioina käytetään epälineaarisia funktioita, joista yleisimpiä ovat Sigmoid- ja ReLu- funktio (Vasilev ym., 2019).

Vasilev ym. (2019) mukaan neuroverkot siis koostuvat eri kerroksista, jotka sisältävät toisiinsa linkittyneitä neuroneita. Jokainen neuroverkko sisältää vähintään syötekerroksen ja ulostulokerroksen. Näiden lisäksi käytetään piilotettuja kerroksia, jotta neuroverkkoon saadaan lisättyä syvyyttä. Tämä johtuu siitä, että saman kerroksen sisällä olevat neuronit käyttävät yleensä samanlaisia aktivointifunktioita, joten kerroksia lisäämällä saadaan sisällytettyä useita erilaisia

aktivointifunktioita (Vasilev ym., 2019). Agatonovic-Kustrin ja Beresford (2000) kuvailevat neuroverkkojen erilaisia arkkitehtuureja. Neuroverkolle valittu arkkitehtuuri määrittelee sen, miten eri kerrokset kommunikoivat toistensa kanssa. Useimmiten kerrosten tuottamat tulosteet kulkeutuvat vain eteenpäin seuraavalle kerrokselle, mutta joissakin arkkitehtuureissa kerroksen tuottamat tulosteet kulkevat myös taaksepäin (Agatonovic-Kustrin & Beresford, 2000).

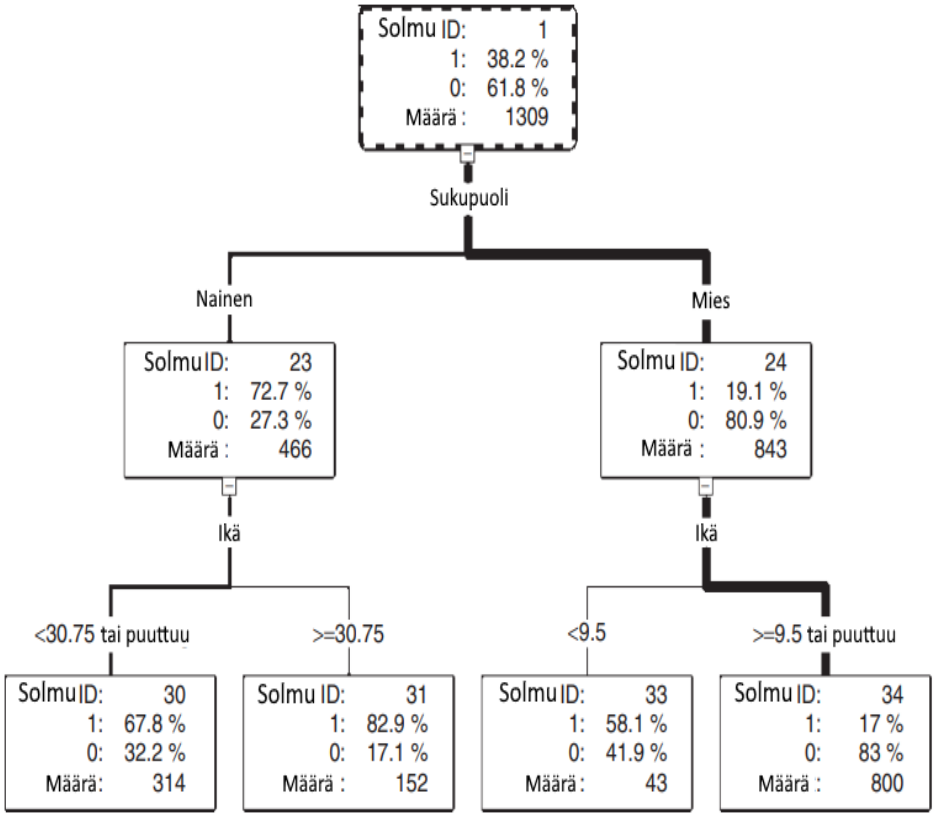
Vasilevin ym. (2019) mukaan neuroverkkojen opetukseen sovelletaan usein ohjattua oppimista eli tavoiteltavat lopputulosteet ovat tiedossa etukäteen. Neuroverkot ovat pääasiassa monikerroksisia, jolloin niissä syntyy useiden välivaiheiden tulosteita, joiden tavoitearvot eivät ole tiedossa. Tästä syystä opetusta täytyy lähteä tekemään ulostulokerroksesta taaksepäin, koska sieltä tulevien tulosteiden halutut arvot ovat tiedossa. Tässä prosessissa yleisin käytetty algoritmi on vastavirta-algoritmi (Vasilev ym., 2019). Vastavirta-algoritmin avulla pystytään laskemaan yksittäisten neuroneiden ja niistä muodostuvien kerrosten vaikutukset virhearvioiden suuruuteen. Tämän ansiosta on mahdollista muokata neuroneiden välisten linkkien vahvuuslukuja oikeaan suuntaa, jotta virhearvioiden suuruus vähenee. Tätä prosessia toistamalla neuroverkot voidaan opettaa tekemään tarkkoja ennusteita. On kuitenkin huolehdittava, ettei vahingossa liiallisella opettamisella poista neuroverkon kykyä toimia erilaisten syötedatojen kanssa (Agatonovic-Kustrin & Beresford, 2000). Candila ja Palazzo (2020) kuvaavat tyypilliseksi ongelmaksi mahdollisen kohinan opetusdatassa mikä saattaa johtaa siihen, että neuroverkko menettää kykynsä yleistää oppimaansa tietoa.

2.3 Satunnaismetsä

Breiman (2001) tiivistää satunnaismetsien (engl. *random forests*) olevan yksittäisistä päätöspuista muodostuvia kokonaisuuksia. Niiden lähtökohtana on rakentaa useita hieman toisistaan eroavia päätöspuita ja valita loppusyötteeksi yksittäisistä päätöspuista saatujen tulosten joukosta yleisin. Tavoitteena on vähentää päätöspuun toimintaan liittyvän satunnaisuuden merkitystä (Breiman, 2001). Käydään seuraavaksi ensin yksittäisen päätöspuun toiminta läpi, minkä jälkeen tarkastellaan satunnaismetsien muodostamista.

De Ville (2013) kuvaa päätöspuun koostuvan eri tasoista, jotka usein kuvataan ylhäältä alaspäin kulkevina solmuina. Kuvio 1 kuvaa yksinkertaisen päätöspuun rakenteen malliesimerkin avulla. Ylimmällä tasolla on yksi solmu, jossa kohdedata on jakautunut yhden muuttujan perusteella. Seuraavilla tasoilla jaetaan aina ylemmällä tasolla muodostuneet solmut uuden muuttujan perusteella. Lohin (2011) mukaan päätöspuu voi toimia sekä luokittelumenetelmänä että regressiomenetelmänä riippuen siitä, että käytetäänkö solmujen jakamiseen diskreettejä vai jatkuvia muuttujia. Tasoilla käytettävien muuttujien valintaan on olemassa useita erilaisia algoritmeja (Loh, 2011; de Ville, 2013). Yhtenä lähtökohtana voidaan esimerkiksi käyttää mahdollisimman paljon erottelua luovaa syötettä (de Ville, 2013). Päätöspuun toiminta loppuu, kun se saavuttaa jonkin

ennalta määrätyn lopetusehdon. Lopetusehtona voi toimia esimerkiksi vähimmäisvaatimus yhdessä solmussa olevien havaintojen määrälle (de Ville, 2013).



KUVIO 1 Titanicin uppoamisen selviytymisanalyysin havainnollistava päätöspuu (de Ville, 2013)

Satunnaismetsät ovat siis useista erillisistä päätöspuista muodostuvia kokonaisuuksia, joissa tulokseksi valitaan päätöspuiden yleisin tuloste (Breiman, 2001). Suuresta määrästä päätöspuita ei kuitenkaan ole hyötyä, jos ne toimivat samalla tavalla. Tästä syystä satunnaismetsässä käytettäviä päätöspuita pyritään erilaistamaan. Breiman (2001) kuvaa tarkemmin yleisesti käytettyä ideaa päätöspuiden erilaistamiseksi. Siinä ensin satunnaistetaan päätöspuille käyttöön päätyvät muuttujat suorittamalla bootstrap-aggregointi alkuperäiselle muuttujajoukolle. Tämä tarkoittaa sitä, että jokaiselle päätöspuulle tulee käyttöön sama määrä muuttujia, mutta ne sisältävät vain osan alkuperäisistä muuttujista, sillä uudessa joukossa sallitaan samojen muuttujien toistuminen. Tämän jälkeen valitaan satunnaisesti jokaiselle päätöspuulle annetuista muuttujista vain osa käytettäviksi päätöspuun rakentamiseksi (Breiman, 2001).

2.4 Konvoluutioneuroverkko

Goodfellow ym. (2016) mukaan konvoluutioneuroverkko (engl. *convolutional neural network*) perustuu aikaisemmin esitettyyn perinteiseen keinotekoiseen neuroverkkoon, mutta konvoluutioneuroverkko sisältää vähintään yhden konvoluutiokerroksen (Goodfellow ym., 2016). Konvoluutioneuroverkko eroaa perinteisestä keinotekoisesta neuroverkosta usealla tavalla. Konvoluutioneuroverkossa neuronit linkittyvät hieman eri tavalla. Vasilev ym. (2019) mukaan neuronit vastaanottavat informaatiota niiden läheisiltä neuroneilta sen sijaan, että ne saisivat syötteensä kaikilta aikaisemman kerroksen neuroneilta. Tämän lisäksi neuronit jakavat osan parametreistaan muiden saman kerroksen neuroneiden kanssa. Kerrosten aktivointifunktioiden määrittely toimii puolestaan hyvin samankaltaisesti kuin perinteisissä neuroverkoissa (Vasilev ym., 2019).

Konvoluutioneuroverkot pystyvät käsittelemään aineistoja useammassa eri muodossa, kuten kaksiulotteisia kuvia tai kolmiulotteisia videoita (Huang & Li, 2021). Tämä pitää ottaa huomioon verkon suunnitteluvaiheessa, mutta näiden ero on hyvin pieni. Huang & Li (2021) kuvaavat, miten yksiulotteisessa konvoluutioverkossa suodatin liikkuu vain yhteen suuntaan ja syötteet, sekä tulokset ovat kaksiulotteisia. Kun taas kaksiulotteisessa konvoluutioverkossa suodatin liikkuu kahteen suuntaan, jolloin syöte ja tulostedata ovat kolmiulotteisia (Huang & Li, 2021).

Yleisimmin konvoluutioneuroverkkoja on hyödynnetty kuvien käsittelyssä eri tieteenaloilla. Sitä on yritetty soveltaa lääketieteessä CT-kuvauksesta saatavien kuvien analysointiin ja luokitteluun. Ibragimov ja Xing (2017) onnistuivat tunnistamaan CT-kuvista elimiä, jotka saattaisivat kärsiä säteilyhoidon aloittamisesta.

3 URHEILUTULOSTEN ENNUSTAMINEN

Urheilu ja sen seuraaminen on ollut osa ihmishistoriaa hyvin pitkän aikaa, mutta varsinkin digitalisaation myötä eri lajien otteluiden seuranta on muuttunut hyvin helpoksi. Tämän lisäksi samaan aikaan otteluiden ennustamiseen liittyvät harrastukset ovat lisääntyneet. Monien kilpasarjojen rinnalla pyöritetään ns. fantasialiigaa, joissa osallistujat pääsevät kilpailemaan toisiaan vastaan esimerkiksi valitsemalla pelaajia, joiden uskoo pärjäävän hyvin tulevissa otteluissa. Tämän lisäksi muun muassa urheiluvedonlyöntimarkkinat jatkavat kasvamistaan. Näiden lisäksi urheiluotteluiden ennustaminen on nykyään myös usein näkyvillä urheilulähetyksissä. Houghton ym. (2019) kuvaavat miten tämä näkyy Yhdysvalloissa valtamedioiden urheilulähetyksissä. Heidän mukaansa lähetyksissä näytetään luotuja ennustuksia muun muassa otteluohjelmien sekä tulostaulukoiden yhteydessä. Jotkut urheilulähetyksiä lähettävät mediaorganisaatiot ylläpitävät myös omia tilastosivuja, jotka antavat tavalliselle katsojalle hieman ideoita ja tilastoja huomioitavaksi (Houghton ym., 2019). Tässä luvussa määritellään ensin urheilutulokset ja tarkastellaan lyhyesti sen erilaisia tyyppejä. Tämän jälkeen käydään vielä läpi yleisimmät lähestymistavat, joita urheilutulosten ennustamiseen sovelletaan.

3.1 Urheilutulosten määritelmä

Urheiluottelut sisältävät useampia erilaisia tuloksia, joista yleisimmin puhuttuja ovat todennäköisesti ottelun mahdolliset voittajat sekä häviäjät. Tämän lisäksi toinen tärkeä tulos on usein lopulliset määreet, kuten maalien määrä tai suorituksen kesto. Nämä vaihtelevat suuresti urheilulajien välillä, mutta ne voidaan kuitenkin pääosin luokitella kahteen tyyppiin.

Tulokset voivat olla luokitteluja tai numeraalisia määreitä. Tämä on oleellista, kun valitaan tapaa niiden ennustamiseksi. Luokitteluongelmissa halutaan määritellä muuttujat tiettyihin luokkiin. Luokat on ennalta määritelty ja ongelmana on sen selvittäminen mihin luokkaan jokin uusi arvo kuuluu (Nagy, 2018).

Yleensä urheiluottelun lopputulosta ennustettaessa lähestytään tilannetta luokitteluongelmana, jossa ennustettavat luokat ovat voitto, tasapeli sekä häviö (Bunker & Thabtah, 2019).

Toisaalta numeraalisten tulosten ennustamista lähestytään yleisesti regressiomenetelmien avulla. Nagyn (2018) mukaan tyypillinen ongelma, johon regressiomenetelmät sopivat on sellainen, missä tiedetään yhden muuttujan arvo ja halutaan selvittää siihen liittyvän toisen muuttujan arvo. Ongelmaksi muodostuu todellisissa tilanteissa se, että useimpiin muuttujiin vaikuttaa monet asiat, joten yhden muuttujan tietämisellä ei voida välttämättä päätellä sen hetkisen toisen muuttujan arvoa (Nagy, 2018).

Regressiomenetelmien yhteydessä käytettävät syötemuuttujat ovat siis rationaalilukuja, kun taas luokitteluongelmissa ne ovat luokkia. Tästä syystä käytettävät syötearvot pitää skaalata samalle tasolle. Tämä tapahtuu yleensä käyttäen jotakin normalisointifunktiota, jonka avulla kaikki syötemuuttujat saadaan arvojen -1 ja 1 välille (Nagy, 2018).

3.2 Erilaiset lähestymistavat

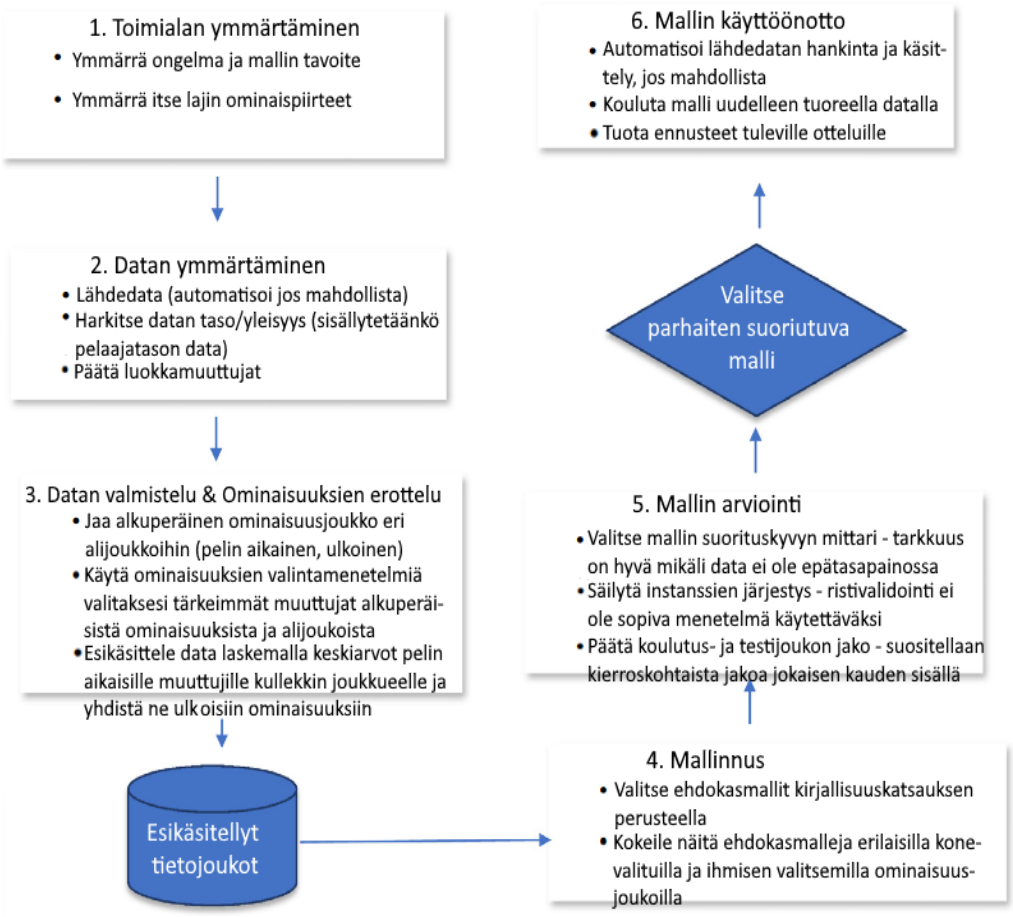
Kyky urheiluotteluiden tuloksien ennustamiseen on oleellinen taito monille eri tahoille eri syistä. Osalle tärkein asia on todenmukaisten ennustettujen tulosten hyödyntäminen esimerkiksi vedonlyönnissä. Toisaalta joukkueenjohtaja saattaa olla kiinnostunut aiheesta enemmän tuloksiin johtavien syiden selvittämisen näkökulmasta. Huang ja Li (2021) nostavat jälkimmäisestä esimerkiksi Michael Lewisin vuonna 2003 julkaiseman kirjan ”Moneyball: The Art of Winning an Unfair Game”, joka heidän mukaansa kuvaa hyvin tilastollisten elementtien hyödyntämistä parhaimman lopputuloksen saavuttamiseksi. Nykyään käytetään valtaosin tilastollisia tietoja urheiluotteluiden analysoimiseksi, mutta ennen digitalisaatiota tämä oli hyvin haastavaa (Huang & Li, 2021).

Siitä huolimatta, että monet urheiluorganisaatiot ovat sisällyttäneet tilastolliset lähestymistavat toimintaansa, niin valtaosa urheilua seuraavista yksilöistä tekevät edelleen ennustuksiansa subjektiivisten mielipiteiden pohjalta. Leung ja Joseph (2014) kuvaavat, miten urheilutuloksia ennustetaan hyvin monesti henkilökohtaisten näkemysten ja tunteiden pohjalta. Tämä näkyy heidän mukaansa usein muun muassa urheiluotteluita edeltävissä lähetyksissä, joissa lajin asiantuntijat antavat omia ennustuksiaan tulevan ottelun kulusta ja lopputuloksesta. Tällainen ennustaminen on kuitenkin usein todella virheherkkää (Leung & Joseph, 2014).

Butler, Butler ja Eakins (2021) tutkivat yksilöiden tekemiä ennustuksia koskien otteluiden voittajaa ja lopullista tulosta jalkapallon valioliigan yhteydessä. Subjektiivisen arviointitavan virheherkkyys korostui tutkimuksen mukaan lopullisten maalimäärien arvioissa. Yksilöt ennustivat huomattavasti vähemmän pienimaalisia tuloksia, kuin mitä niitä todellisuudessa tuli. Tämän taustalla he näkivät sen, että katsojat haluavat mieluummin nähdä suurimaalisia otteluita (Butler, Butler & Eakins, 2021).

Näistä syistä johtuen on suositeltavaa pyrkiä perustamaan ennustuksensa myös tilastoihin. Tilastollisten keinojen lisäksi toimintaa voi myös tehostaa ja automatisoida hyödyntämällä tekoälyä tilastojen tulkinnassa. Tämä tarkoittaa käytännössä tekoälymallien luomista. Thabtah ym. (2019) kuvaa urheiluotteluiden ennustamiseen tarkoitettujen mallien strategian olevan usein samankaltainen. Ensin selvitetään käytettävässä olevista muuttujista lopputulokseen eniten positiivisesti vaikuttavat. Tämän jälkeen käytetään oleellisiksi todettuja syötemuuttujia mallin opettamiseen. Kun malli on opetettu, sille voidaan antaa ennustettavaan otteluun liittyvät muuttujat syötteinä, ja se pystyy opetusdatan ansiosta ennustamaan kyseisille muuttujille todennäköisimmät tulosteet (Thabtah ym., 2019).

Bunker ja Thabtah (2019) kehittivät artikkelissaan oman viitekehyksen tekoälymallien kehittämiseksi. He käyttivät viitekehyksen pohjana tiedonlouhinnassa yleisesti tunnettua prosessimallia nimeltä CRISP-DM. Kuvio 2 sisältää heidän kehittämän viitekehyksen vaiheet sekä lyhyen kuvauksen vaiheen tärkeimmistä tehtävistä.



KUVIO 2 SRP-CRISP-DM Viitekehys tekoälymallien kehittämiseen (Bunker & Thabtah, 2019)

4 TEKOÄLY ENNUSTAMISESSA

Tekoälyn soveltamisesta urheilutulosten ennustamiseksi on tehty paljon kirjallisuutta. Suurin osa näistä on tapaustutkimuksia, joissa kirjoittajat ovat kehittäneet itse aikaisempaan tietoon perustuvia tekoälymalleja. Näiden perusteella pitäisi olla mahdollista saada hyvä yleiskuva tekoälymallien mahdollisuuksista ja rajoituksista. Tässä luvussa käydään aluksi läpi muutamia asioita, joita pitää ottaa huomioon eri tapaustutkimuksien tuloksia vertailtaessa. Toisena asiana on tarkastella millaisia vaikutuksia syötemuuttujien oikeanlaisella valitsemisella voi olla. Tämän lisäksi selvitetään myös, miten alan kirjallisuudessa on tekoälymallien syötemuuttujien valintaa lähestytty. Luvun viimeisenä asiana on tarkastella tapaustutkimuksissa raportoituja tarkkuuslukemia, joiden avulla pystytään saamaan hyvä käsitys eri tekoälymallien potentiaalista.

4.1 Tarkkuuden mittaaminen

Kun vertaillaan tekoälymallien suorituksia eri tutkimuksien välillä, niin tulisi ottaa huomioon useampia asioita. Ensimmäiseksi pitää ottaa huomioon se, miten tutkimuksissa on laskettu ennustustarkkuus. Suurin osa näissä tutkimuksissa esiintyvistä tekoälymalleista ovat luokittelevia, joten yleisimmin käytetty tapa tarkkuuden mittaamiseksi on yksinkertaisesti jakaa oikein luokitellut ennustukset kaikkien tehtyjen ennustusten määrällä. Tällä tavoin saadaan prosenttiluku, joka kertoo kuinka suuri osa tehdyistä ennustuksista on mennyt oikein. Toisaalta regressiomenetelmien yhteydessä yleisesti käytetty keino tarkkuuden arvioimiseksi on laskea tekoälymallin tekemien ennustusten keskineliövirhe. Tämä tarkoittaa sitä, että lasketaan jokaisen tehdyn ennustuksen erotus tiedettyyn tavoitearvoon. Nämä yhdistämällä saadaan käsitys siitä, kuinka lähelle tavoitearvoja tekoälymalli on onnistunut ennustamaan.

Tällaisia tarkkuuksia ei kuitenkaan ole julkistettu kaikissa tutkimuksissa, mikä myös vaikeuttaa niiden välisiä vertailuja. Tällaisissa tutkimuksissa on yleensä kehitetty ja testattu useampia tekoälymalleja, joita on sitten arvioitu

vertailemalla niitä keskenään. Toisaalta näiden tutkimusten hyvä puoli on se, että niissä käytettyjen tekoälymallien kehitykseen ja opetukseen käytettävät tiedot ovat olleet yhdenmukaisia. Tästä syystä voitaisiin olettaa, että vertailukelpoisimmat tulokset ovat samassa tutkimuksessa tehtyjen eri tekoälymallien väliset vertailut.

Seuraava asia mikä pitää ottaa huomioon, onkin siis tutkimuksissa tekoälymallien opetukseen käytettävät tiedot. Huang ja Li (2021) mukaan samankaltaisten tekoälymallien tarkkuuksissa voi ilmentyä eroja sen perusteella, kuinka onnistuneesti tekijät ovat kasanneet opetustiedot. He tarkastelivat amerikkalaisen pesäpallon pääsarjan otteluiden ennustamiseen liittyviä tutkimuksia ja huomasivat sen, että näissä on käytetty useiden eri vuosien tilastoja, jotka eivät välttämättä ole täysin verrannollisia. Tärkeimmäksi huomioksi he kuitenkin mainitsivat tutkimusten väliset erot sopivien syötemuuttujien valitsemisessa (Huang & Li, 2021).

Viimeisenä oleellisena asiana on se, että tutkimuksissa sovelletaan tekoälymalleja useampiin eri urheilulajeihin. Fialho ym. (2019) mukaan tietyt urheilulajit ovat ennustettavimpia kuin toiset. Heidän mukaansa esimerkiksi tasapelin mahdollistavat urheilulajit ovat vaikeampia ennustaa kuin tasapelittömät (Fialho ym., 2019).

4.2 Syötemuuttujien valinta

Hyvin toimivan tekoälymallin edellytyksenä on onnistunut syötemuuttujien valinta. Tekoälymallien tärkeä ominaisuus on niiden kyky yleistää oppimaansa uuteen tietoon. Du Jardin (2009) mukaan tämän kyvyn saavuttamiseksi pitää pystyä pitämään käytettävien syötemuuttujien määrää mahdollisimman vähäisenä. Toisaalta käytettävien syötemuuttujien karsimisella nähdään olevan myös positiivinen vaikutus tekoälymallien tarkkuuteen (du Jardin, 2009; Thabtah ym., 2019; Gao & Kowalczyk, 2021).

Thabtah ym. (2019) yrittivät tutkimuksessaan ennustaa NBA otteluiden lopputuloksia tekoälyn avulla. He keskittyivät tutkimuksessaan tarkastelemaan oikeanlaisten syötemuuttujien merkitystä. He loivat viisi erilaista syötemuuttujajoukkoa hyödyntämällä erilaisia algoritmeja. Tämän jälkeen he opettivat kolmea erilaista tekoälymallia kaikilla eri syötemuuttujajoukoilla. Näiden tekoälymallien tekemien ennustusten perusteella he korostivat sitä, että oikeanlaisten syötemuuttujien valinnalla on positiivinen vaikutus ennustuksien tarkkuuteen (Thabtah ym., 2019).

Myös Gao ja Kowalczyk (2021) keskittyivät tutkimuksessaan syötemuuttujien arviointiin. Heidän mukaansa he kokosivat tähän mennessä suurimman tietokannan koskien ammattilaistennisotteluita, mitä he sitten käyttivät kolmen eri tekoälymallin opetukseen. Tämän tuloksena he pystyivät tunnistamaan tennikseen liittyvät avainsyötemuuttujat ja arvioimaan eri mallien tarkkuutta. He testasivat satunnaismetsän tarkkuutta eri syötemuuttujilla ja lisäämällä niitä yksi kerrallaan he pystyivät paikallistamaan suurimman positiivisen vaikutuksen

sisältämät syötemuuttujat. Oleellista tässä on se, että joidenkin syötemuuttujien lisääminen huononsi tarkkuutta. Parhaimman tarkkuuden he saavuttivat käyttämällä kymmentä eri syötemuuttujaa, kun taas alkuperäisessä tarkastelussa oli mukana reilu kaksikymmentä syötemuuttujaa. Tärkeimmäksi yksittäiseksi syötemuuttujaksi he löysivät syötön vahvuuden (Gao & Kowalczyk, 2021). Myös Candila ja Palazzo (2020) kehittivät tutkimuksessaan ammattilaistennisotteluiden tuloksia ennustavan tekoälymallin ja he päätyivät samankaltaisiin johtopäätöksiin valitessaan tärkeimpiä syötemuuttujia. Eli myös heidän mukaansa yksi tärkeimmistä yksittäisistä tilastoista tekoälymallin ennustavuudelle oli urheilijan oman syötön vahvuus suhteessa vastustajan palautusvahvuuteen (Candila & Palazzo, 2020).

Yleistämiskyvyn ja tarkkuuden paranemisen lisäksi syötemuuttujien optimoinnilla on myös oleellinen vaikutus tekoälymallien suorituskykyyn (Radhika & Syed Masood, 2022; Huang & Li, 2021). Radhika ja Syed Masood (2022) testasivat tutkimuksessaan kolmen eri tekoälymallin ennustustarkkuutta jalkapallon valioliigan yhteydessä. He toteuttivat tämän käyttämällä hyvin yksinkertaisia syötemuuttujia, kuten joukkueiden tehtyjen sekä päästettyjen maalien määrät. Heidän mukaansa tällaisella lähestymistavalla pystyy huomattavasti parantamaan tekoälymallein oppimistehokkuutta (Radhika & Syed Masood, 2022).

Sopivien syötemuuttujien valinta ei kuitenkaan aina ole yksinkertaista. Fialho ym. (2019) huomauttavat esimerkiksi siitä, miten erilaisia oleelliset syötemuuttujat ovat eri urheilulajien välillä. Tämä johtuu yksinkertaisesti siitä syystä, että tietyt urheilulajit ovat täysin erilaisia keskenään. Tiettyjen urheilulajien yhteydessä tämä voi tuottaa vaikeuksia ennustavien tekoälymallien luomisessa, jos kyseisen urheilulajin merkityksellisistä syötemuuttujista ei ole aikaisempaa selvitystä (Fialho ym., 2019).

4.3 Tekoälymallien suoriutuminen

Seuraavaksi käydään läpi, miten eri tekoälymallit ovat vertautuneet toisiinsa alan tutkimuksissa. Vaikkakin tutkimukset eroavat toisistaan muun muassa sovellettavien tekoälymallien sekä kohdelajien osalta, niin niiden tuloksissa toistuvat samankaltaiset huomiot. Merkittävin näistä on se, ettei erilaisten tekoälymallien välillä näytä useinkaan olevan kovinkaan isoja eroja suoritustarkkuudessa. Käydään ensin läpi keinotekoisten neuroverkkojen sekä konvoluutioneuroverkkojen suoriutumista, minkä jälkeen tarkastellaan vielä satunnaismetsän tuloksia.

Candila ja Palazzo (2020) kehittivät keinotekoisen neuroverkon ennustaakseen tennisotteluiden tuloksia ja erityisesti tietyn ottelijan todennäköisyyttä voittaa ottelu. He käyttivät neuroverkossaan vain yhtä piilotettua kerrosta, koska heidän mukaansa tällainen arkkitehtuuri sopii erityisen hyvin ennustaville tekoälymalleille, kunhan piilotettu kerros sisältää tarpeeksi neuroneita. He vertasivat neuroverkon tarkkuutta muihin tenniksessä yleisesti aiemmin käytettyihin ennustusmalleihin, joista vertailuun he valitsivat viisi. Heidän kehittämä neuroverkko oli näistä neljää selvästi parempi ennustustarkkuudessaan, mutta Lisin ja

Zanellan (2017) kehittämä logistinen regressiomalli saavutti lopulta parhaimman tarkkuuden. Candila ja Palazzo (2020) kuitenkin korostivat, että nämä kaksi tekoälymalli olivat hyvin lähellä toisiaan tarkkuudessa varsinkin testimäärien kasvaessa.

Myös Wilkens (2021) kehitti tutkielmassaan ammattilaistennisotteluita ennustavia tekoälymalleja. Hän sisälsi vertailuun myös edellisen tutkimuksen kaltaisesti logistisen regressiomallin sekä keinotekoiset neuroverkot. Näiden tarkkuudet olivat erittäin lähellä toisiaan yhdessä muiden tutkimukseen sisällytettyjen tekoälymallien kanssa. Hän sisällytti vertailuun myös yksinkertaisen vertailulinjan, joka ennusti jokaisessa ottelussa suosikkia. Tekoälymallit saavuttivat opetusvaiheessa noin 70 % tarkkuuden ja ennustamisvaiheessa noin 69 % tarkkuuden. Wilkens (2021) epäili tekoälymallien ennustavan liian usein ottelusuosikin voittoa, minkä takia tarkkuus jäi vain hieman korkeammaksi kuin vertailulinjan tuottama 66 % tarkkuus.

Huang ja Li (2021) tarkastelivat sekä perinteisen neuroverkkojen että yksiulotteisen konvoluutioneuroverkon tarkkuutta ennustamalla amerikkalaisen pesäpallon pääsarjan otteluiden voittajia ja häviäjiä. He käyttivät testaamiseen kahta erilaista tietojoukkoa. Ensimmäisellä tietojoukolla molemmat mallit saavuttivat tarkkuudeksi noin 91 %, perinteisen neuroverkon saavuttaessa hieman korkeamman tarkkuuden. Toisella tietojoukolla tarkkuudet nousivat parhaimmillaan 94 % asti. Heidän mielestään oli tärkeää huomioida se, että vaikkakin perinteiset neuroverkot saavuttivat hieman paremman tarkkuuden niin konvoluutioneuroverkon käytöllä saattaa säästää ajallisesti. Heidän mukaansa konvoluutioneuroverkon yhteydessä ei tarvitse tehdä aikaa ja resursseja vievää syötemuuttujien valintaa yhtä tarkasti, koska se pystyy myös hyvin itse valitsemaan sopivat annetusta joukosta. He nostivat myös esille huomion siitä, ettei konvoluutioneuroverkkojen soveltamisesta urheilutulosten ennustamiseen ole kovinkaan paljon aikaisempaa kirjallisuutta (Huang & Li, 2021).

Hubáček ym. (2019) puolestaan käyttivät perinteistä neuroverkkoa sekä konvoluutioneuroverkkoa yhdessä toisiaan täydentäen. He kehittivät ammattilaiskoripalloon sopeutuvan ottelun voittajaa ennustavan tekoälymallin. Heidän ideanansa oli käyttää konvoluutioneuroverkkoa pelaajakohtaisten syötemuuttujien käsittelemiseen ja perinteistä neuroverkkoa joukkuekohtaisten käsittelyssä. Kehitettyä tekoälymallia he testasivat Pohjois-Amerikan pääsarjan yhteydessä hyödyntämällä yli kymmenen vuoden takaisia ottelutilastoja. Lopulliseksi tarkkuudeksi he onnistuivat saavuttamaan noin 67 % (Hubáček ym., 2019).

Myös satunnaismetsillä on saavutettu tutkimuksissa erinomaisia tuloksia. Radhika ja Syed Masood (2022) ennustivat tutkimuksessaan jalkapallon valioliigan kauden lopullista tulostaulukkoa. He sovelsivat satunnaismetsän lisäksi tehtävään logistista regressioanalyysia sekä tukivektorikonetta. Ennustetuissa sarjataulukoissa oli melko isojakin eroja joukkueiden sijoituksissa, mutta parhaimpaan tulokseen ylsi satunnaismetsällä luotu ennustus (Radhika & Syed Masood, 2022).

Gao ja Kowalczyk (2021) käyttivät samoja tekniikoita kuin Radhika ja Syed Masood (2022) ennustaessaan ammattilaistennisotteluita. He saavuttivat selvästi

parhaimman tarkkuuden satunnaismetsän avulla yltäen hieman yli 80 % asti. He vertailivat tekoälymalleja myös hieman tarkemmin kuin pelkästään tarkkuuden kannalta. He laskivat ennustuksille myös erilliset pisteet ottamalla huomioon sen, kuinka itsevarmoja tekoälymallit olivat ennustuksissaan. Vaikkakin satunnaismetsällä tehdyt ennustukset olivat usein oikein, niin niiden luottamustasot olivat alhaiset (Gao & Kowalczyk, 2021).

Baboota ja Kaur (2019) sovelsivat useita eri tekoälymalleja jalkapallon Valioliiga-otteluiden ennustamiseen. Heidän tuloksistaan erottautui kaksi selvästi paremmin soveltuvaa tekoälymallia, jotka olivat satunnaismetsä sekä gradienttitehostettu luokittelija (engl. *gradient boosted classifier*). Satunnaismetsä saavutti testissä 57 % tarkkuuden ja gradienttitehostettu luokittelija vielä hieman paremman (Baboota & Kaur, 2019). Gradienttitehostettu luokittelija on kuitenkin hyvin samankaltainen kuin satunnaismetsä, sillä sekin perustuu useiden päätöspuiden hyödyntämiseen. Ne eroavat päätöspuiden erilaistamisessa ja opettamisessa, mutta molempien perustuessa päätöspuihin, voidaan todeta päätöspuiden soveltuvan erityisen hyvin jalkapallo-otteluiden voittajien sekä häviäjien ennustamiseen. Kapadia, Abdel-Jaber, Thabtah ja Hadi (2020) sovelsivat satunnaismetsää kriketin ammattiotteluiden voittajien ennustamiseen. Heidän tulosten perusteella voidaan todeta satunnaismetsän toimineen kyseisessä tehtävässä paremmin kuin perinteisemmät tilastolähtöiset ennustuskeinot.

Tuloksista voidaan yleisesti päätellä tarkasteltujen tekoälymallien soveltuvan hyvin urheilutulosten ennustamiseen. Ne saavuttivat vertailukelpoisia tuloksia monissa urheilulajeissa. Tulosten perusteella ei kuitenkaan voida todeta tekoälymallien suoriutuvan paremmin kuin perinteiset tilastolliset menetelmät.

Tekoälymallien tutkiminen ja kehittäminen on kuitenkin tuore aihe ja tekniikat kehittyvät jatkuvasti. Tulevaisuudessa yksi tarkkuuksia parantava tekoälymalli saattaa olla takaisinkytketty neuroverkko (engl. *recurrent neural network*). Kyseistä tekoälymallia on jo sovellettu muutamilla aloilla erinomaisin tuloksin. Hou ja Tian (2022) sovelsivat takaisinkytkettyä neuroverkkoa ennustamaan 3000 metrin juoksijoiden suoriutumista kilpailuissa. He pystyivät erilaisten fysiologisten arvojen perusteella ennustamaan juoksijoiden päivän suoriutumiskykyä. Vetukuri, Sethi ja Rajender (2020) puolestaan hyödynsivät erinomaisin tuloksin takaisinkytkettyä neuroverkkoa krikettijoukkueiden pelaajavalintaan. Jatkossa olisikin erittäin tärkeää tutkia lisää vastaavien syväoppimismallien soveltamismahdollisuuksia urheilutulosten ennustamiseksi

5 YHTEENVETO

Urheiluotteluista kerättävän tiedon määrä on kasvanut teknologian ja internetin myötä suuresti. Tämä on mahdollistanut useiden eri urheilulajien analysoimisen tilastolähtöisten keinojen avulla. Samaan aikaan urheiluvedonlyönti on yleistynyt netin välityksellä, mikä on omalta osaltaan kannustanut ihmisiä analysoimaan urheiluotteluita tulevien tulosten ennustamiseksi. Tähän tarkoitukseen on useita erilaisia lähestymistapoja, joista yhtenä uusimmista on tekoälyn hyödyntäminen. Aiheesta tehdään aktiivisesti tutkimusta ja tekoälymalleja yritetään jatkuvasti kehittää.

Tässä kandidaatintutkielmassa on perehdytty tekoälyn hyödyntämiseen urheilutulosten ennustamiseksi. Tutkielmassa on keskitytty erityisesti kolmeen tekoälymalliin, joita ovat keinotekoiset neuroverkot, konvoluutioneuroverkot sekä satunnaismetsät. Ensimmäisenä läpikäytävänä asiana oli näiden tekoälymallien perusteet sekä toiminnallisuudet. Tämän jälkeen määriteltiin urheilutulosten käsite sekä käytiin läpi yleisimpiä lähestymistapoja, joiden avulla urheilutuloksia yritetään ennustaa. Tämän lisäksi tarkasteltiin myös, miten urheilutulosten ennustukset tulevat nykyään usein näkyville urheiluotteluiden seuraamisen yhteydessä. Viimeisessä sisältöluvussa käytiin ensimmäiseksi läpi rajoitteita, jotka vaikuttivat eri tutkimuksien tuloksien suoraan vertailuun. Toisena asiana oli tekoälymallien toiminnallisuuden kannalta hyvin kriittisessä asemassa olevien syötemuuttujien valitseminen. Luvun viimeisenä asiana tarkasteltiin vielä tekoälymallien avulla eri tutkimuksissa saavutettuja tarkkuuksia.

Tämän kandidaatintutkielman tarkoituksena oli vastata kahteen tutkimuskysymykseen. Ensimmäinen tutkimuskysymys liittyi tekoälyn hyödyntämiseen urheilutulosten ennustamisessa. Haluttiin selvittää millä erilaisin tavoin tekoälyä on hyödynnetty urheilutulosten ennustamiseen. Toinen tutkimuskysymys puolestaan liittyi näiden tekoälymallien tarkkuuteen urheilutuloksia ennustettaessa. Vastaaminen tutkimuskysymyksiin tapahtui pääosin viimeisessä sisältöluvussa, mutta vastauksia ensimmäiseen tutkimuskysymykseen sivuttiin myös aikaisemmissa sisältöluvuissa.

Erilaisia tekoälymalleja on hyödynnetty yhteiskunnan eri osa-alueilla hyvällä menestyksellä jo useita vuosia. Myös urheilutulosten ennustamiseen

liittyen on tehty tutkimuksia, mutta käytännön tasolla useimmat perinteisemmät tilastolliset keinot ovat vielä suositumpia. Tutkielmassa opittiin, miten keinotekoiset neuroverkot toimivat hyödyntäen erilaisia kerroksia, jotka sisältävät toisiinsa linkittyneitä neuroneita. Opittiin myös tarkemmin neuroneiden välisestä kommunikaatiosta linkkien välityksellä ja neuroneiden aktivoitumisesta. Tutustuttiin myös yksittäisistä päätöspuista muodostuviin satunnaismetsiin ja näiden toimintaperiaatteisiin.

Tutkielmassa käsiteltiin myös yleisesti urheilutulosten ennustamista ja sen näkymistä urheiluotteluiden yhteydessä. Selvästi yleisimmäksi tavaksi osoittautui subjektiivisiin päätelmiin perustuva ennustustyyli. Tämä tapa oli kuitenkin hyvin virheherkkä eikä siksi sopiva tarkkojen ennustuksien toistuvaan tekemiseen. Sen sijaan kannattavammaksi todettiin perustaa ennustukset erilaisiin tilastoihin. Tässä yhteydessä tekoälyllä on potentiaalia automatisoida ja parantaa jo hyväksi todettuja tilastollisia ennustuskeinoja.

Viimeisimmässä sisältöluvussa keskityttiin tekoälymallien toimintaan urheilutulosten ennustamisessa. Oikeanlaisten syötemuuttujien käyttäminen osoittautui hyvin oleelliseksi osaksi tarkkojen ennustusten mahdollistamisessa. Osa tekoälymalleista suoriutui kuitenkin myös melko hyvin ilman syötemuuttujien valintaa ja joissakin tapauksissa tarkkuus jopa parani, kun tekoälymallin annettiin itse löytää optimaaliset syötemuuttujat. Tällaisia tekoälymalleja olivat neuroverkkoihin perustuvat ratkaisut.

Tulosten perusteella voidaan todeta tekoälymallien sopivan hyvin urheilutulosten ennustamiseen. Näiden hyödyt korostuvat entisestään, mitä monimutkaisempaan urheilulajiin niitä sovelletaan. Ei kuitenkaan ollut selvää yltävätkö nämä tekoälymallit korkeampiin tarkkuuslukemiin kuin alalla tällä hetkellä yleisesti käytettävät tilastolliset keinot. Paremman tarkkuuden sijaan tekoälymallit saattavat tehostaa urheilutulosten ennustamista automatisoimalla toimintoja. Varsinkin konvoluutioneuroverkko osoittautui tutkimuksissa edukseen omien syötemuuttujien valinnallaan.

Tekoälymallien hyödyntämistä urheilutulosten ennustamiseksi ei kuitenkaan ole vielä keretty tutkimaan kovin pitkään. Suurin osa kirjallisuudesta koskee koneoppimismalleja, minkä takia syväoppimismallien toiminnasta ei ole kattavaa tietoa. Syväoppimismallit ovat saavuttaneet jo useilla muilla aloilla suurta menestystä, ja ne ovat suoriutuneet paremmin kuin koneoppimismallit. Tästä syystä olisi hyvin tärkeää saada jatkossa syväoppimismalleista tutkimuksia myös urheilutulosten ennustamiseen liittyen.

LÄHTEET

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic Concepts of Artificial Neural Network (ANN) modeling and its application in Pharmaceutical Research. *Journal of Pharmaceutical and Biomedical Analysis*, *22(5)*, 717–727. https://doi.org/10.1016/s0731-7085(99)00272-1

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using Machine Learning Approach for English Premier League. *International Journal of Forecasting*, *35(2)*, 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Breiman, L. (2001). Random forests. *Machine Learning*, *45(1)*, 5–32. https://doi.org/10.1023/a:1010933404324

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15(1)*, 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Butler, D., Butler, R., & Eakins, J. (2021). Expert performance and crowd wisdom: Evidence from English Premier League predictions. *European Journal of Operational Research*, *288(1)*, 170–182. https://doi.org/10.1016/j.ejor.2020.05.034

Candila, V., & Palazzo, L. (2020). Neural Networks and betting strategies for tennis. *Risks, 8(3),* 68. https://doi.org/10.3390/risks8030068

de Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5(6)*, 448–455. https://doi.org/10.1002/wics.1278

du Jardin, Philippe (2009). Bankruptcy prediction models: How to choose the most relevant variables? *Bankers, Markets & Investors, 98*, 39-46.

Gao, Z., & Kowalczyk, A. (2021). Random Forest model identifies serve strength as a key predictor of tennis match outcome. *Journal of Sports Analytics, 7(4),* 255–262. https://doi.org/10.3233/jsa-200515

Goodfellow, I., Bengio, Y., & Courville, A. (2018). *Deep learning*. MITP.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of Artificial Intelligence. *California Management Review*, *61(4)*, 5–14. https://doi.org/10.1177/0008125619864925

Hou, J., & Tian, Z. (2022). Application of recurrent neural network in predicting athletes' sports achievement. *The Journal of Supercomputing, 78(4),* 5507–5525. https://doi.org/10.1007/s11227-021-04082-y

Houghton, D. M., Nowlin, E. L., & Walker, D. (2019). From fantasy to reality: The role of fantasy sports in sports betting and online gambling. *Journal of Public Policy & Marketing*, *38*(3), 332–353. https://doi.org/10.1177/0743915619841365

Huang, M.-L., & Li, Y.-Z. (2021). Use of machine learning and deep learning to predict the outcomes of Major League Baseball matches. *Applied Sciences, 11(10),* 4499. https://doi.org/10.3390/app11104499

Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting, 35(2),* 783–796. https://doi.org/10.1016/j.ijforecast.2019.01.001

Ibragimov, B., & Xing, L. (2017). Segmentation of organs-at-risks in head and neck CT images using Convolutional Neural Networks. *Medical Physics, 44(2),* 547–557. https://doi.org/10.1002/mp.12045

Kapadia, K., Abdel-Jaber, H., Thabtah, F., Hadi, W. (2020). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics, 18(3/4),* 256–266. https://doi.org/10.1016/j.aci.2019.11.006

Kollár, A. (2021). Betting models using AI: A review on ANN, SVM, and Markov Chain [Preprint]. *Open Science Framework.* https://doi.org/10.31219/osf.io/mr2v3

Leung, C. K., & Joseph, K. W. (2014). Sports data mining: Predicting results for the college football games. *Procedia Computer Science, 35,* 710–719. https://doi.org/10.1016/j.procs.2014.08.153

Lisi, F., & Zanella, G. (2017). Tennis betting: Can statistics beat bookmakers*? Electronic Journal of Applied Statistical Analysis, 10,* 790–808.

Loh, W. Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery, 1(1),* 14–23. https://doi.org/10.1002/widm.8

Radhika, A., & Syed Masood, M. (2022). Premier League Table Prediction Using Machine Learning Algorithms. *Webology, 19(1),* 6379-6395.

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). Glocalx - from local to global explanations of black box AI models. *Artificial Intelligence, 294,* 103457. https://doi.org/10.1016/j.artint.2021.103457

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science, 6(1),* 103–116. https://doi.org/10.1007/s40745-018-00189-x

Vasilev, I., Slater, D., Spacagna, G., Roelants, P., & Zocca, V. (2019). *Python deep learning: Exploring deep learning techniques and neural network architectures with pytorch, Keras, and tensorflow.* Packt Publishing.

Vetukuri, V. S., Sethi, N., & Rajender, R. (2020). Generic model for automated player selection for cricket teams using recurrent neural networks. *Evolutionary Intelligence, 14(2),* 971–978. https://doi.org/10.1007/s12065-020-00488-4

Wilkens, S. (2021). Sports prediction and betting models in the Machine Learning Age: The case of tennis. *Journal of Sports Analytics*, *7(2),* 99–117. https://doi.org/10.3233/jsa-200463