Yhteenveto tutkimusraporteista

Äyrämö Sami[1*]

[1] Faculty of Information Technology, University of Jyvaskyla, Finland

Abstract — Tekoäly on tällä hetkellä yksi nopeimmin kehittyvistä teknologioista. Modernit tekoälysovellukset perustuvat valtaosin dataa mallintaviin koneoppimismenetelmiin. Suomessa kerätään suuria määriä tekoälyn hyödynnettäväksi soveltuvaa dataa erilaisten terveyden, hyvinvoinnin ja lääketieteen sovellusten kautta. Tämän raportin tarkoitus on esittää "*Value from health data with cognitive computing*"-projektin (TEKES) koneoppimiseen perustuvien tekoälyteknologioiden kyvykkyyksiä selvittäneiden osatutkimusten tuloksia. Tutkimukset suoritettiin käyttäen projektin yhteistyökumppaneiden tai Jyväskylän yliopiston keräämiä aineistoja. Tässä osassa käydään ensin läpi yleisellä tasolla mistä koneoppimispohjaisessa tekoälyssä on kysymys ja sen jälkeen vedetään yhteen yhdeksän teknisen osaraportin tulokset. Tulokset osoittavat, että datapohjaisille sovelluksille on Suomessa tarve ja mahdollisuus. Huolimatta "*Big data*"-ilmiöstä, käytännössä tutkimusta ja innovaatioita hidastaa edelleen myös riittävän suuren ja laajan datan puuttuminen johtuen niiden pirstaleisuudesta. Tähän voidaan vastata parhaiten yhdistämällä aineistoja esimerkiksi Omakannan Omatietovaranto -tyyppisiin ympäristöihin.

1 Johdanto

Tekoälystä (engl. *artificial intelligence*) on tullut yksi keskeisimmistä teknologioista pohdittaessa terveydenhuollon kehittämistä ja tehostamista yhteiskunnan muutosten ja haasteiden keskellä. Tekoälysovelluksia on kehitetty terveyshuollon ja lääketieteen sovelluksiin 1980-luvulta alkaen. Liikunta- ja hyvinvointisektoreiden innovaatiot ovat kuitenkin yleistyneet vasta viime vuosina, sillä kuluttajatason sensori- ja datantallennusratkaisut ovat kehittyneet vasta tämän vuosituhannen puolella. Suomella on vahvat lähtökohdat sekä tekoälyn kehittämiseen että hyödyntämiseen korkean osaamistason ja laajasti digitalisoituvan terveydenhuollon ansiosta (Lehto and Neittaanmäki, 2017). Lisäksi suomalaiset yritykset ovat kehittäneet terveys- ja hyvinvointialalle lukuisia digitaalista dataa ja signaaleja kerääviä sensori-, mittalaite-, ja tiedonkeruuratkaisuja. Omadataa hyödyntävän tekoälyn kehitysmahdollisuudet tulevat paranemaan entisestään muun muassa Omakannan Omatietovarannon käyttöönoton ja datan toissijaisen käytön mahdollistumisen myötä (Niinimäki et al., 2017).

Yksilöllisen vaihtelun huomioivat lähestymistavat potilaiden hoidossa, riskien arvioinnissa ja sairauksien ennaltaehkäisemisessä, englanninkielisin termein ilmaistuna *precision medicine*[1] tai *personalized medicine*, perustuvat suurien tietomassojen ja koneoppimisen hyödyntämiseen. Perinteisesti tutkimustiedon tuottaminen on perustunut valtaosin "one-size-fits-all"-strategiaan, jossa populaatiossa keskimäärin toimivaksi todetun menettelyn oletetaan yleistyvän valtaosin kaikkiin yksilöihin. Tämä johtaa luonnollisesti suuriin yksilöllisiin eroihin hoitoa saaneiden tai vaikkapa liikuntareseptiä noudattaneiden henkilöiden kokemissa vasteissa. Yksilöllisen lääketieteen tavoitteena on ottaa huomioon yksilön perimään, ympäristöön ja elämäntyyliin liittyvät tekijät, mikä mahdollistaa yksilöllisesti tarkemmat riskiarviot ja diagnoosit, vaikuttavammat terveys- ja liikuntasuositukset ja tehokkaan hoidon suunnittelun sairauden osuessa kohdalle. Henkilökohtainen lääketiede

*Yhteyshenkilö: sami.ayramo@jyu.fi, P.O.Box 35, FI-40014 University of Jyvaskyla, Finland

[1] https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

ei luonnollisesti tarkoita menetelmien kehittämistä yksilöittäin vaan yhtenäisesti käyttäytyvien osajoukkojen tai alityyppien tunnistamista suuremmasta datamassasta koneoppimisen ja data-analytiikan keinoin.

Dr. Linda A. Winters-Miner (Linda, 2014) on kuvannut seitsemän tapaa, joilla ennustava mallintaminen voi edistää terveyttä ja hyvinvointia. Vapaasti käännettynä ne voitaneen tulkita seuraavasti:

1. Tarkemmat diagnoosit
2. Yksilöllisempien riskiarvioiden avulla edistyvä ennaltaehkäisevä terveydenhuolto ja kansanterveys
3. Yksilöllisemmät hoitomenetelmät
4. Työterveyden palveluiden vakuutuskustannusten ennustaminen
5. Tiedon aiempi jalkauttaminen kliiniseen toimintaan jatkuvaan oppimiseen kykenevien mallien avulla
6. Lääkkeiden tarvepohjainen kehittäminen
7. Potilaan aktiivisempi rooli terveytensä edistämisessä

Tekoälymenetelmien kehittäminen yksilöllisemmän terveydenhoidon edistämiseksi edellyttää älykkäiden algoritmien lisäksi suuret määrät monipuolista ja riittävän laadukasta dataa. Modernit kone- ja syväoppimisalgoritmit (engl. *machine learning, deep learning*) pystyvät älykkäästi mallintamaan/oppimaan datassa piileviä monimutkaisiakin yhteyksiä ja riippuvuuksia. Menetelmien älykkyydestä ja oppimiskyvystä huolimatta varsinainen sovellusalakohtainen tieto on luonnollisesti kuitenkin datassa. Tekoälylgoritmit eivät pysty lisäämään tai luomaan oppimaansa malliin sellaista tietoa, jota niiden opettamiseen käytettävä data ei sisällä. Tekoälyn laatu ja kattavuus on siis varsin pitkälle datasta riippuvaista. Toisaalta ilman data- ja koneoppimisalan ammattitaitoa suurten datamassojen valtava potentiaali on mahdotonta hyödyntää. Tällaisen keskinäisen riippuvuuden vallitessa kehittämis- ja tutkimustyö edellyttää tiivistä yhteistyötä tekoäly- ja terveys-/hyvinvointialojen osaajien välillä.

2 Koneoppimiseen perustuva tekoäly

Kiinnostus tekoälyä kohtaan on elänyt eräänlaisessa aaltoliikkeessä nykyaikaisten ohjelmoitavien tietokoneiden keksimisestä lähtien. Aluksi tietokoneiden "viisautta" hyödynnettiin pääasiassa formaaliin esitystapaan helposti taipuvien matemaattisten tehtävien ratkomiseen (Goodfellow et al., 2016). Tällaisten tehtävien ratkominen koneen avulla oli hyödyllistä, sillä tehtävät olivat ihmiselle vaikeita, mutta tietokoneille helppoja ja nopeita ratkaista. Klassinen tekoäly perustuu tietämyksen esittämiseen hyvin määriteltyjen symbolien ja loogisten päättelysääntöjen muodossa (Nilsson, 1998). Todellinen maailma on kuitenkin osoittautunut varsin vaikeasti mallinnettavaksi ohjelmoitavan tekoälyn muodossa.

Vaikka tietokone voidaan opettaa pelaamaan älykkäänoloisesti shakkia, on pelin maailma varsin rajallinen ja yksinkertainen verrattuna reaalimaailman ongelmiin. Tosielämän tapausten mallintamiseen vaadittavan kokonaisvaltaisen informaation kerääminen voi olla äärimmäisen vaikeaa. Vaikka monimutkaisten matemaattisten tehtävien ratkaiseminen tietokoneen avulla on nopeuttanut maailman kehittymistä monin tavoin, muodostavat ihmiselle intuitiivisen oloisten mutta vaikeasti formaalissa muodossa esitettävien tehtävien ratkominen edelleen tekoälylle suuren haasteen (Goodfellow et al., 2016). Ollakseen älykäs ja vuorovaikutuskykyinen täytyy tietokoneen siten pystyä oppimaan asioita kuvista, äänistä, teksteistä, sähkösignaaleista, tapahtumasekvensseistä jne.

eli suurista matalan abstraktiotason tietomassoista. Ihmisestä tällaista dataa tuotetaan valtavia määriä biosignaalien, digitaalisten kuvien, geenisekvenssien ym. muodossa.

2.1 Perinteiset koneoppimisen menetelmät

Koneoppiminen on keino opettaa tietokone tunnistamaan, luokittelemaan ja ennustamaan reaalimaailman ilmiöitä ja hahmoja (Bishop, 2006). Koneoppiminen jaetaan pääasiassa kahteen paradigmaan: ohjattuun (engl. *supervised learning*) ja ohjaamattomaan oppimiseen (engl. *unsupervised learning*) (Friedman et al., 2001). Ensin mainitussa kone opetetaan ennustamaan tai luokittelemaan määrättyjä vasteita. Vasteet voivat olla luokiteltuja muuttujia, kuten kasvaimen tyyppi (hyvänlaatuinen = 0 / pahanlaatuinen = 1), tai jatkuvia muuttujia, kuten riskikerroin väliltä 0-1. Ohjaamattomassa oppimisessa datasta pyritään hahmottamaan sen sisäinen rakenne ja tunnistamaan uusia riippuvuuksia ja suhteita. Vaikka osaraporteissa on sovellettu myös ohjaamatonta oppimista, on tämän kokoomaraportin päähuomio ohjatussa koneoppimisessa.

Keinotekoiset neuroverkot (engl. *artificial neural networks*) ovat tekoälyn yksi tunnetuimmista menetelmistä (Haykin et al., 2009). Teoreettisesti keinotekoiset neuroverkot ovat houkuttelevia suuren oppimiskapasiteettinsa ansiosta. Universaalin approksimaatioteoreeman perusteella yhdestä piilokerroksesta koostuva epälineaarinen monikerros- eli MLP-neuroverkko voi mallintaa minkä tahansa jatkuvan funktion (Hornik, 1991). Vaikka kiinnostus neuroverkkoja kohtaan on ajan saatossa vaihdellut, niin nyt 60-70 vuotta alan pioneerien McCulloughin, Pittsin ja Rosenblattin esittämien ensioivallusten jälkeen neuroverkkoja on jo sovellettu lukuisissa sovelluksissa niin lääketieteessä kuin terveys- ja hyvinvointialoilla. 1990-luvun puoliväliin mennessä sovelluksia oli toteutettu diagnostiikassa, kuvantamisessa, biosignaalien analysoinnissa, hoitovasteen ennustamisessa, patologiassa ja farmakologiassa (Baxt, 1995). 2000-luvun edistysaskeleet syväoppimisessa, datan määrän kasvu, uudet opetusalgoritmit ja laskentatehon kasvu ovat antaneet viimeisimmän sysäyksen neuroverkkotutkimukselle. Muita esimerkiksi kliinisissä tekoälysovelluksissa käytettäviä perinteisiä koneoppimisen menetelmiä ovat muun muassa päätöspuut (engl. *decision tree*) (Bae, 2014), k-lähimmän naapurin menetelmä (engl. *k-Nearest Neighbor* tai *kNN*) (Altilio et al., 2017), Lasso regressio (Zemmour et al., 2015), tukivektorikone (engl. *support vector machine*) (Bisenius et al., 2017), Naïvi Bayes-luokitin(Pershad et al., 2017), Bayes-verkot (Takenaka and Aono, 2017) ja satunnaismetsät (engl. *random forest*) (Shaikhina et al., 2017).

On hyvä muistaa, että yksinkertaisissa tapauksissa jopa perinteinen lineaarinen tai logistinen regressiomalli voi edelleen olla toimivin ennustusmenetelmä. Ylipäätään koneoppimisessa tulee noudattaa yksinkertaisuuden periaatetta, koska yksinkertaisin toimiva ratkaisu yleistyy uusille mallin opetusvaiheessa tuntemattomille tapauksille parhaiten. Osassa tämänkin kokoomaraportin tutkimustapauksista ei pystytty löytämään lineaariselle mallille tarkempaa vaihtoehtoa mallin kompleksisuutta lisäämällä. Myös datan vähäinen määrä on peruste käyttää yksinkertaisia malleja, jotta vältyttäisiin mallin ylioppimiselta. Mallinnusmenetelmää valittaessa on tärkeää huomioida myös datan laatu, jotta tunnistettaisiin laatutekijöistä aiheutuvat vaatimukset menetelmien robustisuudelle ja puuttuvien arvojen käsittelylle (Äyrämö, 2006).

Vahvistusoppiminen (engl. *reinforcement learning*) on vähemmän huomiota saanut koneoppimisen paradigma, jossa älykäs agentti oppii onnistumisten ja virheiden kautta toimiessaan vuorovaikutuksessa ympäristönsä kanssa (Sutton and Barto, 1998). Agentti saa valintojensa perusteella negatiivista tai positiivista palautetta. Agentti pyrkii maksimoimaan positiivisen palautteensa, mikä johtaa jatkuvaan oppimiseen. Agentin on päätöksissään tasapainoiltava aiemmin koettujen palkittujen vaihtoehtojen ja uusien tuntemattomien vaihtoehtojen välillä. Vahvistusoppimista voidaan pitää koneoppimisen kolmantena paradigmana ohjatun ja ohjaamattoman oppimisen rinnalla. Vahvistusoppiminen eroaa selkeästi ohjatusta oppimisesta, sillä agentin opettamiseen ei sovelleta varsinaista mahdollisimman yleistyvään malliin pyrkivää opetusaineistoa ja -algoritmia. Paradigma ei myöskään vastaa ohjaamatonta oppimista, sillä agentin tavoitteena ei ole löytää datassa

piileviä rakenteita tai riippuvuuksia vaan maksimoida positiivisten palautteiden kertymä.

Modernit tekoälysovellukset perustuvat siis pääosin datasta oppimiseen. Miten tämä sitten eroaa perinteisestä dataa selittävästä analytiikasta (engl. *explanatory modeling*)? Tekoälytehtävät ovat pääasiassa tavalla tai toisella ennustustehtäviä (engl. *predictive modeling*). Selittävän mallintamisen avulla pyritään ymmärtämään selittävien ja selitettävien muuttujien väliset riippuvuudet ja kausaliteetit. Ennustavassa mallintamisessa taas pyritään mahdollisimman tarkkaan ennustustulokseen syötteiden ja vasteiden välisten yhteyksien luonteen jäädessä useimmiten mustalaatikkotyyppisen mallin taakse (Breiman et al., 2001). Tämän vuoksi ennustavassa mallintamisessa, ja siten koneoppimiseen perustuvassa tekoälyssäkin, datasta opetettujen mallien validointi riippumattomilla aineistoilla on äärimmäisen tärkeää ennen niiden käytäntöön viemistä.

2.2 Syväoppiminen

Kuten edellä mainittiin, neuroverkkotutkimus on kiihtynyt voimakkaasti 2000-luvulla ja tekoäly on noussut "hypetasoiseksi" teknologiatrendiksi erityisesti syvien neuroverkkojen (engl. *deep neural networks*) keskuudessa tapahtuneen kehityksen seurauksena (LeCun et al., 2015). Perinteiset koneoppimisen menetelmät mahdollistivat koneen opettamisen datapohjaisesti älykkyyden ohjelmoimisen sijaan, mutta edelleen soveltamisen haasteena on ollut mallien opettamiseen tarvittavien piirteiden muodostamisen vaikeus. Älykkään koneen toivottaisiin usein ymmärtävän puhetta, ääniä, kuvia tai sensorin mittaamia biosignaaleita, mutta näiden syöttäminen koneoppimisalgoritmille on haastavaa, sillä perinteiset luokittelu- ja ennustusmenetelmät pystyvät ottamaan syötteinä pääasiassa vain asiantuntijan valmiiksi "käsin" suunnittelemia ja laskemia piirteitä. Piirteiden muodostaminen onkin ollut erittäin tutkittu osa-alue esimerkiksi konenäössä (esim. (Nixon and Aguado, 2012)).

Ainakin osaratkaisun moniin datan piirteytyshaasteisiin ovat tarjonneet syvät neuroverkot, jotka ovat käytännössä piirteenmuodostuksen sisältäviä monikerroksisia neuroverkkoja. Keksintönä syvät neuroverkot eivät ole täysin uusia, sillä niitä on tutkittu 1960-luvulta lähtien (Schmidhuber, 2015). Uusien syvien neurolaskentaopetusalgoritmien, grafiikkaprosessorien ja hajautetun laskennan kehittyminen antoivat 2000-luvulla voimakkaan sysäyksen nykyiselle tekoälykehitykselle. Erityisesti konvolutiivisia ja rekurrentteja neuroverkkoja on käytetty konenäön ja luonnollisen kielenkäsittelyä vaativissa sovelluksissa datan spatiaalisten tai temporaalisten rakenteiden mallintamiseen.

Syviä neuroverkkoteknologioita hyödyntävät nykyään useat yritykset, kuten IBM, Google, Microsoft, Apple ja NVIDIA. Lääke- ja terveystieteen puolella syviä neuroverkkoja on testattu lukuisissa sovelluksissa, kuten lääketieteellisessä kuvantamisessa (Shen et al., 2017), aivosignaalien käsittelyssä (Schirrmeister et al., 2017), genomiikassa (Angermueller et al., 2016), rintasyövän tunnistamisessa (Kooi et al., 2017), sydäninfarktin tunnistamisessa (Acharya et al., 2017) ja liikedatan parametrisoinnissa vaikkapa Alzheimer-taudin seurantaa varten (Hannink et al., 2017).

Äkkiseltään syväoppimisesta voi saada jopa sellaisen kuvan, että ilmiöiden opettaminen tietokoneelle neuroverkkojen avulla olisi jo varsin suoraviivaista raakamuotoisen opetusdatan syöttämistä, mutta käytännössä haasteena on edelleen opettamisen edellyttämä datan määrä. Syvät neuroverkot voivat sisältää miljoonia säädettäviä parametreja, minkä vuoksi myös opetusdataa tarvitaan valtavasti. Esimerkiksi Nature-lehdessä vuonna 2017 julkaistussa artikkelissa esitetty konvoluutioverkkoon ja valokuviin perustuva ihosyövän tunnistusmenetelmä opetettiin 129 450 kuvalla (Esteva et al., 2017). Verkko oli tätä ennen esiopetettu yli miljoonalla kuvalla ja 1000 luokituskategorialla. Datamäärän ollessa pieni verkot ylioppivat helposti opetusaineiston eivätkä tulokset tällöin yleisty uusiin ennalta tuntemattomiin havaintoihin. Ylioppimisen välttämiseksi käytetään verkon opettamisessa myös regularisointimenetelmiä (esim. (Srivastava et al., 2014)). Datan ja parametrien määrän kasvu puolestaan tekee verkon opettamisesta laskennallisesti raskaampaa, mutta nykyiset laskenta-alustat ja opetusalgoritmit suoriutuvat näistä tehtävistä varsin hyvin.

Suomalainen terveysdata tarjoaa paljon mahdollisuuksia tekoälyn hyödyntämiselle. Useat "state-of-the-art"-kone- ja syväoppimismenetelmät ovat saatavilla avoimien kirjastojen kautta. Python-ohjelmointikielestä kirjastoineen on tullut käytetyin työväline koneoppimis- ja tekoälysovellusten kehittämiseen. Python onkin noussut kone- ja syväoppimisen trendiohjelmointikieleksi. Yleisesti käytettyjä syväoppimiskirjastoja ovat Caffe[2], Theano[3], Torch7[4] ja TensorFlow[5].

3 Testitapaukset

Tässä kokoomateoksessa esitellään tuloksia "Value from health data with cognitive computing"-projektin (Tekes) terveys- ja liikunta-alan aineistoilla toteutetuista testitapauksista. Projektin osatyöpakettina oli selvittää modernien kognitiivisten tekoälyteknologioiden potentiaalia suomalaisessa terveys- ja hyvinvointidataympäristössä. IBM kuvaa kognitiivista teknologiaa neljän kyvykkyyden kautta: *Ymmärtää*, *päättelee*, *oppii* ja *vuorovaikuttaa*. Tämän raportin piirissä kuvatuissa testeissä ei testattu yhtään järjestelmäkokonaisuutta, joka tarjoaisi nämä kaikki ominaisuudet. Sen vuoksi testasimme erilaisia kyvykkyyksiä ongelma- ja aineistolähtöisesti.

Testiesimerkkien tarkoituksena on tiedon tuottaminen tekoälyn ja koneoppimisen sovellutusmahdollisuuksista kotimaisessa terveysdatamaisemassa yhteistyökumppaneilta saatuja reaalimaailman aineistoja hyödyntäen. Testiesimerkit edustivat karkeasti ottaen seuraavia osa-alueita:

- Kliinisen lääketieteen ja sairaalaympäristön sovellutukset
- Lasten ja nuorten hyvinvointi ja kokonaisvaltainen kehittyminen
- Ennaltaehkäisevä terveydenhuolto ja älykkäät sensorit ja apuvälineet

Kaikki testitapauksemme tutkivat koneoppimismenetelmien kykyä oppia syötetystä datasta. Tutkimme esimerkiksi kuinka koneoppimisalgoritmit voidaan opettaa ymmärtämään syöte- ja vastemuuttujien välisiä riippuvuuksia raakamuotoisesta syöpäkuva-aineistosta opettamalla tekoälyalgoritmi laskemaan kuvasta syöpäsolujen lukumäärä. Lisäksi testasimme kuinka ohjaamaton koneoppimismenetelmä osaa muodostaa koululaisaineistosta erilaisia oppijoiden alityyppejä. Testasimme myös kuinka tekoälyalgoritmi voidaan opettaa päättelemään raakamuotoisesta liikeanalyysikuvadatasta juoksijan askeltyyppi tai jalkapallonuoren mahdollisuudet edetä kansainväliseen valmennusakatemiaan. Vuorovaikutusominaisuuksia ei tutkittu, koska tämä olisi käytännössä vaatinut suomenkielisen käyttöliittymän eikä kyseinen kyvykkyys ollut aineistojen ja sovellusten potentiaalin tutkimisen kannalta relevantti.

Testiesimerkkien toteuttamiseksi aineistoja ja/tai asiantuntemustaan projektin käyttöön antoivat Keski-Suomen Sairaanhoitopiiri, Itä-Suomen yliopisto, Eerikkilän urheiluopisto, LIKES, Liikuntatieteellinen tiedekunta, UKK-instituutti, KIHU ja HUS. Lisäksi hyödynnettiin yhtä avointa aineistoa syväoppimismenetelmien tutkimiseksi. Emme tunnistaneet kaupallisista ohjelmistoista valmiiksi koulutettuja malleja tai mallien kouluttamista erityisesti edistäviä ohjelmistoja, joten testit toteutettiin pääosin avoimen lähdekoodin kirjastoja käyttäen. Keskeiseksi kysymykseksi suljetuissa ohjelmistoissa havaittiin mm. mallin validointimenetelmien luotettavuus.

Tutkimuksiin kartoitettiin menetelmät avoimien (Python kirjastot) ja kaupallisten ratkaisujen joukosta (IBM, Mathworks). Projektissa on räätälöity testejä varten tekoäly- ja koneoppimiskirjastojen, kuten Python Scikit ja Panda, perusmenetelmiä mallien luomiseen rakenteisesta datasta sekä kartoitettu ja testattu myös syväoppimiskirjastojen (Keras, MxNet, TensorFlow, Lasagne) menetelmiä raakamuotoisen datan (syöpäkuvat ja biomekaaniset signaalit) opettamisessa koneelle.

[2] http://caffe.berkeleyvision.org/
[3] http://deeplearning.net/software/theano/
[4] http://torch.ch/
[5] https://www.tensorflow.org/

Suurten kuva- ja biomekaanisten signaalien vaatiman laskennan skaalaamiseksi IT-tiedekuntaan on rakennettu GPU-yksiköllä varustettu tietokone sekä luotu GPU-palvelin syväoppimisverkkoja hyödyntävien sovellusten opettamisen nopeuttamiseksi.

Kunkin aineiston tutkiminen on noudattanut pääosin seuraavaa prosessia:

1. Tarpeen tunnistaminen ja tavoitteen asettaminen
2. Aineiston luovuttamisesta sopiminen ja siirto
3. Menetelmien ja mallien kartoittaminen (ml. IBM ratkaisut ja avoimet kirjastot)
4. Datan mallintaminen
5. Tulosten tulkinta yhteistyössä datan omistajatahon edustajan kanssa
6. Tulosten raportointi

3.1 Osaraportit

Seuraavassa kuvataan lyhyesti osaraporttien pääkysymykset ja -tulokset. Projektissa tuotettiin yhteensä yhdeksän teknistä tutkimusraporttia, joista yksi on kirjallisuuskatsaus ja lopuissa raportoidaan koneoppimistestien tuloksia käytännön datalla.

1. Girka, A., Kulmala, J-P., & Äyrämö, S., **Prediction of ground reaction force impact peak with deep learning.** Biomekaanista 3D-liikeanalyysia käytetään sekä sairaaloissa että liikuntabiomekaniikassa eri liikkumismuotojen ja työasentojen kuormittavuuden mittaamiseen sekä sairauksien monitorointiin. Tässä tutkimuksessa koulutettiin syvä CNN-tyyppinen neuroverkko tunnistamaan 76 %:n tarkkuudella juoksuaskeleesta rasitusvammariskiä lisäävä kuormituspiikki.

2. Jauhiainen, S., Forsman, H., Äyrämö, S., & Kauppi, J-P., **Talent detection in soccer using a One-Class Support Vector Machine.** Poikkeamantunnistusmenetelmiä voidaan soveltaa populaation valtavirrasta poikkeavien ominaisuuksien, oireiden tai käyttäytymisen havaitsemiseen. Tässä tutkimuksessa koulutettiin One-Class-SVM-menetelmä tunnistamaan ulkomaille edenneet nuoret jalkapalloharrastajat tuhansien pelaajien joukosta.

3. Moilanen, H., Kankaanranta, M. & Äyrämö, S., **Detecting pupils' preferred learning styles and different types of personalities by unsupervised machine learning.** Ohjaamattoman koneoppimisen avulla voidaan tunnistaa koululaisten joukosta eri oppimistyylin omaavia alijoukkoja. Tässä tutkimuksessa löydettiin yläkoululaisten joukosta kolme erilaista oppijaprofiilia käyttäen opetusdatana liikunnallisen työpajaopetuksen yhteydessä suoritettua kyselyaineistoa. Tietoa voidaan hyödyntää yksilöllisempien opetusmuotojen suunnittelemisessa.

4. Moilanen, H., Jauhiainen, S., & Äyrämö, S., **Predicting school success from student's wellbeing data.** Liikunnallisella aktiivisuudella on todettu olevan yhteys koulumenestykseen. Tässä tutkimuksessa selvitettiin henkilökohtaisin sensorein kerättyjen mittausten potentiaalia koulumenestyksen ennustamisessa. Koulumenestystä ennustavia tekijöitä ei kuitenkaan koneoppimisen keinoin onnistuttu tunnistamaan. Tulokseen vaikutti suurella todennäköisyydellä merkittävästi otoksen varsin pieni koko ja lyhyt seurantajakso.

5. Niinimäki. E., Pasanen, K., Leppänen, M., Vasankari, T., Parkkari, J., & Äyrämö,S., **Determining risk factors for acute lower extremity injuries in team sports with lasso regression.** Nuorena koettujen alaraajavammojen on todettu lisäävän nivelrikon, ylipainon ja alentuneen elämänlaadun riskiä. Tämän tutkimuksen tavoitteena oli etsiä uusia

prediktoreita polvi- ja nilkkavammariskille UKK-instituutin keräämästä liikuntavammatutkimusaineistosta. Aineistosta löydettiin teorian pohjalta tunnetut riskitekijät sekä alustavia indikaatioita uusista prediktoreista. Uusien prediktorien löytäminen ja vahvempien mallien opettaminen edellyttänee lisämuuttujien mukaanottamista.

6. Ojala, T., Wirta, E-V., Äyrämö, S., Kuopio, T., & Pölönen, I., **Cell counting with gradient boosted convolutional neural networks.** Histopatologisten kuvien tulkinta vaatii paljon resursseja. Tässä tutkimuksessa selvitettiin syvien konvolutiivisten neuroverkkojen soveltuvuutta syöpäsolujen lukumäärän laskemiseen suolistokuvista. Testiaineistossa oli keskimäärin 122 syöpäsolua ja parhaan mallin absoluuttinen keskivirhe oli 16,2.

7. Rautiainen, I., Joensuu, L., Tammelin, T., Kujala, U., & Äyrämö, S., **Finding predictors of aerobic fitness level change in children and adolescents using machine learning.** Nuorten liikunta-aktiivisuus putoaa iän myötä. Tämän tutkimuksen tavoitteena oli selvittää koneoppimisen potentiaalia kestävyyskunnon ennustamisessa ja prediktorien valinnassa. Kestävyyskunnon ennustamista testattiin useilla menetelmillä, mutta valittujen muuttujien ennustusvoima oli vain välttävä. Uusia systemaattisia prediktoreita ei onnistuttu tunnistamaan. Lineaaristen mallien ennustusvoima todettiin riittäväksi nykyisessä aineistossa eivätkä kompleksisemmat mallit parantaneet tuloksia.

8. Rautiainen, I. & Äyrämö, S., **Predicting overweight and obesity in later life from childhood data: A survey.** Lasten ylipainon lisääntyminen on vakava kansanterveysriski, joka kuormittaa myös kansantaloutta. Tässä raportissa kartoitettiin mitä koneoppimisen menetelmiä on sovellettu lasten ylipainon ennustamiseen. Tuloksia voidaan hyödyntää kehitettäessä uusia kotimaisiin kasvukäyräaineistoihin perustuvia tekoälymenetelmiä ylipainon varhaiseen ennustamiseen.

9. Saarela, M., Ryynänen, O-P, & Äyrämö, S., **Predicting hospital associated disability using supervised learning.** Tässä tutkimuksessa vertailtiin koneoppimispohjaisia menetelmiä akuutisti sairastuneiden vanhusten hoitovasteen ennustamiseen. Useita malleja tutkien nykyisen aineiston ennustusvoimaisimmaksi prediktoreiksi havaittiin itsenäistä selviytymistä kuvaava muuttuja. Aineisto todettiin vielä riittämättömäksi ennustemallien luomiseen ja myös lisämuuttujia täytyy vielä kerätä.

4 Pohdinta

Projektin aikana saatiin mallien osalta testattua loppuun asti kahdeksan testitapausta. Testeissä saavutettiin kohtuullisia koneoppimistuloksia testitapausten luokittelu- ja ennustustehtävien osalta vaikka aineistot ovat vielä varsin yleisesti liian pieniä tarkkojen ja yleistymiskykyisten mallien opettamiseen. Joissakin tapauksissa pohdittiin myös, olivatko käytetyt prediktorikandidaatit relevantteja ennustustehtävien kannalta. Mallien yleistymiskyky voidaan varmistaa tehokkaimmin suurella relevantilla aineistolla ja yksinkertaisimmalla datan sisältämän informaation kuvaavalla ratkaisulla.

Usein asiantuntijaorganisaatioiden keräämiä aineistoja on tutkittu teoriatietoon perustuvien hypoteesien ja selittävän tilastotieteen keinoin. Tämän projektin tutkimuksissa aineistojen hyödyntämistä on lähestytty ohjatun koneoppimisen kautta, mikä mahdollistaa sekä uuden tiedon tuottamisen datan sisältämistä monimutkaisista riippuvuuksista syöte- ja vastemuuttujien välillä että automaattiseen päättelyyn kykenevien tekoälysovellusten kehittämisen. Tekoälyn taustalla olevat algoritmit oppivat ennustusmalleja, tunnistavat prediktoreita ja pystyvät etsimään aineistoista jopa niin monimutkaisia datapohjaisia hypoteeseja, että niiden muodostaminen voi olla mahdotonta ihmiselle (Shmueli et al., 2010). Koneoppimismenetelmien täyden potentiaalin hyödyntämiseksi

tekoälyratkaisujen kehittämisessä onkin aineistojen valinnassa tärkeää pyrkiä välttämään rajoittumista liian vahvoihin aiempaan tietämykseen perustuviin oletuksiin. Tuloksia tulkittaessa yhdessä asiantuntijoiden kanssa päädyttiinkin usein siihen, että jatkossa menetelmien opettamiseen tarvittavaa dataa pyritään valitsemaan entistä laajemmin ja olettamuksia välttäen.

Tekoäly- ja koneoppimisalgoritmit/-menetelmät ovat kattavasti yritysten ja tutkijoiden saatavilla avoimien kirjastojen kautta, jolloin voidaan helposti räätälöidä sovelluskohtaisia ratkaisuja. Tekoälymenetelmien räätälöiminen esitettyihin ongelmiin ja mallin opettaminen kohtuullisessa ajassa ovat harvoin kynnyskysymyksiä käytännön kehityshaasteiden liittyessä enemminkin datan määrään ja laatuun. Epätäydellinen data, eli datassa on puuttuvia arvoja, on edelleen haaste joidenkin koneoppimistyökalujen kohdalla.

Nykyään kynnys puhua big datasta on varsin matala. Nykyiset terveysdatalähteet ovat vielä varsin pirstaleisia, sillä ne ovat usein yritysten ja organisaatioiden keräämiä yksittäisiä otoksia. Etenkin julkisella puolella olisi tärkeää pyrkiä keskitettyyn ja koordinoituun datan keräämiseen. Mitä pienempiin kohderyhmiin mennään sitä tärkeämpää on datan keskitetty yhdistäminen. Muutoin datan riittävyys tulee rajoittamaan tekoälymallien opettamista ja käyttöä tulevaisuudessakin. Tämä pätee etenkin trendiksi nousseiden syvien neuroverkkojen opettamiseen. Suomessa on teknologisen ja terveysalan osaamisen sekä Omakannan Omatietovarannon kaltaisten ympäristöjen kehittymisen myötä erinomaiset edellytykset tuottaa älykkäitä dataa hyödyntäviä sovelluksia.

Viitteet

Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., 2017. Application of deep convolutional neural network for automated detection of myocardial infarction using ecg signals. Information Sciences 415, 190–198.

Altilio, R., Paoloni, M., Panella, M., 2017. Selection of clinical features for pattern recognition applied to gait analysis. Medical & biological engineering & computing 55 (4), 685–695.

Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O., 2016. Deep learning for computational biology. Molecular systems biology 12 (7), 878.

Äyrämö, S., 2006. Knowledge mining using robust clustering. University of Jyväskylä.

Bae, J.-M., 2014. The clinical decision analysis using decision tree. Epidemiology and health 36.

Baxt, W. G., 1995. Application of artificial neural networks to clinical medicine. The lancet 346 (8983), 1135–1138.

Bisenius, S., Mueller, K., Diehl-Schmid, J., Fassbender, K., Grimmer, T., Jessen, F., Kassubek, J., Kornhuber, J., Landwehrmeyer, B., Ludolph, A., et al., 2017. Predicting primary progressive aphasias with support vector machine approaches in structural mri data. NeuroImage: Clinical 14, 334–343.

Bishop, C. M., 2006. Pattern recognition and machine learning. springer.

Breiman, L., et al., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16 (3), 199–231.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Vol. 1. Springer series in statistics New York.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Hannink, J., Kautz, T., Pasluosta, C. F., Gaßmann, K.-G., Klucken, J., Eskofier, B. M., 2017. Sensor-based gait parameter extraction with deep convolutional neural networks. IEEE journal of biomedical and health informatics 21 (1), 85–93.

Haykin, S. S., Haykin, S. S., Haykin, S. S., Haykin, S. S., 2009. Neural networks and learning machines. Vol. 3. Pearson Upper Saddle River, NJ, USA:.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural networks 4 (2), 251–257.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lehto, M., Neittaanmäki, P., 2017. Suomen terveysdataympäristö. Informaatioteknologian tiedekunnan julkaisuja (35).

Linda, A. W.-M., 2014. Seven ways predictive analytics can improve healthcare. published in Elsevier.

Niinimäki, E., Pöyhönen, J., Äyrämö, Neittaanmäki, 2017. Omadata terveydenhuollon tietointensiivisessä rakenteessa. Informaatioteknologian tiedekunnan julkaisuja (40).

Nilsson, N. J., 1998. Artificial intelligence: a new synthesis. Elsevier.

Nixon, M. S., Aguado, A. S., 2012. Feature extraction & image processing for computer vision. Academic Press.

Pershad, Y., Govindan, S., Hara, A. K., Borad, M. J., Bekaii-Saab, T., Wallace, A., Albadawi, H., Oklu, R., 2017. Using naïve bayesian analysis to determine imaging characteristics of kras mutations in metastatic colon cancer. Diagnostics 7 (3).

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for eeg decoding and visualization. Human brain mapping 38 (11), 5391–5420.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks 61, 85–117.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., 2017. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annual Review of Biomedical Engineering (0).

Shmueli, G., et al., 2010. To explain or to predict? Statistical science 25 (3), 289–310.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research 15 (1), 1929–1958.

Sutton, R. S., Barto, A. G., 1998. Reinforcement learning: An introduction. Vol. 1. MIT press Cambridge.

Takenaka, S., Aono, H., 2017. Prediction of postoperative clinical recovery of drop foot attributable to lumbar degenerative diseases, via a bayesian network. Clinical Orthopaedics and Related Research® 475 (3), 872–880.

Zemmour, C., Bertucci, F., Finetti, P., Chetrit, B., Birnbaum, D., Filleron, T., Boher, J.-M., 2015. Prediction of early breast cancer metastasis from dna microarray data using high-dimensional cox regression models. Cancer informatics 14 (Suppl 2), 129.