



Vaasan yliopisto
UNIVERSITY OF VAASA

Matias Parkkinen

Koneoppiminen ja tekoäly sijoittajan työkaluina

Tarkastelussa uhat ja mahdollisuudet

Laskentatoimen ja rahoituksen
akateeminen yksikkö
Taloustieteen
kandidaatintutkielma
Koulutusohjelma

Vaasa 2026

VAASAN YLIOPISTO**Laskentatoimen ja rahoituksen akateeminen yksikkö**

Tekijä:	Matias Parkkinen
Tutkielman nimi:	Koneoppiminen ja tekoäly sijoittajan työkaluina: Tarkastelussa uhat ja mahdollisuudet
Tutkinto:	Kauppätieteiden kandidaatin tutkinto
Oppiaine:	Taloustiede
Työn ohjaaja:	Petri Kuosmanen
Valmistumisvuosi:	2026
Sivumäärä:	46

TIIVISTELMÄ:

Rahoitusmarkkinoiden teknologinen murros, erityisesti koneoppimisen ja generatiivisen tekoälyn nopea kehitys, asettaa sijoittajalle sekä uusia analyyttisiä mahdollisuuksia että tunnistamattomia uhkia. Ymmärrys siitä, mitä nämä teknologiat käytännössä voivat tehdä ja missä niiden käyttö voi mennä pieleen, on noussut keskeiseksi taidoksi niin yksityissijoittajille kuin rahoitusalan ammattilaisille.

Tässä kandidaatintutkielmassa selvitetään kirjallisuuskatsauksen menetelmin, miten koneoppiminen soveltuu osakekurssien ennustamiseen ja miten generatiivinen tekoäly täydentää näitä kvantitatiivisia menetelmiä laadullisen informaation käsittelyssä. Tutkimuskysymys kattaa molempien teknologioiden mahdollisuudet ja ongelmat sijoittajan näkökulmasta. Tiedonhankinta on toteutettu kansainvälisiä tieteellisiä tietokantoja ja vertaisarvioituja lähteitä hyödyntäen.

Kirjallisuuskatsauksen perusteella koneoppimismallit, erityisesti LSTM-pohjainen syväoppiminen ja ensemble-menetelmät, kykenevät tuottamaan tilastollisesti merkitsevää ennustetarkkuutta lyhyen aikavälin osakekurssikehityksessä. Generatiivinen tekoäly tarjoaa puolestaan systemaattisen väylän laadulliseen informaatioon, jota perinteiset numeeriset mallit eivät tavoita. Teknologiat täydentävät toisiaan, eivät kilpaile keskenään. Kumpaankin liittyy kuitenkin kriittisiä rajoitteita: koneoppimisessa ylisovittuminen ja läpinäkyvättömyys, generatiivisessa tekoälyssä hallusinaatioalttius. Näiden ymmärtäminen on ehdoton edellytys vastuulliselle soveltamiselle.

AVAINSANAT: koneoppiminen, tekoäly, osakekurssien ennustaminen, suuret kielimallit, sijoittajan työkalut, sentimenttianalyysi

Sisällys

1	Johdanto	5
1.1	Tavoite ja tutkimusmenetelmät	6
1.2	Tutkimuksen rakenne	9
2	Koneoppimisen määrittely	10
2.1	Koneoppimismallit ja kolmijako	11
2.1.1	Ohjattu oppiminen	11
2.1.2	Ohjaamaton oppiminen	13
2.1.3	Vahvistusoppiminen	13
2.2	Syväoppiminen ja neuroverkkoarkkitehtuurit	14
2.2.1	Keinotekoisien neuroverkon perusteet	14
2.2.2	Takaisinkytketyt neuroverkot ja LSTM	15
2.2.3	Konvoluutioneuroverkot	16
2.2.4	Transformerit ja tarkkaavaisuusmekanismit	17
2.3	Opetusmenetelmät ja metodologiset näkökohdat	17
2.3.1	Ylisovittaminen ja regularisointi	18
2.3.2	Mallin valinta ja hyperparametrien viritys	19
2.3.3	Ensemble-menetelmät	19
2.3.4	Käytännön haasteet ja ”musta laatikko” -ongelma	20
3	Osakekurssien ennustaminen koneoppimismallien avulla	22
3.1	Rahoitusaikasarjojen ominaispiirteet ja ennustamisen haasteet	23
3.2	Neuroverkot ja ei-lineaarinen mallinnus	25
3.2.1	Monikerroksiset perceptron-verkot	25
3.2.2	Rekurrentit verkot ja LSTM-arkkitehtuurit	26
3.3	Ensemble-menetelmät ja korkeaulotteinen osakedata	27
3.4	Vertailu klassisiin malleihin	28
3.5	Tutkimustulosten ja menetelmien arviointi	28
3.6	Tulosten analyysi	29
4	Generatiivinen tekoäly sijoittajan työkaluna	32

4.1	Transformer-arkkitehtuurista suuriin kielimalleihin	32
4.2	Suurten kielimallien sovellukset rahoitusmarkkinoilla	33
4.3	Generatiivinen tekoäly suhteessa perinteiseen koneoppimiseen	34
4.4	Generatiivisen tekoälyn rajoitteet ja riskit sijoituskontekstissa	35
4.5	Yhteenveto	37
5	Johtopäätökset	38
5.1	Löydösten merkitys eri sidosryhmille	39
5.2	Jatkotutkimusmahdollisuudet ja tutkimuksen rajoitteet	41
	Lähteet	42

1 Johdanto

Koneoppiminen ja tekoäly eivät ole enää pelkästään teknologia-alan erikoisuus, vaan ne muuttavat perusteellisesti tapaa, jolla rahoitusmarkkinoilla tehdään päätöksiä. Algoritmit käsittelevät hinta-aineistoa, uutisvirtaa ja tilinpäätöksiä nopeudella, joka ylittää inhimillisen kapasiteetin moninkertaisesti. Yksityissijoittajan saataville on avautumassa analyyttisiä valmiuksia, jotka aiemmin edellyttivät institutionaalisten toimijoiden resursseja. Tämä kehitys ei kuitenkaan ole yksiviivainen mahdollisuus: teknologioihin liittyy systemaattisia heikkouksia, joiden tunnistaminen on keskeistä niiden vastuulliselle hyödyntämiselle. Tässä tutkielmassa systematisoidaan, mitä koneoppiminen ja generatiivinen tekoäly voivat sijoittajalle käytännössä tarjota ja missä niiden käyttö voi mennä pieleen.

Analyysi ja päätöksenteko rahoitusmarkkinoilla ovat kokeneet merkittävän murroksen teknologian ja datan käytön kehittyessä. Tekoäly ja koneoppiminen (ML, Machine Learning) ovat teknologisista työkaluista keskiössä suurien datamäärien käsittelykykynsä ansiosta. Tämä on erityisen merkityksellistä nykyisellä big datan aikakaudella. Big datalla (tai massadatalla) tarkoitetaan kaikkea sitä tietoa, joka tallentuu lähestulkoon jokaisesta tekemästämme liikkeestä digitaalisilla alustoilla järjestelemättömiin datamassoihin, joita puolestaan hyödyntämällä voimme saada keskeistä tietoa niin rahoitusinstrumenttien käyttäytymisestä kuin myös makrotaloudellisista suhdanteista (Aldridge & Avellaneda, 2021).

Määrällisen käsittelykykynsä lisäksi koneoppiminen mahdollistaa epälineaaristen kuvioiden ja monimutkaisesti rakentuneen datan hyödyntämisen siinä, missä perinteiset lineaariset mallit olisivat samaan tehtävään liian kömpelöitä (Bartram ym., 2021). Osakemarkkinoiden dynaaminen luonne vaikeuttaa staattisten analyysimallien soveltamista, mutta syväoppimisarkkitehtuurit, erityisesti toistuvat neuroverkot ja LSTM (Long Short-Term Memory), ovat osoittaneet merkittävää ennustevoimaa lyhyen aikavälin osakekurssimuutoksissa (Fischer & Krauss, 2018).

Koneoppimisen rinnalle on noussut generatiivinen tekoäly, jonka keskeisimpänä ilmentymänä toimivat suuret kielimallit (Large Language Models, LLM). Siinä missä perinteinen koneoppiminen on optimoitu strukturoidun numeerisen datan analyysiin, suuret kielimallit on suunniteltu luonnollisen kielen ymmärtämiseen ja tuottamiseen, mikä avaa sijoittajalle pääsyn laadullisiin informaatiolähteisiin, kuten tilinpäätöksiin, tulospuheluihin ja uutisvirtaan (Brown ym., 2020). Koneoppiminen ja generatiivinen tekoäly muodostavat siten toisiaan täydentävän kokonaisuuden, ja tätä kokonaisuutta tarkastellaan tässä tutkielmassa sijoitustoiminnan näkökulmasta.

Käytännön tasolla tämä tarkoittaa, että koneoppiminen voi tunnistaa toistuvia tilastollisia kuvioita historiadatasta ja rakentaa niiden perusteella ennustemalleja, mutta se ei pysty ennakoimaan ennennäkemättömiä markkinashokeja. Generatiivinen tekoäly puolestaan voi tiivistää tulosraportteja, luokitella uutisvirtaa ja vastata sijoittajan kysymyksiin luonnollisella kielellä, mutta saattaa samalla tuottaa sisällöllisesti uskottavaa mutta faktuaalisesti virheellistä tietoa. Mahdollisuuksien ja riskien tasapainoinen arviointi on se viitekehys, jonka tämä tutkielma pyrkii tarjoamaan.

1.1 Tavoite ja tutkimusmenetelmät

Tämä kandidaatintyö on kirjallisuuskatsaus, joka tarkastelee, miten koneoppiminen ja generatiivinen tekoäly soveltuvat sijoittajan työkaluiksi. Koneoppiminen käydään läpi sekä teoreettiselta pohjalta että sen soveltamisen kautta osakekurssien ennustamisessa. Lisäksi tarkastellaan, miten generatiivinen tekoäly, etenkin suuret kielimallit, täydentää koneoppimismenetelmiä laadullisen informaation käsittelyssä. Tavoite on arvioida molempien teknologioiden mahdollisuuksia ja rajoitteita rahoitusalan näkökulmasta sekä verrata niitä perinteisesti käytettyihin analyysimenetelmiin. Työ pyrkii vastaamaan tutkimuskysymykseen:

Miten sijoittaja voi hyötyä koneoppimisen ja generatiivisen tekoälyn soveltamisesta osana sijoitusprosessia, mitä ongelmia näiden teknologioiden käyttämiseen liittyy, ja miten koneoppimismenetelmät suhteutuvat perinteisiin analyysimenetelmiin ja generatiiviseen tekoälyyn sijoittajan näkökulmasta.

Tiedonhankinta tätä kandidaatintyötä varten on tehty hyödyntäen Google Scholar hakukonetta, Tritonian Finna e-kirjastoa ja Tritonian LibGuidesista löytyviä tietokantoja, kuten Scopus ja SpringerLink. Tämän lisäksi relevanttien tieteellisten lähteiden löytämistä varten, lähteiden itsensä tutkimiseen ja oikeinkirjoituksen tarkistamiseen on hyödynnetty generatiivista tekoälyä (ChatGPT, ScholarGPT, Google Gemini, Claude Sonnet 4.5 ja ChatPDF). Kaikki tiedonhakutyö tutkielmaa varten on suoritettu englanniksi hakutulosten maksimoimiseksi.

Tieteellisiä lähteitä tätä tutkielmaa varten löytyi runsaasti, ja tiedonhakua on suoritettu sekä aika- että relevanssijärjestyksessä. Lähteiksi on priorisoitu aiheesta julkaistut ja vertaisarvioidut kirjat, joita seuraa luotettaviin ja arvostettuihin alustoihin julkaistut tieteelliset artikkelit. Tärkeä osa tätä kandityötä on generatiivisen tekoälyn tehokas hyödyntäminen osana tutkimustyötä. Relevanttien lähteiden löytäminen ja tieteellisten tekstien informaation omaksuminen vaatii generatiivisen tekoälyn kouluttamista antamalla palautetta ja uudelleenkysymällä oikeita kysymyksiä (prompting). Taulukoihin 1 ja 2 on koottu käytetyimmät generatiiviselle tekoälylle annetut kehotteet (promptit) sekä hakukoneissa käytetyt hakusanat tutkielman toistettavuuden varmistamiseksi. Tekoälyn ehdottamat lähteet on hakujen jälkeen tutkittu perusteellisesti generatiivisen tekoälyn hallusinoinnin estämiseksi.

Taulukko 1. Tiedonhaussa käytetyt generatiivisen tekoälyn kehotteet (promptit)

Teach me everything there is to know about X and only use peer reviewed scientific sources. Cite your sources.	Find me the most relevant scientific and peer reviewed sources about machine learning in investments
Teach me all the key points about this pdf	Prioritize books, then peer reviewed articles and find me the most important sources for my thesis about machine learning in stock price predictions, and artificial intelligence in the investment process
Give me an in depth breakdown of the following article/book	What are the most important scientific and peer reviewed sources for machine learning and artificial intelligence in investing?
Find me the pages from this book that cover subject x	Find me scientific AND peer reviewed sources for my bachelors thesis; Machine learning and artificial intelligence as an investors tool; Literature review about stock price predictions with machine learning models and the use of artificial intelligence as a part of the investment process
How can machine learning/artificial intelligence be applied to x? ONLY use scientific peer reviewed sources and give me the links to all the sources used.	Give me more examples and this time focus on X X=(/machine learning/neural networks/portfolio optimization/stock price prediction/investing artificial intelligence/)

Taulukko 2. Tietokantahauissa käytetyt hakusanat

Machine learning AND investments	Machine learning AND portfolio
Machine learning AND arbitrage	Machine learning AND stocks
Machine learning AND finance	Machine learning AND investing
Neural networks AND stocks	Neural networks AND finance
Artificial intelligence AND investing	Artificial intelligence AND stocks
LLM AND investing	LLM AND stocks
Deep learning AND portfolio	Deep learning AND stock prices
Algorithms AND investing	Statistics AND finance
Data science AND investing	Data science AND finance

Sekä hakukoneissa että generatiivisen tekoälyn kanssa hakusanoista ja prompteista on käytetty myös niiden yhdistelmiä ja variaatioita.

1.2 Tutkimuksen rakenne

Tämä kandidaatintutkielma jatkuu johdannon jälkeen koneoppimisen määritelmään ja eri koneoppimismalleihin syventymällä. Luvun tarkoitus on rakentaa teoreettinen pohja, jossa lukijalle muodostuu ymmärrys koneoppimisen perusperiaatteista ja keskeisistä arkkitehtuureista. Kolmannessa luvussa tutkitaan osakekurssien ennustamista koneoppimisen avulla: analysoidaan, millaisia malleja ja rakenteita kirjallisuudessa on hyödynnetty ennustamistehtävissä, millaisia haasteita rahoitusmarkkinoiden ennustettavuuteen liittyy, ja mitä koneoppimiskirjallisuuden tulokset kertovat suhteessa perinteisiin menetelmiin. Neljäs luku on omistettu generatiiviselle tekoälylle sijoittajan työkaluna: luvussa käydään läpi suurten kielimallien kehitys, konkreettiset sovellukset rahoitusmarkkinoilla, sekä niiden rajoitteet ja riskit kriittisesti arvioiden. Viidennessä luvussa tehdään johtopäätökset, joissa tarkastellaan molempien teknologioiden laajempaa merkitystä eri sidosryhmien, kuten yksityissijoittajien, finanssialan instituutioiden, sijoituspalveluammattilaisten ja alan opiskelijoiden, näkökulmasta.

2 Koneoppimisen määrittely

Koneoppiminen on yksi 2000-luvun vaikuttavimmista teknologioista, joka on muuttanut perusteellisesti tapaa, jolla tietokoneet prosessoivat informaatiota ja tekevät päätöksiä. Tämän kappaleen tarkoituksena on tarjota lukijalle ymmärrys koneoppimisen perusperiaatteista ja eri koneoppimisarkkitehtuureista. Kappale toimii teknisenä perustana myöhemmin tutkielmassa esiteltäville sovelluksille rahoituksen alalla. Perusteiden ymmärtäminen on olennaista, jotta koneoppimisen menetelmiä voidaan arvioida sijoitustoiminnan ja markkinaennusteiden kontekstissa.

Koneoppiminen on tietotekniikan osa-alue, joka tutkii algoritmien ja tilastollisten mallien kehittämistä ja niiden vaikutusta siihen, miten tietokoneohjelma pystyy parantamaan suoriutumiskykyään tietyssä tehtävässä ilman uutta ohjelmointia annetun datan perusteella. Eräs klassinen ja yleisesti käytetty määritelmä koneoppimiselle on:

”Tietokoneohjelman sanotaan oppivan kokemuksesta E tietyn tehtäväluokan T suhteen ja suorituskymittarin P mukaan, jos sen suorituskky tehtävissä T , mitattuna mittarilla P , paranee kokemuksen E myötä.” (Mitchell, 1997, s. 2; suomennos tekijän)

Määritelmästä on hyvä nostaa esille oppimisen mitattavuus, joka erottaa koneoppimisen muusta sääntöpohjaisesta ohjelmoinnista.

Koneoppiminen voidaan nähdä myös tilastollisena haasteena, jossa syötteiden (*input*) ja vasteiden (*output*) välistä riippuvuutta pyritään estimoimaan funktiolla siten, että sama malli on myöhemmin yleistettävissä aikaisemmin näkemättömään dataan (Hastie ym., 2009). Käytetyimmät menetelmät tämän saavuttamiseksi ovat yleensä tilastollinen päättely, todennäköisyysteoria ja optimointi. Jotta saatu malli olisi laajemmin käyttökelpoinen, keskiössä ovat joustavuus ja yleistettävyys.

Koneoppiminen voidaan ymmärtää myös tilastollisen mallinnuksen evoluutiona. Kuten Goodfellow ym. (2016) toteavat: *”koneoppiminen on pohjimmiltaan sovellettua*

tilastotiedettä, jossa painotetaan tietokoneiden käyttöä monimutkaisten funktioiden tilastolliseen estimointiin ja vähennetään painoarvoa näiden funktioiden luottamusvälien todistamiselta” (s. 96).

Tästä määritelmästä voidaan tulkita, että vaikka koneoppiminen jakaa perusteita klassisen tilastotieteen kanssa, se asettaa laskennalliset lähestymistavat ja monimutkaisten suhteiden mallintamisen perinteisten päättelymenetelmien edelle. Koneoppimisen laskennallinen painotus mahdollistaa sellaisen datan analysoimisen, joka olisi mahdotonta perinteisillä tilastotieteellisillä menetelmillä.

2.1 Koneoppimismallit ja kolmijako

Tyypillisesti koneoppiminen jaetaan kolmeen alalajiin: ohjattu oppiminen (supervised learning), ohjaamaton oppiminen (unsupervised learning) ja vahvistusoppiminen (reinforcement learning) (Hastie ym., 2009). Ohjatussa oppimisessa mallia opetetaan syöte-vaste-pareilla. Tätä hyödynnetään yleisimmin luokittelu- ja regressiotehtävissä. Ohjaamattomassa oppimisessa vasteita ei anneta valmiiksi, vaan tavoitteena on antaa koneen tunnistaa datasta rakenteita ja ryhmiä. Vahvistusoppimisessa agentti vuorovaikuttaa dynaamisesti ympäristönsä kanssa. Agentille ei anneta suoraan toivottuja vasteita, vaan se ohjaa toimintaansa palautteen muodossa saatavien palkkioiden ja rangaistusten perusteella, pyrkien kokeilun ja erehdyksen kautta maksimoimaan kumulatiivisen hyödyn (Sutton & Barto, 2018).

2.1.1 Ohjattu oppiminen

Ohjattu oppiminen kattaa sisällään laajimmin käytetyt koneoppimismallit rahoituslallalla. Siinä algoritmeja opetetaan niin sanottujen merkittyjen aineistojen avulla, joissa jokaiselle syötehavainnolle on liitetty vastaava kohdemuuttuja. Oppimisen tavoitteena on muodostaa kuvausfunktio, joka pystyy ennustamaan uusien, aiemmin näkemättömien syötteiden tulokset mahdollisimman tarkasti koulutusaineistosta tunnistettujen rakenteiden perusteella. Tämä lähestymistapa soveltuu erityisen hyvin

rahoituksen ennustetehtäviin, joissa historiallista dataa tunnetuilla toteumilla on saatavilla (Bishop, 2006).

Ohjattu oppiminen jakautuu kahteen päätehtäväluokkaan: luokitteluun ja regressioon. Luokittelutehtävissä ennustetaan toisistaan eroavia kategorisia lopputuloksia, kuten nouseeko vai laskeeko osakkeen arvo tai onko yksittäinen maksutransaktio luonnoltaan epäilyttävä tai vilpillinen. Regressiotehtävissä puolestaan ennustetaan jatkuvia numeerisia arvoja, kuten rahoitusinstrumentin tulevaa hintaa tai sijoitusportfolion tuottoa. Valinta luokittelu- ja regressiomuotoilun välillä riippuu ennusteiden tavoitteesta ja siitä, miten tuloksia hyödynnetään sijoituspäätöksenteossa (Hastie ym., 2009).

Rahoituksessa yleisesti käytettyjä algoritmeja ovat muun muassa lineaarinen ja logistinen regressio, jotka tarjoavat tulkittavia perusmalleja. Tukivektorikoneet (Support Vector Machines, SVM), jotka soveltuvat monimutkaisten päätösrajojen muodostamiseen; päätöspuut ja random forest -menetelmät, jotka kykenevät mallintamaan epälineaarisia riippuvuuksia; gradienttitehostusmenetelmät (esim. XGBoost), jotka saavuttavat usein huipputason ennustustarkkuuden; sekä neuroverkot, jotka voivat mallintaa erittäin monimutkaisia rakenteita, mutta edellyttävät usein suuria aineistoja ja huolellista regularisointia (Bartram ym., 2021). Ohjatun oppimisen tehokkuus riippuu ratkaisevasti koulutusaineiston laadusta ja edustavuudesta, mikä voi olla haastavaa rahoitusympäristössä, jossa markkinaolosuhteet muuttuvat.

Ohjatun oppimisen sovellukset rahoitusosalalla ulottuvat ennustetehtävien ulkopuolelle. Lee (2007) osoitti tukivektorikoneiden soveltuvan luottoluokitusten ennustamiseen yritysdatalla, saavuttaen paremman tarkkuuden kuin logistinen regressio tai neuroverkko vastaavalla aineistolla. Tulokset havainnollistavat koneoppimisen laajaa käyttöpotentiaalia rahoitusalan luokittelutehtävissä, joissa päätösrajat ovat epälineaarisia.

2.1.2 Ohjaamaton oppiminen

Ohjaamattomassa oppimisessa käytetään merkitsemättömiä aineistoja, joista malli pyrkii löytämään sisäisiä rakenteita, säännönmukaisuuksia ja keskinäisiä riippuvuuksia ilman ennalta määritettyjä kohdemuuttujia. Ohjaamattoman oppimisen algoritmit tutkivat dataa löytääkseen merkittäviä riippuvuuksia, vähentääkseen moniulotteisuutta tai havaitakseen poikkeamia. Tämä koneoppimisen alalaji on erityisen hyödyllinen rahoitusalan alueilla, joilla merkityn datan saaminen on haastavaa, kallista, tai joilla tutkiva analyysi on ensisijaista (Bishop, 2006).

Merkittävimpiä tekniikoita ohjaamattomassa oppimisessa ovat klusterointi (clustering) ja ulottuvuuden pienentäminen (dimensionality reduction). Klusterointialgoritmit, kuten k-means, jakavat aineiston samankaltaisten havaintojen ryhmiin. Rahoituksessa tätä on hyödynnetty esimerkiksi osakkeiden ryhmittelyyn kaupankäyntikäyttäytymisen perusteella tai sijoittajien segmentointiin. Ulottuvuuden pienentämismenetelmät, kuten pääkomponenttianalyysi (Principal Component Analysis, PCA) ja autoenkooderit, muuntavat korkeaulotteisen datan tiiviimpään muotoon säilyttäen olennaisen informaation. Näitä voidaan käyttää esimerkiksi riskitekijöiden tunnistamiseen ja kohinan vähentämiseen (Consoli ym., 2021).

2.1.3 Vahvistusoppiminen

Vahvistusoppiminen eroaa tyyliltään merkittävästi kahdesta edellisestä mallista. Sen sijaan, että oppiminen perustuisi staattiseen aineistoon, koneoppimisagentti kokeilee erilaisia toimintoja ja saa palautetta palkkioiden tai rangaistusten muodossa. Oppimisen tavoitteena on löytää strategia, joka maksimoi kumulatiivisen pitkän aikavälin palkkion. Tätä viitekehystä mallinnetaan usein Markovin päätösprosessin (*Markov Decision Process*, MDP) avulla, joka tarjoaa matemaattisen rakenteen peräkkäiseen päätöksentekoon epävarmuuden vallitessa (Sutton & Barto, 2018).

Vahvistusoppiminen on uutta tietoa etsivän ja vanhaa tietoa hyödyntävien toimintojen tasapainoa. Koneoppimisagentin on toisaalta kokeiltava uusia toimintoja (eksploraatio) ja toisaalta hyödynnettävä jo tunnettuja, toimiviksi havaittuja toimintatapoja (eksploraatio). Tämä jatkuvasti muuttuva prosessi tekee vahvistusoppimisesta erinomaisesti soveltuvan ongelmiin, joissa päätökset tehdään peräkkäin, kuten algoritmisessa kaupankäynnissä tai dynaamisessa portfolioalokaatiossa (Zhang ym., 2020). Viimeaikaiset edistysaskeleet syvävahvistusoppimisessa tarjoavat lupaavia mahdollisuuksia mukautuvien kaupankäyntistrategioiden kehittämiseen (Bartram ym., 2021).

2.2 Syväoppiminen ja neuroverkkoarkkitehtuurit

Syväoppiminen on koneoppimisen alahaara, joka perustuu keinotekoisiiin neuroverkkoihin (ANN), joissa on useita toisiinsa kytkettyjä solukerroksia (*nodes*). Tämä mahdollistaa datan hierarkkisten esitystapojen automaattisen oppimisen. Tutkimukset ovat osoittaneet, että syväoppimistekniikat, erityisesti takaisin syöttävät ja konvoluutioneuroverkot, päihittävät johdonmukaisesti perinteiset koneoppimismenetelmät rahoitusmarkkinoiden ennustetehtävissä (Consoli ym., 2021).

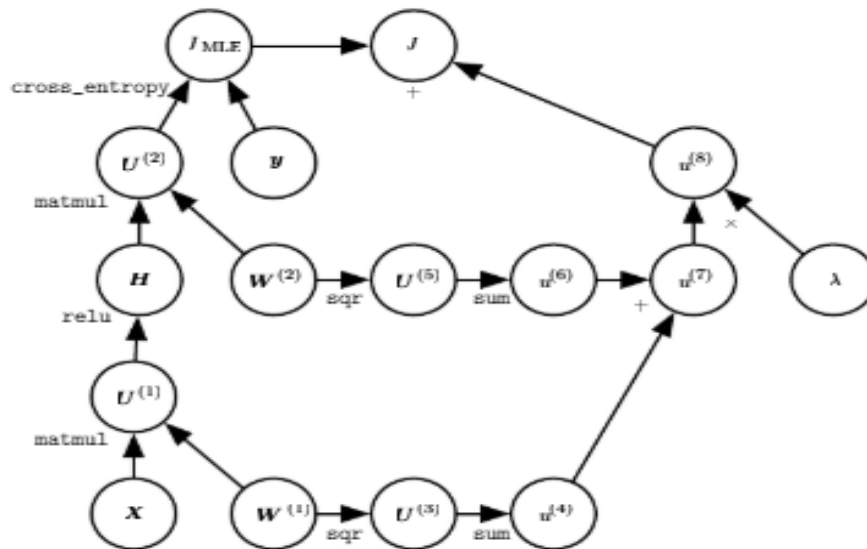
2.2.1 Keinotekoisien neuroverkon perusteet

Keinotekoiset neuroverkot (*Artificial Neural Networks*, ANN) koostuvat toisiinsa kytketyistä prosessointiyksiköistä, neuroneista, jotka on organisoitu kerroksiksi. Jokaisella kytkennällä on painokerroin w , joka säätelee signaalin voimakkuutta. Koulutusprosessissa näitä painoja säädetään vastavirta-algoritmilla (*backpropagation*), joka hyödyntää gradienttimenetelmää virhefunktion (*loss function*) minimointia varten (Goodfellow ym., 2016).

Tyypillinen neuroverkko käsittää syötekerroksen, yhden tai useamman piilokerroksen sekä ulostulokerroksen. Verkon syvyys, eli piilokerrosten lukumäärä, määrittää sen kyvyn oppia hierarkkisia rakenteita. Siinä missä matalat verkot ovat rajoittuneita, syvät verkot

taas kykenevät tavoittamaan yhä abstraktimpia piirteitä peräkkäisillä tasoilla (Bishop, 2006). Epälineaariset aktivointifunktiot, kuten ReLU (Rectified Linear Unit), mahdollistavat monimutkaisten suhteiden mallintamisen, joita ilman syväkin verkko vastaisi vain lineaarista mallia (Goodfellow ym., 2016).

Kuvio 1. Yhden piilokerroksen MLP-verkon kustannusfunktion laskentagraafi



Lähde: Goodfellow ym. (2016, s. 216).

Kuviossa 1 esitetty laskentagraafi havainnollistaa vaiheittain, miten eteenpäin syöttävän neuroverkon kokonaiskustannus muodostuu, kun yhden piilokerroksen monikerroksista perceptronin (MLP) opetetaan ristientropiahäviön ja painojen vaimennuksen (weight decay) avulla (Goodfellow ym., 2016).

2.2.2 Takaisinkytketyt neuroverkot ja LSTM

Takaisinkytketyt neuroverkot (*Recurrent Neural Networks*, RNN) on suunniteltu erityisesti sekvenssidatan käsittelyyn. Toisin kuin eteenpäin kytketyissä verkoissa, RNN-verkot ylläpitävät sisäistä tilaa eli muistia, joka tavoittaa tietoa sekvenssin aiemmista elementeistä (Goodfellow ym., 2016). Rahoitusmarkkinoilla, joilla data on luonteeltaan

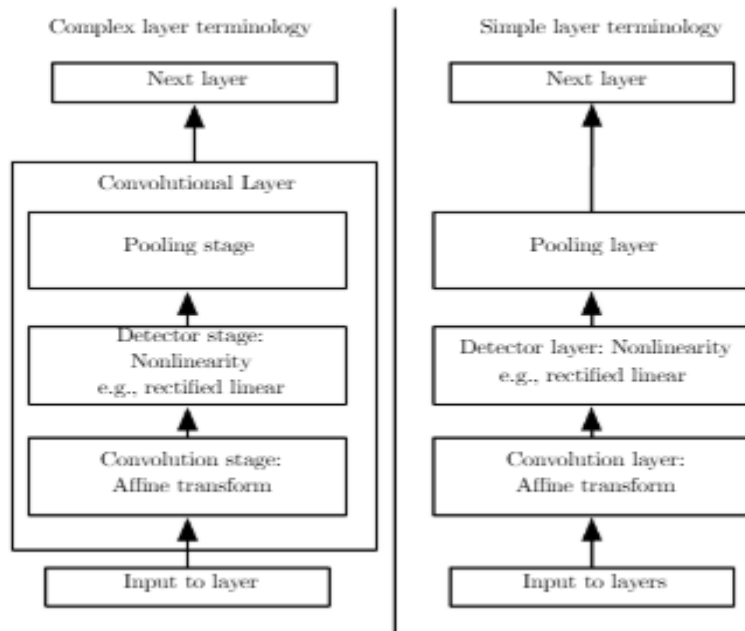
peräkkäistä ja aikariippuvaista, perinteiset mallit ovat usein riittämättömiä (Gu ym., 2020).

Tavalliset RNN-verkot kärsivät kuitenkin katoavan gradientin ongelmasta (vanishing gradient), mikä vaikeuttaa pitkän aikavälin riippuvuuksien oppimista. Hochreiter ja Schmidhuber (1997) kehittivät LSTM-verkoston (Long Short-Term Memory) ratkaisemaan tämän ongelman monimutkaisen porttimekanismin avulla. LSTM-yksikkö sisältää seuraavat portit: unohdusportin (forget gate), joka määrittää, mitä tietoa aiemmasta tilasta hylätään; syöteportin (input gate), joka kontrolloi uuden tiedon tallentamista; sekä ulostuloportin (output gate), joka säätelee solun tilasta välitettävää informaatiota (Hochreiter & Schmidhuber, 1997; Goodfellow ym., 2016). Ta ym. (2020) osoittavat, että LSTM-arkkitehtuuri on erityisen tehokas kvantitatiivisessa kaupankäynnissä, sillä se kykenee säilyttämään relevanttia markkinainformaatiota pitkien aikasarjojen yli samalla, kun se suodattaa pois merkityksetöntä kohinaa.

2.2.3 Konvoluutioneuroverkot

Konvoluutioneuroverkot (Convolutional Neural Networks, CNN) hyödyntävät konvoluutiokerroksia, jotka soveltavat opittuja suodattimia paikallisten rakenteiden havaitsemiseen. Keskeisiä innovaatioita ovat paikallinen kytkeytyvyys, painojen jakaminen ja koontioperaatiot (*pooling*), jotka tarjoavat invarianssia pienille siirtymille (Goodfellow ym., 2016). Konvoluutiokerroksen vaiheita, eli konvoluutiota, epälineaarista aktivointia ja koontia, havainnollistetaan kuviossa 2. Vaikka CNN-verkot kehitettiin kuvantunnistukseen, Consoli ym. (2021) toteavat niiden soveltuvan rahoitukseen, kun aikasarjoja käsitellään yksiulotteisina, tai kun teknisiä indikaattoreita visualisoidaan kuvamuotoon. Tutkimukset ovat osoittaneet, että CNN-kerrosten käyttö piirteiden erottamiseen yhdistettynä LSTM-kerrosten ajalliseen mallinnukseen voi parantaa ennustetarkkuutta merkittävästi verrattuna yksittäisiin arkkitehtuureihin (Consoli ym., 2021).

Kuvio 2. Konvoluutioneuroverkon kerroksen vaiheet kahden eri terminologian mukaan



Lähde: Goodfellow ym. (2016, s. 336).

2.2.4 Transformerit ja tarkkaavaisuusmekanismit

Transformer-lähestymistapa on tuore kehityssaskel, jossa takaisinkytketyt (rekurrentit) rakenteet korvataan tarkkaavaisuusmekanismilla (*attention mechanism*). Tämän ansiosta malli kykenee määrittämään eritasoisia tärkeysarvoja sekvenssin eri elementeille riippumatta siitä, kuinka kaukana toisistaan ne sijaitsevat. Tämä puolestaan mahdollistaa rinnakkaislaskennan ja pitkän aikavälin riippuvuuksien tehokkaamman käsittelyn verrattuna perinteisiin RNN-malleihin (Goodfellow ym., 2016). Aldridge ja Avellaneda (2021) huomauttavat, että vaikka Transformer-mallit vaativat suuria datamääriä, niiden kyky yhdistää tietoa useista erityyppisistä datalähteistä (kuten uutis- ja hintadatasta) tekee niistä lupaavan työkalun moderniin rahoitusanalyysiin.

2.3 Opetusmenetelmät ja metodologiset näkökohdat

Koneoppimismallien onnistunut soveltaminen rahoitusmarkkinoilla edellyttää huolellisuutta mallien opetus- ja validointiprosesseissa. Koska rahoitusdata on

luonteeltaan kohinaista ja epästationaarista, mallin ennustetarkkuus opetusdatassa ei takaa sen asianmukaista toimintaa todellisissa markkinaolosuhteissa.

2.3.1 Ylisovittaminen ja regularisointi

Eräs koneoppimisen keskeisistä haasteista on ylisovittaminen (*overfitting*). Siinä malli oppii opetusdatan liian tarkasti ja poimii siitä kohinaa sekä satunnaisia kaavoja, jotka eivät yleisty uuteen dataan. Tämän seurauksena malli suoriutuu erinomaisesti opetusdatassa, mutta epäonnistuu hyvien tulosten saavuttamisessa aiemmin näkemättömällä datalla. Ongelma on erityisen vakava rahoitussovelluksissa, sillä markkinoiden perustavanlaatuiset muutokset tarkoittavat, etteivät menneisyyden kuviot välttämättä toistu tulevaisuudessa. Koska rahoitusaineistot ovat erittäin korkeaulotteisia ja modernit mallit poikkeuksellisen joustavia, näennäisten säännönmukaisuuksien (*spurious patterns*) havaitsemisen riski on huomattava (Hastie ym., 2009).

Regularisointina tunnetut tekniikat on suunniteltu vähentämään ylisovittamista rajoittamalla mallin monimutkaisuutta. Yleisimpiä menetelmiä ovat L1- (lasso) ja L2- (ridge) regularisointi, joista kumpikin lisää virhefunktioon (*loss function*) sakkotermejä suurista parametriarvoista. L1-regularisointi edistää harvuutta (*sparsity*) ajamalla osan painokertoimista nolleen, jolloin se toimii samalla piirteiden valintana. Sitä vastoin L2-regularisointi hillitsee painokertoimien kasvua tasaisemmin. Neuroverkoissa käytetään usein Dropout-menetelmää, jossa neuroneja kytketään satunnaisesti pois päältä verkon opetuksen aikana, jotta estetään niiden liiallinen keskinäinen yhteissopeutuminen (*co-adaptation*) ja rohkaistaan kestävämpien piirteiden oppimista. Aikainen pysäyttäminen (*early stopping*) puolestaan katkaisee opetusprosessin heti, kun suorituskky validointidatassa alkaa heikentyä (Goodfellow ym., 2016).

Rahoituksessa asianmukainen validointimenetelmä on välttämätön, sillä perinteinen k-kertainen ristiinvalidointi (k-fold cross-validation) on harvoin soveltuva aikariippuvuuksien ja ennakkotiedon vinouman (look-ahead bias) vuoksi. "Walk-forward"-validointi, jossa malli opetetaan varhaisemmalla datalla ja testataan sitten sitä

seuraavalla ajanjaksolla, tarjoaa realistisempia arvioita suorituskyvystä. Tutkijoiden on kuitenkin oltava valppaina, sillä mallien toistuva hienosäätö validointitulosten perusteella voi johtaa epäsuoraan ylisovittamiseen. Tämän vuoksi lopullisen, täysin erillisen testijoukon (hold-out set) käyttö on välttämätön keino suojautua tältä riskiltä (Aldridge & Avellaneda, 2021).

2.3.2 Mallin valinta ja hyperparametrien viritys

Koneoppimismalleissa on tyypillisesti lukuisia hyperparametreja, eli asetuksia ja rakennevalintoja, jotka tutkijan on määritettävä ennen varsinaisen opetusprosessin alkamista ja joita malli ei opi suoraan datasta. Syvissä neuroverkoissa tällaisia ovat esimerkiksi piilokerrosten lukumäärä, gradienttioptimoinnin oppimisaskel (learning rate) sekä aktivaatiofunktioiden valinta. Koska kyseiset parametrit säätelevät mallin joustavuutta ja monimutkaisuutta, liian aggressiivinen tai validointidataan liiallisesti sovitettu parametrien optimointi kasvattaa merkittävästi ylisovittamisen riskiä (Bishop, 2006).

Vakiintuneisiin optimointimenetelmiin lukeutuu perinteinen ruudukkohaku (grid search), joka arvioi jokaisen ennalta määritellyn parametriyhdistelmän. Usein tehokkaammaksi vaihtoehdoksi muodostuu kuitenkin satunnaishaku (random search) tai edistyneempi bayesilainen optimointi, joka kykenee tilastollisesti ohjaamaan hakua kohti lupaavimpia parametriavaruuden osia. Rahoitusmarkkinoiden kontekstissa mallin ja hyperparametrien valinnan on oltava poikkeuksellisen vakaata: mallin on kyettävä yleistymään (generalization) myös historiallisesta aineistosta poikkeaviin markkinasuhdanteisiin (market regimes), eikä se saa pohjautua yksinomaan tietyn historiallisen ajanjakson ominaispiirteisiin (Hastie ym., 2009).

2.3.3 Ensemble-menetelmät

Sen sijaan, että pyrittäisiin löytämään yksi täydellinen malli, useiden eri mallien tai päätössääntöjen ennusteita voidaan yhdistää niin sanottujen ensemble-menetelmien

avulla. Lähestymistavan matemaattinen perusta lepää siinä, että useiden, osittain toisistaan riippumattomien mallien yhdistäminen kumoaa yksittäisten mallien tekemiä satunnaisia virheitä, mikä pienentää lopullisen ennusteen varianssia ja parantaa mallin vakautta (robustness).

Hastie ym. (2009) jakavat ensemble-menetelmät kolmeen pääkategoriaan. Bootstrap-aggregointi eli bagging pyrkii vähentämään varianssia opettamalla useita rinnakkaisia malleja alkuperäisen opetusdatan satunnaisotoksilla ja laskemalla lopulta niiden tuottamien ennusteiden keskiarvon (tunnetuimpana esimerkkinä satunnaismetsät eli Random Forests). Boosting puolestaan perustuu mallien peräkkäiseen opettamiseen, jossa kukin uusi iteraatio keskittyy korjaamaan aiempien heikkojen mallien (weak learners) tekemiä virheitä. Kolmas lähestymistapa, stacking, käyttää erillistä metamallia oppiakseen, mikä on optimaalinen tapa painottaa ja yhdistää eri perusmallien tuottamat ennusteet.

2.3.4 Käytännön haasteet ja ”musta laatikko” -ongelma

Vaikka edellä mainitut edistyneet menetelmät parantavat mallien ennustustarkkuutta, koneoppimisen käytännön soveltamiseen rahoitusallalla liittyy merkittäviä haasteita. Erityisesti syvissä neuroverkoissa esiintyy kriittinen 'musta laatikko' (black box) -ongelma: päätösten taustalla oleva monimutkainen päättelyketju on ihmiselle usein täysin läpinäkymätön (Bartram ym., 2021). Päätöksentekoprosessi voi olla rakenteeltaan joko täysin piilossa tai yksinkertaisesti liian korkeaulotteinen tulkittavaksi. Rahoitusmarkkinoilla tämä läpinäkymättömyys on vakava ongelma, sillä institutionaalisten sijoittajien ja sääntelyviranomaisten on pystyttävä ymmärtämään ja auditoimaan riskin taustatekijät, eikä pelkkä mallin tuottama lopputulema ole riittävä peruste sijoituspäätökselle.

Lisäksi koneoppimismallien toimivuus ja luotettavuus ovat täysin riippuvaisia niille syötetyn datan laadusta. Mikäli opetusdata on virheellistä, vinoutunutta (biased) tai ei

edusta riittävällä laajuudella erilaisia markkinaolosuhteita, mallin tuottamat tulokset ovat väistämättä puutteellisia (Bartram ym., 2021).

3 Osakekurssien ennustaminen koneoppimismallien avulla

Osakekurssien tarkka ennustaminen on ollut vuosikymmenten ajan yksi sijoittajien ja tutkijoiden keskeisimmistä tavoitteista. Samalla se kuuluu rahoitusalan kiistanalaisimpiin aiheisiin, sillä kysymys ennustettavuudesta kytkeytyy kiinteästi markkinoiden tehokkuuden teoriaan. Tehokkaiden markkinoiden hypoteesi (Efficient Market Hypothesis, EMH) esittää, että arvopapereiden hinnat heijastavat välittömästi kaikkea saatavilla olevaa informaatiota, minkä vuoksi systemaattisten ylituottojen saavuttaminen on mahdotonta ilman lisäriskin ottamista (Fama, 1970). Mikäli tämä oletus pitäisi ehdottomasti paikkansa, osakekurssien suunnan ennustaminen olisi periaatteessa mahdotonta.

Käytännön empiirinen tutkimus on kuitenkin osoittanut, ettei ajatus täydellisestä tehokkuudesta aina päde. Esimerkiksi momentum-ilmiö, erilaiset arvoanomaliat sekä riskipreemioihin kytkeytyvät systemaattiset tekijät osoittavat, että markkinahinnoissa voi esiintyä vähintäänkin tilapäistä ennustettavuutta. Akateemisen keskustelun ydin ei siten ole pelkästään siinä, voidaanko osakekursseja ennustaa, vaan pikemminkin siinä, missä määrin anomaalista ennustettavuutta voidaan hyödyntää, millä aikahorisonteilla ja millaisin menetelmin.

Koneoppimisen nopea yleistyminen rahoitustutkimuksessa vastaa suoraan tähän haasteeseen. Siinä missä perinteiset ekonometriset mallit nojaavat usein tiukasti määriteltyihin lineaarisiin oletuksiin ja kiinteisiin funktionaalsiin muotoihin, koneoppimismenetelmät ovat vahvasti datalähtöisiä. Ne kykenevät oppimaan monimutkaisia ja epälineaarisia riippuvuuksia suoraan havaintodatasta ilman voimakkaita teoreettisia ennako-oletuksia (Hastie ym., 2009). Koska rahoitusmarkkinoiden ilmiöt ovat tyypillisesti erittäin korkeaulotteisia ja dynaamisia, tämä menetelmällinen joustavuus muodostaa keskeisen teoreettisen perusteen koneoppimisen käytölle.

Koneoppimisen soveltaminen sijoitustoiminnassa ei siten ole pelkkä tekninen kysymys, vaan se heijastaa syvällisempää teoreettista näkemystä markkinoiden rakenteesta. Se tosiasia, että koneoppimismallit kykenevät systemaattisesti parantamaan ennustetarkkuutta ja tuottamaan riskikorjattua ylituottoa, viittaa markkinoilla esiintyviin monimutkaisiin säännönmukaisuuksiin, joita perinteiset lineaariset mallit eivät pysty tavoittamaan. Tässä kontekstissa koneoppiminen toimii paitsi laskennallisena ennustetyökaluna, myös tieteellisenä välineenä markkinadynamiikan ymmärtämiseksi.

Viime vuosina koneoppimisen ja markkinatehokkuuden välinen suhde onkin noussut laajemman teoreettisen kiinnostuksen kohteeksi. Kuten Gogas ja Papadimitriou (2021) painottavat, koneoppimisen arvo taloustieteessä ja rahoituksessa ulottuu huomattavasti pelkkää hintojen ennustamista pidemmälle: se tarjoaa välineitä taloudellisten taustarakenteiden, säännönmukaisuuksien ja sijoittajakäyttäytymisen tunnistamiseen. Tästä näkökulmasta osakekurssien ennustaminen asettuu luontevaksi osaksi laajempaa modernia datatieteellistä tutkimusperinnettä.

3.1 Rahoitusaikasarjojen ominaispiirteet ja ennustamisen haasteet

Osakekurssien ja -tuottojen mallintaminen eroaa luonteeltaan huomattavasti monista muista aikasarjaongelmista. Rahoitusaikasarjoissa signaali-kohinasuhde on tyypillisesti erittäin matala, jolloin markkinoiden satunnaisvaihtelu usein peittää alleen mahdolliset systemaattiset kuviot. Lisäksi tuottojen jakaumille ovat ominaisia paksut hännät (fat tails) ja volatilitetilla on taipumus klusteroitua ajallisesti (volatility clustering). Näiden ominaispiirteiden vuoksi perinteinen lineaarinen mallinnus osoittautuu usein riittämättömäksi.

Klassiset ARIMA-mallit perustuvat oletukseen, että aikasarjaa voidaan kuvata autoregressiivisten (AR) ja liukuvien keskiarvojen (MA) komponenttien avulla. Menetelmän soveltaminen edellyttää joko datan alkuperäistä stationaarisuutta tai vähintään sarjan muuntamista stationaariseksi differensoimalla (Box & Jenkins, 1976). Rahoitusaikasarjoissa stationaarisuusoletus on kuitenkin yleensä vain likimääräinen.

Lisäksi ARIMA-mallien lineaarinen perusrakenne rajoittaa huomattavasti niiden kykyä mallintaa dynaamisia, ei-lineaarisia markkinasuhteita.

Ehdollisen heteroskedastisuuden huomioivat GARCH-mallit (Bollerslev, 1986) kehitettiin vastaamaan nimenomaan volatilitietin klusteroitumisen ongelmaan, ja ne ovat osoittautuneet erittäin hyödyllisiksi riskin ennustamisessa. Vaikka GARCH-arkkitehtuurit kykenevät mallintamaan tuottovarianssin ajallista dynamiikkaa tehokkaasti, niiden pääasiallinen käyttötarkoitus on riskinmallinnus, ei instrumentin hintatason tai sen kehityssuunnan ennustaminen. Tämän vuoksi niiden käyttö kvantitatiivisessa sijoitustoiminnassa painottuu ensisijaisesti salkun riskienhallintaan ja optioiden hinnoitteluun tuottohakuisten strategioiden sijaan.

Neuroverkot tarjoavat joustavan ja tehokkaan vaihtoehdon volatilitietin arviointiin. Hamid ja Iqbal (2004) osoittivat tutkimuksessaan, että neuroverkot kykenevät ennustamaan S&P 500 -futuuriin implisiittistä volatilitietettä perinteisiä tilastollisia malleja tarkemmin. Heidän tuloksensa vahvistavat, että neuroverkkojen kyky epälineaariseen mallinnukseen on poikkeuksellisen arvokas ominaisuus käsiteltäessä monimutkaisia rahoitusaikasarjoja, joille on tyypillistä volatilitietin epäsymmetrinen klusteroituminen.

Teknisten mallinnushaasteiden lisäksi osakekurssien ennustamista vaikeuttaa rahoitusmarkkinoiden adaptiivinen luonne. Mikäli markkinoilta havaitaan historiallinen ennustekuvio ja sitä aletaan hyödyntää laajasti kaupankäynnissä, arbitraasimekanismit ja kilpailupaine pyrkivät nopeasti hävittämään kyseisen anomalian. Tämän seurauksena koneoppimismallin suorituskky ei ole staattinen ominaisuus, vaan se elää jatkuvassa vuorovaikutuksessa markkinaympäristön kehityksen kanssa. Historiallisella aineistolla todennettu mallin aiempi toimivuus ei siten takaa tulevaa menestystä, mikä asettaa merkittäviä vaatimuksia mallien jatkuvalle päivittämiselle ja markkinadynamiikan ymmärtämiselle.

3.2 Neuroverkot ja ei-lineaarinen mallinnus

3.2.1 Monikerroksiset perceptron-verkot

Monikerroksiset perceptron-verkot (MLP) kuuluvat rahoitusalan varhaisimpiin ja tutkituimpiin koneoppimismalleihin. Niiden toiminta perustuu arkkitehtuuriin, jossa syötemuuttujat muunnetaan useiden piilokerrosten läpi ei-lineaaristen aktivaatiofunktioiden avulla. Tämä joustavuus mahdollistaa monimutkaisten funktionaalisten suhteiden approksimoinnin, mihin perinteiset mallit eivät kykene.

Maciel ja Ballini (2010) havaitsivat empiirisessä tutkimuksessaan, että MLP-verkot saavuttivat useiden kansainvälisten indeksien kohdalla noin 75 prosentin suuntatarkkuuden, kun taas ehdollisen varianssin GARCH-malli ylsi vain 55–60 prosenttiin. Tämä ei-lineaarisen mallinnuksen paljastama rakenteellinen etu ei jäänyt vain tilastolliseksi, vaan se johti simuloiduissa kaupankäyntistrategioissa huomattavasti korkeampiin kumulatiivisiin tuottoihin. Vastaavia käytännön hyötyjä havaitsivat Desai ym. (2011) tutkiessaan Intian osakemarkkinoita: neuroverkkolähestymistapa tuotti keskimäärin noin 29 prosentin vuotuisen tuoton passiivisen strategian jäädessä 18 prosenttiin. Ennustetarkkuuden parantuminen heijastui myös merkittävästi korkeampana riskikorjattuna tuottona eli Sharpen lukuna.

MLP-verkkojen vankka suorituskkyky on herättänyt laajaa vertailevaa tutkimusta. Guresen ym. (2011) tutkivat kolmea erilaista keinokekoisten neuroverkkojen (ANN) arkkitehtuuria NASDAQ-indeksin ennustamisessa ja osoittivat, että joissakin tilanteissa yksinkertainen perceptron-verkko ylitti monimutkaisempien dynaamisten arkkitehtuurien suorituskyyvyn. Havainto korostaa, ettei mallin monimutkaisuus itsessään takaa parempia tuloksia, vaan arkkitehtuurivalinta on aina sovitettava käsillä olevaan datarakenteeseen. Kuten Tkáč ja Verner (2016) kahden vuosikymmenen kirjallisuutta kattavassa katsauksessaan vahvistavat, neuroverkot ovat vakiinnuttaneet paikkansa erityisesti rahoitusalan luokittelu- ja ennustetehtävissä.

Yksi neuroverkkojen keskeisistä vahvuuksista on niiden kyky hyödyntää historiatietoa joustavasti. Moghaddam ym. (2016) tarkastelivat NASDAQ-indeksille tehtyjä ennusteita muuttamalla syöteikkunan pituutta. Kun havaintoikkunaa pidennettiin neljästä päivästä yhdeksään, mallin selitysaste (R^2) kasvoi 0,89:stä 0,96:een, ja samanaikaisesti keskineliövirheen neliöjuuri (RMSE) pieneni systemaattisesti. Tulokset todistavat mallien kyvyn oppia ja hyödyntää dynaamisesti muuttuvaa aikasarjadataa.

3.2.2 Rekurrentit verkot ja LSTM-arkkitehtuurit

Rahoitusaikasarjojen peräkkäinen eli sekventiaalinen luonne tekee takaisinkytketyistä neuroverkoista (RNN) teoreettisesti houkuttelevan vaihtoehdon. Toisin kuin eteenpäin kytketyissä monikerroksisissa perceptroneissa (MLP), RNN-verkoissa on sisäinen muistimekanismi, jonka avulla aikasarjan aiemmat havainnot vaikuttavat suoraan tuleviin ennusteisiin.

Wu ym. (2017) totesivat tutkimuksessaan, että RNN-malli saavutti CSI 300 -indeksin ennustamisessa pienemmän keskineliövirheen neliöjuuren (RMSE) ja keskimääräisen suhteellisen itseisarvovirheen (MAPE) verrattuna perinteiseen MLP-malliin. Tämä havainto korostaa RNN-verkkojen rakenteellista etua aikariippuvuuksien mallintamisessa.

Erityisen vakuuttavaa on näyttö, joka perustuu pitkiin otoksen ulkopuolisiin (out-of-sample) arviointijaksoihin. Fischer ja Krauss (2018) testasivat LSTM-mallia S&P 500 -indeksin osakkeilla pitkällä aikavälillä ja havaitsivat sen tuottavan tilastollisesti merkitsevää ylituottoa (alfaa) silloinkin, kun transaktiokustannukset huomioitiin. Tämä viittaa siihen, että LSTM-verkkojen ennustekyky on siirrettävissä käytännön kaupankäyntistrategioihin, joskin markkinadynamiikan muutokset edellyttävät mallin säännöllistä päivittämistä. Lisäksi Khodayari ym. (2020) osoittivat, että neuroverkot kykenevät mallintamaan markkinavolatiliteetin ja sijoittajien päätöksenteon välistä epälineaarista dynamiikkaa tavalla, jota lineaariset mallit systemaattisesti aliarvioivat.

Kuten luvussa 2.2.2 käsiteltiin, LSTM-arkkitehtuuri kehitettiin ratkaisemaan katoavan gradientin ongelma (vanishing gradient problem), joka heikentää huomattavasti perinteisten RNN-verkkojen oppimiskykyä (Hochreiter & Schmidhuber, 1997). Bao ym. (2017) veivät tätä kehitystä eteenpäin yhdistämällä aallokemuunnoksen (wavelet transform), pinotut autoenkooderit ja LSTM-ennustimen edistyneeksi hybridimalliksi. Heidän lähestymistapansa saavutti alle 1,2 prosentin MAPE-arvoja useille tutkituille indekseille. Mallin mittaama Theilin U -epävarmuuskerroin oli systemaattisesti perinteisiä ARIMA- ja MLP-malleja pienempi, mikä heijastui selvästi parempana suorituskynä kaupankäyntisimulaatioissa.

Tällaiset hybridimallit edustavat nykytutkimuksen kärkeä: ne yhdistävät datan esikäsittelyn, kuten kohinan suodatuksen, piirteiden erottamisen ja sekventiaalisen ajallisen mallinnuksen yhdeksi yhtenäiseksi prosessiksi. Lähestymistapa huomioi rahoitusdatan monikerroksisen luonteen ja monimutkaisuuden tehokkaammin kuin erillisten arkkitehtuurien käyttö yksinään.

3.3 Ensemble-menetelmät ja korkeaulotteinen osakedata

Puupohjaiset menetelmät, kuten satunnaismetsät (Breiman, 2001) ja gradient boosting (Friedman, 2001), perustuvat useiden heikkojen oppijoiden (weak learners) yhdistämiseen. Näiden menetelmien suurin etu on niiden kyky käsitellä suuria ja monimutkaisia piirreavaruuksia tekemättä rajoittavia parametrisia ennakkoolotuksia datan rakenteesta.

Tämän osoittivat vakuuttavasti Gu, Kelly ja Xiu (2020), jotka analysoivat laajaa Yhdysvaltain osakeaineistoa käyttäen 94 yrityskohtaista ennustemuuttujaa sekä lukuisia makrotaloudellisia indikaattoreita. Heidän tuloksensa osoittivat, että boosting- ja neuroverkkopohjaiset mallit saavuttivat huomattavasti pienemmän otoksen ulkopuolisen ennustevirheen ja korkeamman selitystason (R^2) verrattuna perinteiseen lineaariseen regressioon. Käytännön tasolla tämä parannus ennustetarkkuudessa heijastui suoraan kaupankäyntistrategioihin: koneoppimiseen perustuvien pitkä-lyhyt-

strategioiden (long-short) annualisoidut alfat ja Sharpe-luvut ylittivät selvästi perinteisten menetelmien tuotot.

Tämä havainto on merkittävä erityisesti aineiston laajuuden ja pitkän arviointijakson vuoksi. Se osoittaa konkreettisesti, että koneoppimisen tuoma lisäarvo korostuu nimenomaan korkeaulotteisissa ympäristöissä, joissa perinteiset lineaariset mallit ovat kykenemättömiä hahmottamaan muuttujien välisiä monimutkaisia ja epälineaarisia vuorovaikutuksia.

3.4 Vertailu klassisiin malleihin

Vaikka joissakin tapauksissa empiirinen näyttö koneoppimisen puolesta on vakuuttavaa, tutkimuskirjallisuus ei anna perusteita yksiselitteiselle johtopäätökselle sen absoluuttisesta yliveraisuudesta perinteisiin malleihin nähden. Makridakis ym. (2018) osoittivat, että klassiset tilastolliset menetelmät säilyttävät vahvan kilpailukykyä lukuisten erilaisten aikasarjojen kohdalla. Vastaavasti Džikevičius ja Stabužytė (2012) havaitsivat osakeindeksien mallinnuksessa, että perinteiset regressiomallit saavuttivat kuukausitason ennusteissa jopa neuroverkkoja parempia tuloksia.

Se, että eri mallien suhteellinen tehokkuus ei ole vakio, vaan riippuu voimakkaasti muun muassa ennustehorisontista, datan rakenteesta ja tutkimusasetelmasta, korostaa menetelmien erilaista luonnetta. Koneoppiminen tarjoaa poikkeuksellista joustavuutta ja kykyä ei-lineaaristen suhteiden mallintamiseen, kun taas klassisten mallien vahvuutena säilyvät niiden teoreettinen läpinäkyvyys, rakenteellinen vakaus ja tehokkuus pienemmillä aineistoilla.

3.5 Tutkimustulosten ja menetelmien arviointi

Koneoppimismenetelmien suorituskyvyn luotettava arviointi rahoitusmarkkinoilla edellyttää huolellista tutkimusasetelmaa. Koska syväoppimisjärjestelmissä parametrien määrä on usein erittäin suuri suhteessa käytettävissä olevaan otoskokoon,

ylisovittamisen (overfitting) riski on huomattava. Perinteinen satunnainen ristiinvalidointi soveltuu rahoitusaikasarjoihin huonosti ennakkotiedon vinouman (look-ahead bias) vuoksi. Tämän vuoksi rullaavaan aikaikkunaan (rolling window) tai ns. walk-forward-menetelmään perustuva validointi on perustellumpaa, sillä se simuloi todellisessa sijoitustoiminnassa kohdattavaa informaatiiovirtaa todenmukaisemmin (Aldridge & Avellaneda, 2021).

Toinen keskeinen haaste liittyy kaupankäyntikustannuksiin, jotka on ehdottomasti huomioitava, jotta akateemisten mallien tulokset olisivat sovellettavissa käytännön salkunhoitoon. Vaikka useissa tutkimuksissa koneoppimisen avulla on saavutettu tilastollisesti merkitsevää ylituottoa (alfa), mallien vaatimasta tiheästä kaupankäynnistä koituvat transaktiokustannukset voivat helposti hävittää strategian nettotuoton (Bartram ym., 2021; Fischer & Krauss, 2018).

Kolmas merkittävä este erityisesti syväoppimismallien laajemmalle institutionaaliselle hyödyntämiselle on niiden heikko tulkittavuus. Siinä missä perinteisten lineaaristen regressiomallien kertoimille voidaan suoraan antaa taloudellinen merkitys, neuroverkkojen päätöksenteko perustuu monimutkaiseen painoarvojen jakautumiseen piilokerroksissa muodostaen niin sanotun mustan laatikon (black box). Tämä läpinäkymättömyys vaikeuttaa sijoituspäätösten perustelemista sekä sääntelyviranomaisille että asiakkaille (Bartram ym., 2021).

3.6 Tulosten analyysi

Tässä luvussa tarkastellun kirjallisuuden perusteella koneoppimisen havaitaan saavuttavan osaketuottojen ennustamisessa paremman tarkkuuden kuin perinteisten lineaaristen mallien, erityisesti lyhyillä ennustehorisonteilla ja korkeaulotteisen datan ympäristöissä. Esimerkiksi Gu ym. (2020) osoittivat laajaa Yhdysvaltain osakedataa hyödyntävässä tutkimuksessaan, että sekä boosting-menetelmät että neuroverkot kykenivät pienentämään otoksen ulkopuolista keskineliövirhettä (MSE) huomattavasti verrattuna lineaariseen regressioon. Havainto on poikkeuksellisen merkittävä, sillä se

todistaa koneoppimismallien tarjoaman ennustuskyvyn parantumisen olevan paitsi tilastollisesti merkitsevää, myös käytännön sijoitustoiminnassa taloudellisesti hyödynnettävää etua

Arkkitehtuurianalyysi osoittaa, että eri malliperheet soveltuvat erilaisiin ennustamistehtäviin. LSTM-pohjaisilla malleilla on erityinen kyky hahmottaa peräkkäisten aikasarjojen riippuvuuksia: Wu ja Tingting (2017) totesivat RNN-mallien saavuttavan matalamman RMSE:n kuin MLP-verkkojen, ja Fischer ja Krauss (2018) osoittivat LSTM-verkkojen tuottavan huomattavaa alfaa osakevalintastrategioissa myös transaktiokustannusten huomioimisen jälkeen. Bao ym. (2017) kehittämän aalloke-LSTM-arkkitehtuurin kaltaiset hybridimallit ylittävät erillismallit yhdistämällä kohinan suodatuksen, piirteiden erottamisen ja ajallisen mallinnuksen yhdeksi yhtenäiseksi prosessiputkeksi. Myös ensemble-menetelmät, erityisesti satunnaismetsät (Breiman, 2001) ja gradient boosting (Friedman, 2001), vähentävät tehokkaasti varianssia pienentäen siten ylisovittumisen riskiä, joka on tyypillistä yksittäisille malleille.

On tärkeää tehdä tietty varaus verrattaessa tuloksia perinteisiin malleihin: olemassa oleva kirjallisuus ei todista koneoppimisella olevan ehdotonta etulyöntiasemaa. Makridakis ym. (2018) osoittivat, että klassiset tilastolliset menetelmät pysyvät kilpailukykyisinä etenkin silloin, kun otoskoko on pieni tai ennustehorisontti pitkä. Džikevičius ja Stabužytė (2012) havaitsivat regressiomallien toimivan paremmin kuukausitason horisonteissa. Mallin suorituskyky riippuu siten suuresti kontekstista: koneoppiminen tarjoaa erityistä lisäarvoa tapauksissa, joihin liittyy suuri ulottuvuus, lyhyen aikavälin ennusteita ja ei-lineaarisia suhteita.

Kirjallisuudessa on kuitenkin selkeitä rajoitteita, jotka on syytä tunnistaa. Empiirinen näyttö tapaa rajautua tiettyihin markkinoihin, tiettyihin otosajanjaksoihin ja rajattuihin data-asetelmiin, mikä vähentää tulosten yleispätevyyttä. Ylisovittaminen on edelleen jatkuva ongelma, sillä mallit, jotka toimivat erinomaisesti historiallisella testidatalla, voivat pettää muuttuvissa markkinaolosuhteissa (Hastie ym., 2009). Empiirisissä

tutkimuksissa ei useinkaan huomioida transaktiokustannuksia riittävästi, mikä saattaa johtaa siihen, että raportoidut alfat liioittelevat todellisia tuottoja (Bartram ym., 2021). Lisäksi läpinäkyvyyden puute estää syväoppimisen käyttöä institutionaalisissa ympäristöissä, joissa sijoituspäätösten on oltava perusteltuja.

Kirjallisuuskatsauksen perusteella voidaan todeta: vaikka koneoppiminen on hyödyllinen työkalu osakekurssien ennustamiseen, se ei ole yleispätevä ratkaisu. Sen hyödyntäminen vaatii huolellista metodologista toteutusta, markkinakontekstin hyvää ymmärtämistä sekä varovaisuutta ennusteiden muuntamisessa varsinaisiksi kaupankäyntistrategioiksi. Seuraavassa luvussa tarkastellaan, miten generatiivinen tekoäly täydentää koneoppimisen mahdollisuuksia sijoittajan näkökulmasta.

4 Generatiivinen tekoäly sijoittajan työkaluna

Edellä käsitellyillä koneoppimismenetelmillä on erityiset vahvuutensa strukturoidun numeerisen datan käsittelyssä. Rahoitusmarkkinoiden informaatioympäristö on kuitenkin paljon tätä laajempi: tilinpäätökset, vuosikertomukset, tulospuhelut, analytikkoraportit, uutiset ja sosiaalinen media tuottavat valtavan määrän laadullista tietoa, jonka johdonmukainen käsittely on perinteisesti vaatinut asiantuntijaa. Generatiivinen tekoäly, erityisesti suuret kielimallit (Large Language Models, LLM), on muuttanut pääsyä tähän laadulliseen tietoon rahoitusmarkkinoilla.

4.1 Transformer-arkkitehtuurista suuriin kielimalleihin

Rahoitustekstin automaattinen käsittely ei ole uusi tutkimusaihe. Loughran ja McDonald (2011) kehittivät rahoitusosalalle räätälöidyn sentimenttisanakirjan, joka osoitti yleiskielen sanalistoihin perustuvan negatiivisuuden mittaamisen yliarvioivan systemaattisesti yritysdokumenttien negatiivisuutta. Tämä havainto käynnisti laajemman tutkimusperinteen, jossa tekstimuotoisen yritysinformaation ja osakemarkkinoiden välisiä yhteyksiä pyrittiin systemaattisesti kartoittamaan.

Varhaiset luonnollisen kielen prosessoinnin (NLP) menetelmät käsitelivät tekstiä pääasiassa sanojen esiintymistiheyksien kautta, mikä irrotti sanat niiden alkuperäisestä asiayhteydestä. Tämä rajoitti menetelmien kykyä ymmärtää monitulkintaisia tai asiayhteydestä riippuvaisia ilmauksia, jotka ovat rahoitusalan viestinnässä keskeisiä. Kuten Goodfellow ym. (2016) kuvaavat, sanaopetusmenetelmät (word embeddings) toivat tähän merkittävän parannuksen mallintamalla sanojen välisiä semanttisia suhteita tiivistetyssä vektorimuodossa. Näiden staattisten upotusten rajoitteena oli kuitenkin se, että ne antoivat sanalle aina saman vektorin riippumatta sen kontekstista, jolloin esimerkiksi osakekursseihin liittyvien termien moniselitteisyys saattoi hämärtyä.

Keskeinen teknologinen harppaus tapahtui, kun Vaswani ym. (2017) esittelivät tarkkaavaisuusmekanismiin (attention mechanism) perustuvan Transformer-

arkkitehtuurin. Toisin kuin aiemmat sekventiaaliset mallit, Transformer kykenee prosessoimaan koko syötteen samanaikaisesti, mikä mahdollisti sekä pitkien tekstiriippuvuuksien paremman mallintamisen että erittäin tehokkaan rinnakkaislaskennan.

Tämän arkkitehtuurin pohjalta rakennettiin modernit suuret kielimallit. Brown ym. (2020) esittelivät 175 miljardin parametrin GPT-3-mallin ja osoittivat sen kykenevän suoriutumaan laajasta kirjosta täysin uusia kielitehtäviä pelkästään muutaman esimerkin ohjaamana (few-shot learning) ilman erillistä hienosäätöä. Rahoituskontekstissa tämä yleinen kielikyky tarkoittaa, että yksi ja sama malli voi joustavasti siirtyä tilinpäätöstekstin tiivistämisestä sijoittajan kysymyksiin vastaamiseen tai riskisignaalien tunnistamiseen.

4.2 Suurten kielimallien sovellukset rahoitusmarkkinoilla

Sentimenttianalyysi on parhaiten dokumentoitu LLM:ien sovelluskohde rahoitusmarkkinoilla. López-Lira ja Tang (2023) tutkivat järjestelmällisesti, kykeneekö ChatGPT:n tuottama sentimenttiluokitus talousuutisille ennustamaan osakekurssien päiväsäisiä tuottoja. Tulokset olivat tilastollisesti merkittäviä: positiivinen sentimenttisignaali korreloi positiivisten osaketuottojen kanssa ja negatiivinen signaali ennakoii negatiivista kehitystä erityisesti selkeän sentimentin uutisten kohdalla. LLM-pohjainen sentimenttiluokitus suoriutui paremmin kuin Loughranin ja McDonaldin (2011) sanakirjapohjainen lähestymistapa, mikä viittaa siihen, että LLM:t pystyvät tavoittamaan semanttisia nuansseja, joita sanapohjainen mittaus ei tunnista.

Ke ym. (2019) osoittivat jo ennen LLM-aikakautta, että yritysten vuosikertomusten tekstisisältö, erityisesti johdon analyysi- ja keskusteluosio (MD&A), sisältää ennakointivoimaa tulevista osaketuotoista, ja että tätä informaatiota voidaan hyödyntää koneoppimismenetelmin. LLM:t tarjoavat luontevan seuraavan askeleen: ne kykenevät prosessoimaan ja tiivistämään pitkien dokumenttien sisällön nopeammin ja laajamittaisemmin kuin aiemmat NLP-menetelmät, mahdollistaen laaja-alaisen tekstianalyysin myös suurten organisaatioiden käyttöön.

Institutionaaliset sijoittajat käsittelevät runsaasti aikaa vieviä dokumentteja, kuten tilinpäätöksiä, osavuositarkastuksia ja analyttikkoraportteja, joiden perusteellinen läpikäyminen on työvoimaintensiivistä. LLM-mallit kykenevät tiivistämään pitkien dokumenttien ydinsisällön, vertailemaan peräkkäisten tilikausien kieliasua muutosten tunnistamiseksi sekä havaitsemaan johdon puhetavoissa tapahtuvia muutoksia, jotka saattavat ennakoita liiketoiminnan suunnanmuutoksia (Aldridge & Avellaneda, 2021). Tämä muodostaa sijoittajan näkökulmasta konkreettisen tehokkuusedun laadullisen analyysin nopeuttamisessa.

Erityisen kiinnostava sovellus on yritysten tulospuheluiden (earnings calls) analyysi. Johdon äänensävy, tiettyjä aiheita koskevat välttämistästrategiat ja suhtautuminen analyttikkojen kysymyksiin voivat sisältää informaatiota, joka ei välity suoraan tilinpäätösluvuista. Puhelulitteroinnin perusteella LLM voi järjestelmällisesti vertailla johdon viestintää aiempiin kausiin, kilpailijoihin tai toimialan yleiseen puhetapaan, tuottaen signaaleja, joiden tunnistaminen käsin olisi käytännössä mahdotonta laajassa mittakaavassa.

Ehkä eniten julkista huomiota on saanut LLM:ien käyttö sijoitusneuvontaan ja päätöstukeen. Chatbot-tyyppiset rajapinnat mahdollistavat yksityissijoittajalle sijoitusterminologian selittämisen, riskiprofiilien hahmottamisen ja erilaisten sijoitusvaihtoehtojen vertailun luonnollisella kielellä (Aldridge & Avellaneda, 2021). Tämä voi potentiaalisesti pienentää tiedollista epäsymmetriaa kokeneiden institutionaalisten sijoittajien ja yksittäisten piensijoittajien välillä, tasoittaen pääsyä laadukkaaseen taloudelliseen analytiikkaan.

4.3 Generatiivinen tekoäly suhteessa perinteiseen koneoppimiseen

Perinteinen koneoppiminen ja generatiivinen tekoäly eroavat toisistaan sekä teknisesti, datavaatimuksiltaan että käyttötarkoituksiltaan. Edellä käsitellyt koneoppimismallit (LSTM, random forest ja gradient boosting) on tyypillisesti koulutettu spesifiin

ennustetehtävään strukturoidulla numeerisella datalla. LLM:t puolestaan ovat geneerisiä kielimalleja, joita ei ole optimoitu mihinkään yksittäiseen tehtävään, mutta jotka kykenevät joustavan luonnollisen kielen käsittelyn kautta mukautumaan eri käyttötarkoituksiin (Brown ym., 2020).

Käytännön sijoitustoiminnassa nämä teknologiat täydentävät toisiaan. Siinä missä Gu ym. (2020) osoittivat koneoppimismallien kykenevän käsittelemään poikkeuksellisen laajaa ja monimutkaista numeerista muuttuja-avaruutta, kielimallit puolestaan avaavat pääsyn laadulliseen dataan. Esimerkiksi López-Liran ja Tangin (2023) osoittamia LLM-sentimenttisiгнаaleja voidaan käyttää lisämuuttujina koneoppimismalleissa, jolloin kvantitatiivinen malli ja tekstuaalinen analyysi yhdistyvät samassa ennusteprosessissa.

Tulkittavuudessa erot ovat kuitenkin huomattavia. Koneoppimismalleissa, erityisesti puupohjaisissa menetelmissä, on kehitetty post-hoc -tulkintamenetelmiä piirteiden tärkeyden arviointiin, mikä osittain helpottaa läpinäkyvyyshaastetta (Hastie ym., 2009). LLM:ien kohdalla yksittäisen tulosteen jäljittäminen tiettyyn päättelyketjuun on huomattavasti haastavampaa: miljardeja parametreja käsittävän mallin toimintaperiaate jää käytännössä läpinäkymättömäksi (Goodfellow ym., 2016).

4.4 Generatiivisen tekoälyn rajoitteet ja riskit sijoituskontekstissa

Suurten kielimallien keskeinen haaste on niiden taipumus tuottaa sisällöllisesti virheellistä mutta vakuuttavalta vaikuttavaa tietoa, niin sanottuja hallusinaatioita. Koska autoregressiiviset kielimallit on koulutettu ennustamaan tilastollisesti todennäköisintä seuraavaa tekstiä eikä hakemaan varmennettuja tietueita tietokannasta (Brown ym., 2020), ne voivat generoida olemattomia faktoja tai virheellisiä lähdeviittauksia. Sijoitustoiminnassa tämä muodostaa huomattavan riskin: malli voi esittää virheellisiä osaketuottoja tai keksiä analyytikkosuosituksia ilman, että käyttäjä havaitsee virhettä tekstin sujuvuudesta.

Suuret kielimallit esikoulutetaan tyypillisesti tietyllä tekstiaineistolla vain tiettyyn ajankohtaan saakka, minkä vuoksi niiden staattinen tietopohja ei oletusarvoisesti kykene prosessoimaan reaaliaikaista tai uutta informaatiota ilman erillistä tiedonhakuominaisuutta (Wang ym., 2025). Rahoitusmarkkinoilla, joilla relevantti tieto päivittyy jatkuvasti, tämä muodostaa rakenteellisen haasteen: malli ei kykene kommentoimaan esimerkiksi viimeisintä tulosjulkistusta tai muuttuneita kurssihintoja ilman ulkopuolista, dynaamisesti päivitettävää tiedonlähdettä. Tietopohjan ajallinen rajallisuus onkin yksi keskeisimmistä suurten kielimallien soveltamisen esteistä reaaliaikaisessa sijoituspäätöksenteossa, mitä on alettu ratkaista yhdistämällä malleihin RAG (Retrieval-Augmented Generation) -tiedonhakuprosesseja (Wang ym., 2025).

Kuten luvussa 2.3.4 todettiin, syväoppimismallit kärsivät usein läpinäkyvyysongelmasta (black box). Suuret kielimallit vievät tämän haasteen uudelle tasolle: miljardeista parametreista muodostuva malli tuottaa vastauksensa monimutkaisessa prosessissa, jota on erittäin vaikea jäljittää yksittäisiin ja tulkittaviin päättelyketjuihin (Al-Otaibi ym., 2025). Rahoitussäätelyssä tämä muodostaa vakavan ongelman, sillä algoritmien tuottamien sijoitussuositusten tai luottopäätösten on oltava selitettävissä ja perusteltavissa sääntelyvaatimusten mukaisesti, eikä sääntely hyväksy päätöksentekoa ilman läpinäkyvää auditointipolkua (Al-Otaibi ym., 2025). Täysin läpinäkymätön ja selittämätön (unexplainable) kielimallin tuottama sijoitus- tai riskipäätös ei siten sovellu sellaisenaan institutionaalisen toimijan päätöksentekoon.

LLM-järjestelmät ovat osoittautuneet haavoittuvaisiksi niin sanotuille prompt injection -hyökkäyksille (syöteinjektioille), joissa haitallinen tekstisyöte ohjaa mallin käyttäytymistä ei-toivottuun suuntaan. Tilanteessa, jossa LLM prosessoi automaattisesti ulkopuolisia dokumentteja, kuten uutisvirtoja, verkkosivuja tai raporttilatauksia, järjestelmä altistuu epäsuorille injektiohyökkäyksille (Greshake ym., 2023). Rahoitusympäristöissä tämä muodostaa vakavan tietoturvariskin: koska markkinaosapuolilla voi olla merkittävä taloudellinen kannustin harhauttaa kilpailijoiden algoritmisia kaupankäyntijärjestelmiä manipuloimalla niille syötettävää dataa (vrt. Aldridge & Avellaneda, 2021), LLM-

pohjaisten päätöksentekojärjestelmien turvallisuuteen on kiinnitettävä erityistä huomiota.

4.5 Yhteenveto

Generatiivinen tekoäly tarjoaa sijoittajalle merkittäviä uusia työkaluja rahoitusmarkkinoiden laadullisen informaation käsittelyyn. Suuret kielimallit kykenevät tiivistämään tilinpäätöksiä, luokittelemaan uutisten sentimenttiä ja syntetisoimaan sijoitustietoa tavalla, joka täydentää perinteisten koneoppimismallien kvantitatiivisia ennusteita. Empiirinen tutkimusnäyttö, kuten López-Liran ja Tangin (2023) havainnot ChatGPT:n sentimenttisihtinaalin ennustevoimasta, osoittaa, että laadullisen analyysin automatisointi voi tuottaa sijoitusprosessiin mitattavaa lisäarvoa.

Samalla generatiivinen tekoäly tuo mukanaan riskejä, jotka eroavat luonteeltaan perinteisen koneoppimisen haasteista. Hallusinaatiot, staattinen tietopohja, päättelyketjujen läpinäkymättömyys sekä manipuloitavuus edellyttävät tiukkaa riskienhallintaa ennen kuin LLM-pohjaisia järjestelmiä voidaan hyödyntää vastuullisesti osana sijoitusprosessia. Teknologioiden rajoitteiden ymmärtäminen ei ole este niiden käytölle, vaan perusedellytys niiden menestyksellään soveltamiselle. Seuraavassa luvussa tarkastellaan näiden havaintojen laajempaa merkitystä rahoitusalan erisidosryhmien näkökulmasta.

5 Johtopäätökset

Tässä kandidaatintutkielmassa on tarkasteltu koneoppimisen ja generatiivisen tekoälyn sovelluskohteita osana sijoitusprosessia. Kirjallisuuskatsauksen perusteella muodostuu selkeä kuva: molemmat teknologiat tarjoavat sijoittajalle mitattavaa lisäarvoa, mutta niiden vahvuudet painottuvat täysin eri analyysin tasoille ja työvaiheisiin.

Kvantitatiivisessa osakekurssien ennustamisessa koneoppiminen parantaa ennustetarkkuutta, joskin saavutetut tulokset ovat vahvasti kontekstisidonnaisia. Erityisesti LSTM-pohjainen syväoppiminen, ensemble-menetelmät, kuten satunnaismetsät (Breiman, 2001) ja gradient boosting (Friedman, 2001), sekä erilaiset hybridimallit ylittävät johdonmukaisesti perinteisten lineaaristen mallien suorituskyvyn lyhyen aikavälin rahoitusaikasarjoissa (Fischer & Krauss, 2018; Gu ym., 2020). Koneoppimisen tuottama ennustekyky ei kuitenkaan ole universaali, vaan se riippuu ratkaisevasti syötedatan rakenteesta, ennustehorisontista ja vallitsevasta markkinasyklistä. Lisäksi mallien käytännön soveltamista rajoittavat merkittävästi ylisovittumisen (overfitting) riski sekä päätöksenteon läpinäkyvättömyys.

Generatiivinen tekoäly täydentää tätä kvantitatiivista perustaa tarjoamalla systemaattisen väylän laadulliseen informaatioon, joka on perinteisesti jäänyt numeeristen mallien ulottumattomiin. López-Liran ja Tangin (2023) tulokset ChatGPT:n tuottamien sentimenttisihtaalien ennustevoimasta sekä Ke ym. (2019) aiemmat havainnot tekstidatan arvosta osoittavat, että kielimallit tuottavat mitattavaa lisäarvoa juuri siellä, missä perinteinen koneoppiminen kohtaa rajansa: massiivisen tekstiaineiston semanttisessa analyysissa. Nämä kaksi teknologiaa eivät siten kilpaile keskenään, vaan muodostavat luontevan, toisiaan täydentävän analyysikonaisuuden.

Kummankin teknologian soveltamiseen liittyy kuitenkin kriittisiä rajoitteita, jotka sijoittajan on tunnistettava. Koneoppimismallien kompastuskivinä ovat erityisesti rahoitusdatan epästationaarisuus, mallien ylisovittuminen historialliselle aineistolle sekä tulosten heikko selitettävyyys. Generatiivisen tekoälyn kohdalla haasteet ovat vielä

monimutkaisempia: arvaamattomat hallusinaatiot, tietopohjan staattisuus ja järjestelmien lähes täydellinen läpinäkymättömyys estävät suoraviivaisen luottamuksen. Vaikka teknologiat tehostavat tiedonkäsittelyä ennennäkemättömällä tavalla, ne eivät korvaa rahoitusalan syvällistä substanssiosaamista, vaan pikemminkin ne korostavat inhimillisen asiantuntijan ja kriittisen arviointikyvyn merkitystä modernissa sijoitusprosessissa.

5.1 Löydösten merkitys eri sidosryhmille

Yksityissijoittajan näkökulmasta koneoppimisen ja generatiivisen tekoälyn yhdistelmä tarjoaa merkittävän mahdollisuuden tiedollisen epäsymmetrian kaventamiseen. Kun LLM-pohjaiset sovellukset mahdollistavat laajan markkinadatan ja monimutkaisten tilinpäätösten nopean syntetisoinnin, piensijoittaja saa käyttöönsä analyyttisiä valmiuksia, jotka ovat perinteisesti vaatineet institutionaalisia resursseja. Teknologian hyödyntäminen vaatii kuitenkin uudenlaista lähdekriittisyyttä. Hallusinaatioiden ja faktavirheiden riskin vuoksi sijoittajan on osattava arvioida tekoälyn tuotoksia asiantuntevasti, eikä sokea luottamus automaatioon ole perusteltua.

Finanssialan instituutioille tutkimuskirjallisuus piirtää selkeän teknologisen kehityskaaren. Kvantitatiiviset koneoppimismallit ovat jo vakiinnuttaneet asemansa muun muassa riskienhallinnassa, luottoluokitusten arvioinnissa ja petostentorjunnassa (Bartram ym., 2021; Lee, 2007). Generatiivinen tekoäly laajentaa tätä työkalupakkia laadullisen prosessoinnin puolelle, mahdollistaen esimerkiksi sääntelydokumenttien (compliance) automatisoidun läpikäynnin, asiakasrajapinnan tehostamisen ja sisäisten tietovarantojen nopeamman hyödyntämisen. Tulevaisuudessa kilpailuetu markkinoilla rakentuukin todennäköisesti niiden organisaatioiden varaan, jotka kykenevät integroimaan nämä teknologiat saumattomasti osaksi tiukasti säänneltyä toimintaympäristöä.

Sijoituspalvelualan ammattilaisten (analyytikkojen, salkunhoitajien ja riskimanagerien) kohdalla tutkielma tukee näkemystä, jonka mukaan koneoppiminen ja generatiivinen

tekoäly toimivat parhaiten asiantuntijan työtä täydentävinä, ei korvaavina teknologioina. LLM kykenee käsittelemään satoja analytiikkoraportteja nopeammin kuin yksittäinen analytiikko, mutta toimialaosaaminen, eettinen arviointi ja asiakasvuorovaikutus pysyvät inhimillisen osaamisen ydinalueina. Ammattilaisille tärkeintä on ymmärtää teknologioiden vahvuudet ja rajoitteet riittävän syvästi osataksaan arvioida niiden tuotoksia kriittisesti.

Valmistuville ekonomisteille ja alan opiskelijoille tämä tutkielma osoittaa selväksi: koneoppimisen ja generatiivisen tekoälyn perusteiden ymmärtäminen on muodostumassa keskeiseksi osaksi rahoitusalan ammatillista osaamista. Ei ole odotettavissa, että kaikkien ekonomistien tulisi olla teknisiä asiantuntijoita, mutta kyky tunnistaa, milloin koneoppimismallit ovat asianmukaisia, arvioida LLM:n tuottaman analyysin luotettavuutta ja viestiä asiantuntijoiden kanssa näistä teknologioista on kilpailuetu tulevilla työmarkkinoilla (Gogas & Papadimitriou, 2021).

Sijoittajan näkökulmasta koneoppimisen ja generatiivisen tekoälyn tuomat mahdollisuudet ja uhat kulkevat rinnakkain. Mahdollisuudet liittyvät ennen kaikkea tiedonkäsittelykapasiteettiin: kykyyn analysoida laajoja tekstiaineistoja, tunnistaa markkinasignaaleja suuresta datamäärästä ja syntetisoida hajanaista informaatiota päätöksenteon tueksi nopeammin kuin aiemmin oli mahdollista. Uhat ovat yhtä konkreettisia: koneoppimismallit voivat tuottaa harhaisesti kalibroituja ennusteita, jos ne on sovitettu liiaksi historialliseen dataan, eivätkä ne sellaisenaan kykene sopeutumaan markkinaympäristön rakenteelliseen murrokseen. Generatiivinen tekoäly puolestaan voi tuottaa faktavirheitä, jotka vaikuttavat uskottavilta ja joiden havaitseminen edellyttää käyttäjältä substanssiosaamista. Molempien teknologioiden turvallinen hyödyntäminen vaatii sijoittajalta kriittistä arviointikykyä.

Laajempaa merkitystä arvioitaessa on tärkeää tarkastella myös sitä, miten teknologioiden laajamittainen käyttöönotto vaikuttaa markkinoiden tehokkuuteen. Tehokkaiden markkinoiden hypoteesin (Fama, 1970) valossa hinnat heijastavat jo kaiken

julkisen informaation. Kun yhä useampi toimija omaksuu samanlaiset koneoppimismallit ja sentimenttisisignaalit, markkinoiden informaationkäsittelykyky kasvaa, ja systemaattiset anomaliat, joita mallit hyödyntävät, alkavat haihtua. Tätä ilmiötä kutsutaan alfan kulumiseksi (alpha decay): mitä laajempaan käyttöön nämä teknologiat leviävät, sitä nopeammin niiden tuottama ylisuorituskyky tasoittuu. Teknologioista saatava hyöty on siis luonteeltaan väliaikainen, ja se edellyttää jatkuvaa innovointia sekä kykyä löytää uusia, vielä hyödyntämättömiä signaaleja. Sijoittajan on tiedostettava, että teknologinen kilpailu markkinoilla on jatkuva prosessi, ei kertaluonteinen ratkaisu.

5.2 Jatkotutkimusmahdollisuudet ja tutkimuksen rajoitteet

Tämä kirjallisuuskatsaus on rajautunut osakekurssien ennustamiseen ja LLM:ien sovelluksiin rahoitusmarkkinoilla. Muutamia jatkotutkimussuuntia on silti nostettavissa esiin. Erityisesti koneoppimismallien ja LLM:ien yhdistäminen yhdeksi integroiduksi päätöstukijärjestelmäksi on empiirisesti melko vähän tutkittu alue: miten tekstuaaliset signaalit tulisi parhaiten muuntaa numeerisiksi syötteiksi koneoppimisalleille ja missä määrin tämä yhdistäminen parantaa ennustekykyä yksinomaisiin malleihin verrattuna? Myös suomalaisten tai pohjoismaisten markkinoiden kontekstissa tehtävä empiirinen tutkimus tarjoaisi paikallisesti relevanttia näyttöä, jota kansainvälisesti painottuneessa kirjallisuudessa on vain vähän.

Kirjallisuuskatsausmenetelmän rajoitteet on myös tunnistettava. Koneoppimisen ja tekoälyn tutkimuskenttä kehittyy poikkeuksellisen nopeasti: tässä tutkielmassa käsitelty kirjallisuus edustaa ajankohtaa, jolloin LLM:ien soveltaminen rahoitusmarkkinoilla on vasta tutkimuksen alkuvaiheessa. Empiirinen näyttö on toistaiseksi osin hajanaista. Myös arviointikäytäntöjen standardisointi on vasta muotoutumassa. Nämä rajoitteet eivät heikennä tutkielman kirjallisuuteen perustuvien johtopäätösten pätevyyttä, mutta korostavat jatkuvan seurannan ja päivittämisen tarvetta.

Lähteet

Aldridge, I., & Avellaneda, M. (2021). *Big data science in finance*. John Wiley & Sons.

Al-Otaibi, R., Al-Zahrani, A., & Al-Qurishi, M. (2025). Exploring explainability in large language models. *Preprints.org*. <https://doi.org/10.20944/preprints202503.2318.v1>

Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*, 12(7), e0180944. <https://doi.org/10.1371/journal.pone.0180944>

Bartram, S. M., Branke, J., De Rossi, G., & Motahari, M. (2021). Machine learning for active portfolio management. *Journal of Financial Data Science*, 3(3), 9–30. <https://doi.org/10.3905/jfds.2021.1.071>

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (Rev. ed.). Holden-Day.

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Consoli, S., Reforgiato Recupero, D., & Saisana, M. (2021). *Data science for economics and finance*. Springer International Publishing.

Desai, J., Desai, K., Joshi, N., Juneja, A., & Dave, A. (2011). Forecasting of Indian Stock Market Index S&P CNX Nifty 50 Using Artificial Intelligence. *Behavioral & Experimental Finance eJournal*.

Dzikevičius, A., & Stabužytė, N. (2012). Forecasting OMX Vilnius stock index – a neural network approach. *Business: Theory and Practice*, 13(4), 324–332. <https://doi.org/10.3846/btp.2012.34>

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 383–417. <https://doi.org/10.1111/j.1540-6261.1970.tb00518.x>

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>

Gogas, P., & Papadimitriou, T. (2021). Machine learning in economics and finance. *Computational Economics*, 57, 1–4. <https://doi.org/10.1007/s10614-021-10094-w>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, F., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. <https://doi.org/10.1145/3605764.3623985>

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, 38(8), 10389–10397. <https://doi.org/10.1016/j.eswa.2011.02.068>

Hamid, S. A., & Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P 500 Index futures prices. *Journal of Business Research*, 57(10), 1116–1125. [https://doi.org/10.1016/S0148-2963\(03\)00043-2](https://doi.org/10.1016/S0148-2963(03)00043-2)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2. painos). Springer. <https://doi.org/10.1007/978-0-387-84858-7>

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Ke, Z. T., Kelly, B., & Xiu, D. (2019). Predicting returns with text data (NBER Working Paper 26186). National Bureau of Economic Research. <https://doi.org/10.3386/w26186>

Khodayari, M., Yaghobnezahad, A., & Eraghi, K. (2020). A neural-network approach to the modeling of the impact of market volatility on investment. *Advances in Mathematical Finance & Applications*, 5(4), 1–16. <https://doi.org/10.22034/amfa.2020.674953>

Lee, Y. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, 33(1), 67–74. <https://doi.org/10.1016/j.eswa.2006.04.018>

López-Lira, T., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. SSRN Working Paper. <https://doi.org/10.2139/ssrn.4412788>

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>

Maciel, L., & Ballini, R. (2010). Neural networks applied to stock market forecasting: An empirical analysis. *Learning and Nonlinear Models*, 8(1), 3–22.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 13(3), e0194889. <https://doi.org/10.1371/journal.pone.0194889>

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Moghaddam, A. H., Moghaddam, M. H., & Esfandyari, M. (2016). Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, 21(41), 89–93. <https://doi.org/10.1016/j.jefas.2016.07.002>

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2. painos). MIT Press.

Ta, V. D., Liu, C. M., & Tadesse, D. A. (2020). Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. *Applied Sciences*, 10(2), 437. <https://doi.org/10.3390/app10020437>

Tkáč, M., & Verner, R. (2016). Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, 38, 788–804. <https://doi.org/10.1016/j.asoc.2015.09.040>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, Z., Li, Y., & Chen, H. (2025). Retrieval-augmented large language models for financial time series forecasting. *arXiv preprint arXiv:2502.05878*. <https://arxiv.org/abs/2502.05878>

Wu, B., & Tingting, D. (2017). A performance comparison of neural networks in forecasting stock price trend. *International Journal of Computational Intelligence Systems*, 10(1). <https://doi.org/10.2991/ijcis.2017.10.1.23>

Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep learning for portfolio optimization. *The Journal of Financial Data Science*, 2(4), 8–20. <https://doi.org/10.3905/jfds.2020.1.042>