Ville Holopainen

Koneoppimisen hyödyntäminen vesijohtoverkostojen vuotojen hallinnassa – Systemaattinen kirjallisuuskartoitus

Tietotekniikan
pro gradu -tutkielma
18. lokakuuta 2022

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Kokkolan yliopistokeskus Chydenius

Tekijä: Ville Holopainen
Yhteystiedot: ville.holopainen1@gmail.com
Puhelinnumero: -
Ohjaaja: Lasse Harjumaa ja Risto Honkanen
Työn nimi: Koneoppimisen hyödyntäminen vesijohtoverkostojen vuotojen hallinnassa – Systemaattinen kirjallisuuskartoitus
in English: Utilization of machine learning in the management of leaks in water supply networks – A systematic mapping study
Työ: Tietotekniikan pro gradu -tutkielma
Sivumäärä: 61+7
Tiivistelmä: Tämä pro gradu -tutkielma vastaa kysymykseen ”Kuinka paljon ja minkälaista tutkimusta on tehty koneoppimisen hyödyntämisestä vesijohtoverkostojen vuotojen hallinnassa?”. Tutkimusmenetelmänä käytettiin systemaattista kirjallisuuskartoitusta, joka sekundäärisenä tutkimusmenetelmänä pyrkii tutkimusalueen jäsentämiseen. Tutkimuksessa tunnistettiin ja luokiteltiin yhteensä 80 koneoppimiseen ja vesijohtoverkostojen vuotojen hallintaan liittyvää tutkimusta ajalta 1/2012-5/2022. Vuosittaisten julkaisumäärien perusteella aiheen tutkiminen on ollut aktiivista ja kasvanut vuodesta 2017 lähtien. Julkaisufoorumien (ml. julkaisupaikkatyypit ja julkaisijat) hajanaisuudesta päätellen aihetta tutkitaan laajalla rintamalla eri tiedeyhteisöissä. Lähes kaikki luokitellut tutkimukset olivat validointitutkimuksia, mikä viittaa siihen, että tutkimustoiminta on vielä teoreettisella tasolla. Viime vuosina suosituimmat koneoppimismenetelmät pohjautuivat neuroverkkoihin, mikä vastaa koneoppimisen yleistä kehityssuuntaa. Käyttötarkoituksen perusteella koneoppimista hyödynnettiin erityisesti vuotojen tunnistamiseen ja/tai paikantamiseen. Koneoppimismallien lähtötietona käytettiin pääasiassa putkiston painetietoa, mutta myös virtaama-, ääni- ja värinätietoja.
Avainsanat: Koneoppiminen, systemaattinen kirjallisuuskartoitus, vesijohtoverkosto, vuodot
Abstract: This Master’s thesis answers the question ”How much and what kind of research has been done on the utilization of machine learning in the management of leaks in water supply networks?”. Systematic mapping study was used as the research method, which is a secondary research method that aims to structure research areas. A total of 80 studies related to the research topic from the period of 1/2012-5/2022 were identified and classified in the study. Based on the annual number of publications research on the topic has been active and growing since 2017. The scat-

teredness of the publication venues (incl. venue types and publishers) indicates that the topic is being studied in various scientific communities. Almost all the studies were validation studies, which suggests that the research is still on a theoretical level. In accordance with the general machine learning trends, the most popular machine learning methods were based on neural networks. The reason to use machine learning was mainly to identify and/or locate leaks. Especially pressure, but also flow, sound and vibration data were used as input data for machine learning algorithms.
Keywords: Machine learning, systematic mapping study, water supply network, leaks

Sanasto

Big data	Termi, jolla viitataan suuriin, vaihtelevan tyyppisiin ja suurella vauhdilla kerääntyviin tietojoukkoihin.
Koneoppiminen	Tietokoneohjelma, joka kehittyy automaattisesti kokemuksen kautta.
Näyttöön perustuva paradigma	Empiiristä tutkimusta, jossa pyritään yhdistämään paras mahdollinen tutkimusnäyttö asiantuntemuksen kanssa.
Ohjaamaton oppiminen	Koneoppimisen muoto, jonka tavoitteena on tutkia dataa, jota ei ole luokiteltu etukäteen millään tavalla.
Ohjattu oppiminen	Koneoppimisen muoto, jonka tavoitteena on ennustaa jotakin käyttäen hyväksi etukäteen luokiteltua dataa.
Puoliohjattu oppiminen	Koneoppimisen muoto, jonka tavoitteena on yleensä parantaa joko ohjatun tai ohjaamattoman oppimisen algoritmeja käyttämällä hyödyksi sekä luokiteltua että luokittelematonta dataa.
Systemaattinen kirjallisuuskartoitus	Sekundäärinen tutkimusmenetelmä, joka pyrkii ensisijaisesti tutkimusalueen jäsentämiseen.
Systemaattinen kirjallisuuskatsaus	Sekundäärinen tutkimusmenetelmä, joka pyrkii syntetisoimaan tutkimusaineistoa.
Syväoppiminen	Termi, jolla viitataan syvien neuroverkkojen koneoppimistekniikoihin.
Syvät neuroverkot	Neuroverkko, jossa on vähintään kaksi kerrosta, jotka eivät ole syöte- eikä tuloskerroksia.
Vahvistusoppiminen	Koneoppimisen muoto, jossa oppiminen tapahtuu ympäristöä tutkimalla.
Vesijohtoverkosto	Putkiverkosto, joka kuljettaa talousvettä kuluttajille.

Sisällys

1 Johdanto

Vesijohtoverkoston vuodot ovat sekä haastava ja kallis ongelma vesilaitoksille että hukkaan menevä luonnonresurssi. Vesijohtoverkostolla tarkoitetaan tavanomaisesti maan alla kulkevaa paineistettua teräs-, valurauta- tai muoviputkistoa, jota pitkin vesilaitokset toimittavat talousvettä kuluttajille. Vesijohtoverkoston muoto ja koko ovat täysin paikasta riippuvaisia. Esimerkiksi Helsingin seudun ympäristöpalvelujen (HSY) mukaan vesijohtoverkoston koko pääkaupunkiseudulla on noin 3000 km. Kyseessä voi olla siis pitkä ja laajalle levittäytynyt verkosto, jossa vuotoja esiintyy jatkuvasti. Vuodot voivat johtua esimerkiksi putkien haurastumisesta tai venttiilien hajoamisesta. Vaikka vuoto olisi suurikokoinen, ei sen tunnistaminen ja paikantaminen ole aina helppoa. Esimerkiksi vuonna 2022 Pieksämäellä etsittiin 350 kuutiometriä vuorokaudessa vuotanutta putkea useiden viikkojen ajan [69]. Tätä tekstiä kirjoittaessa Hattulassa vuotaa juomakelpoista vettä maastoon noin 200 kuutiometriä vuorokaudessa [70]. Paikallinen vesilaitos on etsinyt vuotokohtia jo kuukausien ajan ja tarjonnut rahapalkkion vuotokohdan löytymiseen johtavasta vihjeestä.

Yksinkertaisin tapa tunnistaa vuoto on luonnollisesti nähdä vettä paikassa, jossa sitä ei kuuluisi olla. Esimerkiksi kaupungin keskustoissa kaduille nousevasta vedestä tulee nopeasti tietoa vesilaitoksille. Tarkan vuotopaikan paikantaminen voi olla kuitenkin hankalaa, sillä vesi saattaa kulkea vuotokohdasta pitkiä matkoja ennen kuin se nousee ylös maanpinnalle. Toinen toimiva keino tunnistaa vuotoja on menevän ja tulevan putkistovirtauksen seuranta eri verkoston osissa. Tätä kutsutaan aluemittaukseksi. Mittaus tapahtuu useimmiten yöllä, jolloin vedenkulutus on tyypillisesti tasaisempaa kuin päivällä. Mikäli menevän ja tulevan virtaaman erotuksessa on epätavallisia poikkeamia, saattaa se tarkoittaa vuotoa verkostossa. Tällä keinolla vuoto pystytään paikantamaan tiettyyn verkoston osaan, mutta ei kovin tarkasti. Tarkempi paikantaminen tapahtuu lähettämällä alueelle ryhmä, joka etsii vuotoa käyttämällä muun muassa putkistoon kiinnitettäviä kuuntelulaitteita. Yökulutuksen seuranta on menetelmänä hidas. Mikäli vuoto sattuu päivällä, havaitaan se vasta yöllä. Keinoja ennakoivaan vuotojen hallintaan on myös olemassa. Esimerkiksi mikäli vesilaitoksella on tiedossa putken tyyppi ja asennusvuosi, voidaan etukäteen arvioida milloin putki alkaa vuotamaan ja vaihtaa se ennen sitä.

Kuten edeltä käy ilmi, on vuotojen hallinnassa runsaasti kehittämisen varaa. Yksi mahdollinen keino, jolla hallintaa voitaisiin parantaa, on koneoppimisen hyödyntäminen. Koneoppimisessa ohjelmaa (ts. algoritmia) suorittava tietokone parantaa suorituskykyään oppimalla itsenäisesti datasta, jolloin ihmisen ei tarvitse kirjoittaa ohjelmaa alusta loppuun. Koneoppimisen historia ulottuu vuosikymmenien taakse, mutta vuosituhannen alusta lähtien alalla on otettu isoja harppauksia eteenpäin. Nykyään koneoppimisen avulla voidaan esimerkiksi luokitella valokuvia ja muuntaa puhetta tekstiksi sellaisella tarkkuudella mikä ei ennen ollut mahdollista. Koneoppimisalgoritmien avulla suuresta massasta dataa voidaan tunnistaa hyvin hienojakoisia trendejä ja piirteitä suurella nopeudella, mihin tavanomaiset tietokoneohjelmat tai edes ihminen eivät kykenisi. Esimerkiksi sähköautovalmistaja Teslan koneoppimiseen pohjautuva autopilotti kykenee samalla ajanhetkellä ennustamaan (ts. tunnistamaan tietyllä todennäköisyydellä) 1000 erillistä geometrista kokonaisuutta auton ympäriltä otetuista raakakuvista [61]. Vesijohtoverkoston vuotojen hallinnan osalta koneoppimista voitaisiin käyttää esimerkiksi tunnistamaan vuotoja keskellä päivää. Päivällä normaali vedenkulutus ja siihen liittyvä verkostodata on vaihtelevampaa kuin yöllä, mikä tekee poikkeamien tunnistamisesta hankalaa perinteisillä menetelmillä. Vesijohtoverkoston vuotoihin ja koneoppimiseen liittyvää tutkimusta on jo tehty. Esimerkiksi Wun ja Zhangin [66] tutkimuksessa osoitettiin, että koneoppimisen avulla voidaan tehokkaasti ratkaista vuotojen paikantamiseen liittyviä ongelmia. Koneoppimisen kyvykkyys sekä vesijohtoverkoston vuotoihin liittyvät todelliset ongelmat luovat hyvän lähtöasetelman tälle tutkimukselle.

Tässä pro gradu -tutkielmassa suoritetaan systemaattinen kirjallisuuskartoitus koskien koneoppimisen hyödyntämistä vesijohtoverkoston vuotojen hallinnassa. Tutkimusta, joka soveltaisi systemaattista kirjallisuuskartoitusta juuri tähän aiheeseen ei ole ennen julkaistu, tai ainakaan sellaista ei löydetty. Kartoituksen päällimmäisenä tarkoituksena on jakaa vesitoimialan organisaatioille, joita pelkästään Suomessa on satoja [63], tietoa aiheeseen liittyvän tutkimustoiminnan laajuudesta, suunnasta, luonteesta ja sisällöstä. Systemaattinen kirjallisuuskartoitus sopii tähän tarkoitukseen hyvin, sillä sen ensisijaisena tarkoituksena on juuri tutkimusalueen jäsentäminen. Menetelmänä systemaattinen kirjallisuuskartoitus on hyvin monikäyttöinen. Sen avulla voidaan esimerkiksi vetää yhteen tutkimustuloksia päätöksentekijöille tai tunnistaa tutkimuskentästä aukkoja. Se eroaa tavanomaisesta kirjallisuuskartoituksesta tai -katsauksesta nimensä mukaisesti sen systemaattisella lähestymistavalla. Menetelmässä seurataan ennalta määriteltyä ja osittain iteratiivista prosessia,

jonka tarkoituksena on minimoida tutkijan subjektiivisuudesta johtuvia vääristymiä. Tutkimuksen jokainen vaihe dokumentoidaan yksityiskohtaisesti, jotta tutkimus olisi mahdollisimman hyvin toistettavissa ja tulokset tarkistettavissa.

Tutkimuksen tuloksista käy ilmi, että aiheesta on tehty ainakin 80 primääritutkimusta 1/2012-5/2022 välillä. Vuosittaisten julkaisumäärien perusteella aiheen tutkiminen on ollut aktiivista ja kasvanut vuodesta 2017 lähtien. Vuonna 2017 tutkimuksia julkaistiin vain kuusi kappaletta, kun vuonna 2021 luku oli jo 19. Eri julkaisupaikkoja oli yhteensä 56 kappaletta, mikä tarkoittaa keskimäärin 1,4 tutkimusta per julkaisupaikka. Tutkimuksia julkaistiin niin tieteellisissä lehdissä, konferensseissa kuin symposiumeissa. Julkaisupaikkojen hajanaisuudesta päätellen aiheen tutkiminen kiinnostaa tutkijoita laajalla rintamalla. Lähes kaikki tutkimukset olivat validointitutkimuksia eli tutkimuksia, johon kuuluu empiirinen arviointi mutta ei käytännön soveltamista. Tämä viittaa siihen, että aihetta koskeva tutkimustoiminta on vielä teoreettisella tasolla. Koneoppimisen yleisen kehityssuunnan mukaisesti viime vuosien suosituimmat koneoppimismenetelmät olivat neuroverkkopohjaisia. Käyttötarkoituksen perusteella koneoppimista hyödynnettiin vuotojen tunnistamiseen ja/tai paikantamiseen, pois lukien muutama tutkimus, joiden tarkoitus oli kunnon arvioiminen. Koneoppimisen lähtötietona käytettiin pääasiassa putkiston painetietoa, mutta myös virtaama-, ääni- ja värinätietoja.

Tämän tutkielman luvussa 2 käsitellään koneoppimista. Luku sisältää johdatuksen aiheeseen, esittelee kehityksen käännekohdat, oppimisprosessin sekä keskeisimmät koneoppimismenetelmät. Koneoppimisen jälkeen luvussa 3 esitellään empiirisessä osiossa käytetty tutkimusmenetelmä eli systemaattiseen kirjallisuuskartoitus. Luku alkaa johdatuksella, jota seuraa näyttöön perustustuvan paradigman sekä systemaattisen kirjallisuuskatsauksen esittely omissa luvuissaan. Luvun 3 lopussa käydään läpi tapoja suorittaa systemaattinen kirjallisuuskatsaus. Edellä mainittujen lukujen tavoitteena on antaa erityisesti aiheisiin perehtymättömällä hyvä ymmärrys siitä, mistä koneoppimisessa ja systemaattisessa kirjallisuuskartoituksessa on kyse. Luvussa 4 esitellään tutkimuksen noudattama prosessi, varsinaiset tutkimuskysymykset sekä kuinka tutkimus suoritettiin vaihe vaiheelta. Tutkimuksen tulokset eli systemaattinen kuvaus on esitelty luvussa 5. Tulosten pohjalta tehdyt johtopäätökset ja pohdinta löytyvät luvusta 6. Viimeisessä luvussa 7 on esitelty yhteenveto tutkimuksesta.

2 Koneoppiminen

Tämän tutkielman empiirinen osa koskee koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa, minkä vuoksi koneoppimiseen liittyvää teoriaa käydään tässä luvussa kattavasti läpi. Luku sisältää johdatuksen aiheeseen, esittelee kehityksen käännekohtia, oppimisprosessin sekä keskeisimmät koneoppimismenetelmät. Luvun luettuaan, myös aiheeseen täysin perehtymättömällä, lukijalla tulisi olla hyvä käsitys siitä, mistä koneoppimisessa pohjimmiltaan on kyse.

2.1 Johdatus koneoppimiseen

Perinteisesti tietokoneet on ohjelmoitu kirjoittamalla niille tarkat ohjeet, miten suorittaa jokin tehtävä. Näitä ohjeita, jotka koostuvat perättäisistä käskyistä, kutsutaan algoritmeiksi [5, s.x]. Perinteisellä tavalla ohjelmien kirjoittaminen ei kuitenkaan ole ollut mahdollista sellaisissa tehtävissä, joissa on ollut puutteellinen ymmärrys tehtävän taustalla olevasta ilmiöstä [45, s.6]. Esimerkiksi ihmisen puheen muuttaminen tekstimuotoon on sellainen tehtävä, jota on vaikea selittää, vaikka ihmiset osaavat helposti tehdä sen itse. Algoritmin kirjoittaminen perinteisellä tavalla voi olla myös käytännössä mahdotonta, mikäli analysoitavaa dataa on liikaa tai se on erittäin monimutkaista [57, s.3]. Edellä mainittujen haasteiden lisäksi perinteisiä ohjelmia rajoittaa niiden staattinen luonne eli ohjelma pysyy muuttumattomana sen jälkeen, kun se on kirjoitettu ja asennettu tietokoneelle [57, s.4]. Tämä on ongelma tilanteissa, joissa ohjelman tulisi sopeutua ympäristön muutoksiin. Esimerkiksi sähköpostin roskapostisuodatin toimii huonosti, mikäli se ei opi tunnistamaan uusia roskapostityyppejä. Edellä mainituissa tilanteissa koneoppiminen (engl. machine learning) voi auttaa.

Koneoppimisella tarkoitetaan tietokoneiden ohjelmointia siten, että tietokone kykenee optimoimaan suoritustaan esimerkkidatan tai aiemmin kokemansa avulla [4, s.3]. Koneoppimisen taustalla oli aikoinaan seuraava ajatus: *"mikäli tietokoneelle on niin vaikea kertoa, miten tietty ongelma tulisi ratkaista, niin miksi emme antaisi ohjeita epäsuorasti - siirretään tarvittavat taidot esimerkkien avulla, jolloin kone oppii!"* [39, s.xi]. Koneoppiminen on poikkitieteellinen ala [57, s.6], joka perustuu vahvasti tietotek-

niikkaan ja tilastotieteeseen, mutta myös psykologia, neurotieteet ja muut oppimista tutkivat tieteenalat liittyvät koneoppimiseen [44]. Koneoppiminen on myös osa tekoälyn kehitystä (mm. [23, s.9] ja [45, s.4]). Goodfellow et al. [23, s.8] mukaan tekoälyjärjestelmiä, jotka toimivat monimutkaisissa tosielämän ympäristöissä, ei voisi toteuttaa ilman koneoppimista. Vahvasta yhteydestä huolimatta näitä kahta käsitettä ei pidä sekoittaa, sillä ne eroavat toisistaan esimerkiksi tavoitteissaan. Tekoälyn pioneeri John McCart määritti tekoälyn vuonna 1955 karkeasti seuraavalla tavalla: *"tekoälyn tavoitteena on kehittää koneita, jotka käyttäytyvät ikään kuin ne olisivat älykkäitä"* [21, s.1]. Toisin kuin tekoäly, koneoppiminen ei pyri rakentamaan jäljitelmää älykkäästä käyttäytymisestä, vaan pyrkii täydentämään ihmisen älykkyyttä, esimerkiksi suorittamalla tehtäviä joihin ihminen ei kykene [57, s.6].

2.2 Kehityksen käännekohta

Jordan et al. [28] mukaan koneoppiminen saattaa olla yksi 21. vuosisadan uudistavimmista teknologioista. Edellä mainitusta vuosisadasta huolimatta koneoppiminen ei ole mikään uusi keksintö, vaan koneoppimisalgoritmeja on kehitetty jo yli neljäkymmenen vuoden ajan [17]. Koneoppimisen historiallinen käännekohta tapahtui kuitenkin vasta 2000-luvun alussa kolmen rinnakkaisen ja toisiaan tukevan trendin ansiosta [22]. Nämä trendit olivat "Big data", tietokoneiden muistin ja rinnakkaislaskennan halventuminen sekä syväoppimiseen (engl. deep learning) liittyvä kehitys.

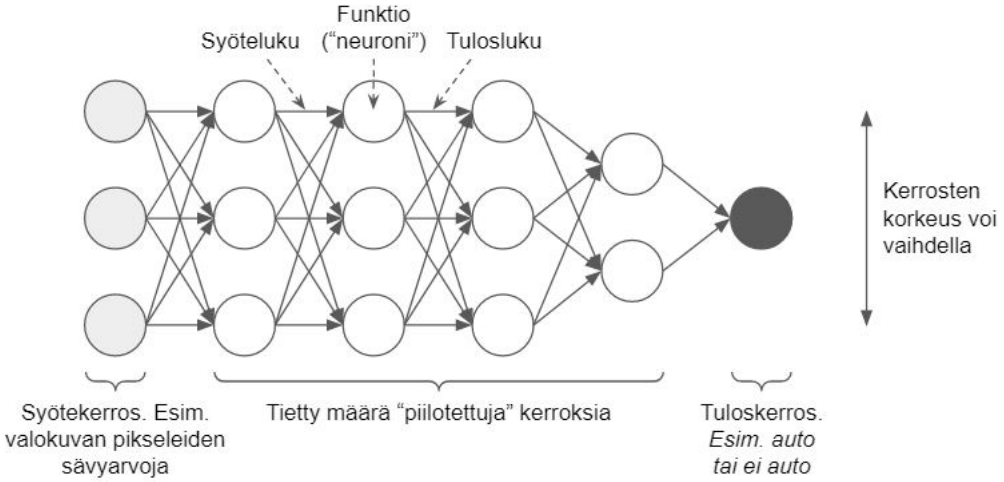
Ensimmäisellä trendillä eli Big datalla viitataan suuriin, vaihtelevan tyyppisiin ja suurella vauhdilla kerääntyviin tietojoukkoihin, joita on vaikea hallita perinteisillä ohjelmistotyökaluilla ja tekniikoilla (esim. tietokannoilla) [20]. Koneoppiminen hyötyy Big datasta, sillä suuret datamäärät mahdollistavat hienojakoisempien piirteiden tunnistamisen datasta, mikä puolestaan mahdollistaa paremmat algoritmit [74]. Vastaavasti Big data hyötyy myös koneoppimisesta. Ennen suuret tietomäärät saatettiin nähdä yrityksissä lähinnä pakollisena kulueränä, mutta koneoppimisen avulla Big datasta on tullut yritysten toimintoja ohjaava tekijä [5, ss.11-12].

Toiseen trendiin, eli muistin ja rinnakkaislaskennan halventumiseen, johti erityisesti kolme asiaa: 1) Googlen MapReduce ja avoimen lähdekoodin (engl. open source) Hadoop-teknologiat, jotka yhdessä mahdollistivat suurien datamäärien rinnakkaislaskennan yksinkertaisilla prosessoreilla, 2) tehokkaiden GPU eli grafiikkasuorittimien käyttäminen koneoppimistarkoituksiin sekä 3) tietokoneen keskus-

muistin eli RAM-muistin halventuminen, mikä mahdollisti suurien datamäärien käsittelyn suoraan keskusmuistissa [22]. Keskusmuistin etu on sen nopeus verrattuna massamuistiin (esim. kiintolevy).

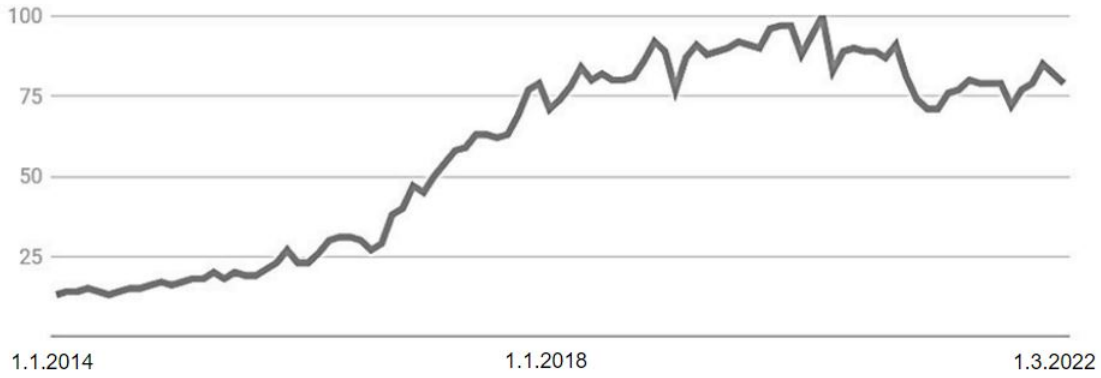
Viimeinen trendi, syväoppiminen, viittaa melko laajaan koneoppimistekniikoiden luokkaan, joka koostuu monitasoisista, epälineaarisista ja hierarkkisista tiedonkäsittelyvaiheista [16]. Yhdessä nämä tiedonkäsittelyvaiheet muodostavat niin sanotun syvän neuroverkon (engl. deep neural network), jonka rakennetta ja toimintaa on havainnollistettu kuvan 2.1 esimerkissä. Sanalla ”syvä” viitataan neuroverkkoon, jossa on vähintään kaksi ”piilotettua” kerrosta (engl. hidden layer) eli kerrosta, jotka eivät ole syöte- eikä tuloskerroksia [47, s.37]. Hierarkkisuudella viitataan siihen, että mitä pidemmälle neuroverkossa mennään, sitä monimutkaisempia ja abstraktimpia asioita kerrokset käsittelevät [47, s.37].

Syväoppimisen avulla on onnistuttu ratkaisemaan tekoälykehitystä vuosikausia jarruttaneita ongelmia ja saavutettu läpimurtoja muun muassa kuvan-, tekstin- ja puheenkäsittelyssä [40]. Justus et al. [29] arvion mukaan syväoppimisesta on tulossa nopeasti paras työkalu eri tekoälyongelmiin johtuen sen kyvykkyydestä verrattuna muihin koneoppimismenetelmiin. Syvillä neuroverkoilla on myös heikkoutensa. Yksi niistä on huono tulkittavuus, josta käytetään myös termiä ”musta laatikko” (engl. black-box) [72]. Huonolla tulkittavuudella tarkoitetaan sitä, että käyttäjä ei pysty ymmärtämään eikä perustelemaan algoritmin muodostamia tuloksia [13].



Kuva 2.1: Yksinkertaistettu esimerkki neuroverkosta, jonka tehtävä on luokitella valokuvia.

Kiinnostusta koneoppimisen käyttöön ovat edellä mainittujen trendien lisäksi lisänneet myös parantuneet apuohjelmat, joiden avulla on voitu nopeuttaa koneoppimissovellusten kehittämistä [17]. Tällaisia apuohjelmia ovat esimerkiksi avoimen lähdekoodin ohjelmistokirjastot (esim. Scikit Learn, TensorFlow ja Apache Spark MLlib) sekä pilvipalvelutarjoajien koneoppimispalvelut.



Kuva 2.2: Google hakumäärien kuukausikohtainen kehitys 1/2014-3/2022 termille ”Machine learning”.

Koneoppimiseen ja tekoälyyn kohdistuva mielenkiinto on tällä hetkellä suurta. Tarkasteltaessa Googlen kuukausittaisia hakumääriä Google Trends -Internet sivustolla [24] termille ”Machine learning” voidaan huomata selkeä kasvu vuoden 2016 jälkeen. Tämä kasvu on esitetty kuvassa 2.2, jonka asteikon arvo 100 kuvaa ajanjakson suosituinta kuukautta. Esimerkiksi arvo 50 kuvaa taas kuukautta, jolloin hakutermillä oli puolet niin paljon hakuja kuin suosituimmalla kuukaudella. Vastaava prosentuaalinen kasvu oli tapahtunut myös ”Artificial intelligence” hauille. Vuosi 2016 olikin tekoälyn ja koneoppimisen kannalta merkittävä vuosi. Tuolloin koneoppimista hyödyntävä AlphaGo-tietokoneohjelma voitti Go-nimisessä pelissä pelin 18-kertaisen maailmanmestarin [15]. Tämä oli merkittävä saavutus sen takia, että Go-peliä on pidetty tekoälylle haasteellisimpana klassisena pelinä johtuen pelin eri siirtosarjojen valtavasta määrästä ja optimaalisten siirtojen arvioimisen vaikeudesta [58].

2.3 Oppimisprosessi

Tietokoneohjelmoinnin keskiössä ovat algoritmit. Algoritmi on lista perättäisiä ohjeita, joita seuraamalla annetuista lähtötiedoista saadaan muodostettua tulos [4, s.1]. Ruokaresepti on yksi klassinen esimerkki algoritmista, jota tietokoneen sijaan suorittaa vain ihminen. Algoritmia eli reseptiä seuraamalla raaka-aineksista saadaan muodostettua ateria. Tietokoneohjelmat koostuvat erilaisista algoritmeista, jotka ohjelmistokehittäjät ovat aiemmin määrittäneet alusta loppuun. Algoritmi voi esimerkiksi laskea käyttäjän syöttämien lukujen summan ja tulostaa ratkaisun tietokoneen ruudulle. Tämän kaltaisen algoritmin kirjoittaminen on ohjelmoijalle vaivatonta. Algoritmi voidaan rakentaa yksinkertaisilla jos-niin-muutoin (engl. if-then-else) ehtolausekkeilla. Sen sijaan, jos tehtävänä olisi kirjoittaa algoritmi, joka esimerkiksi tunnistaa eläinaiheisista valokuvista kissoja, olisi tämä jo huomattavasti vaikeampaa.

Vielä 1970-luvulla uskottiin yleisesti, että älykkäitä systeemejä tulee rakentaa ehtolausekkeiden avulla insinöörien ja alan asiantuntijoiden yhteistyöllä. Muutamia onnistumisia lukuun ottamatta tämä malli kuitenkin toimi huonosti. Tämä johtui muun muassa tiedon jakamiseen liittyvistä haasteista insinöörien ja asiantuntijoiden välillä. Myös ehtolausekkeiden kirjoittaminen koettiin hankalaksi, koska niitä saattoi olla tuhansia. 1980-luvulla suunta kuitenkin muuttui, kun vaihtoehtoinen idea alkoi kerätä suosiota. Sen sijaan, että koneelle kerrotaan tarkalleen, kuinka ratkaista jokin tietty ongelma, ohjeet voisi antaa tietokoneelle epäsuorasti käyttäen esimerkkejä. Näiden esimerkkien avulla kone ”oppii”. Tämä yksinkertainen idea johti nopeasti koneoppimisen tieteenalan syntymiseen. [39, s.xi]

Ethem Alpaydinin [4, s.1] mukaan koneoppimisalgoritmin muodostaminen tapahtuu karkeasti näin: koneoppimisessa aloitetaan erittäin yleisestä mallista, joka pitää sisällään monia säädettäviä parametreja. Oppimisprosessissa näiden parametrien arvoja säädetään automaattisesti siten, että malli vastaa parhaiten koulutuksen aikana näkemäänsä dataa. Koulutusdatan ja parametrien säädön kautta tämä yleinen malli erikoistuu suorittamaan tiettyä tehtävää. Koulutuksen kautta muodostettu versio eli yleisestä mallista luotu instanssi (/ilmentymä) on tehtävää suorittava algoritmi.

Yksinkertaisessa (regressio)tehtävässä malli voi olla esimerkiksi seuraavanlainen:

$$y = \alpha x + \beta$$

Yhtälössä y on tulos, x on syöte ja α ja β ovat säädettävät parametrit. Parametre-

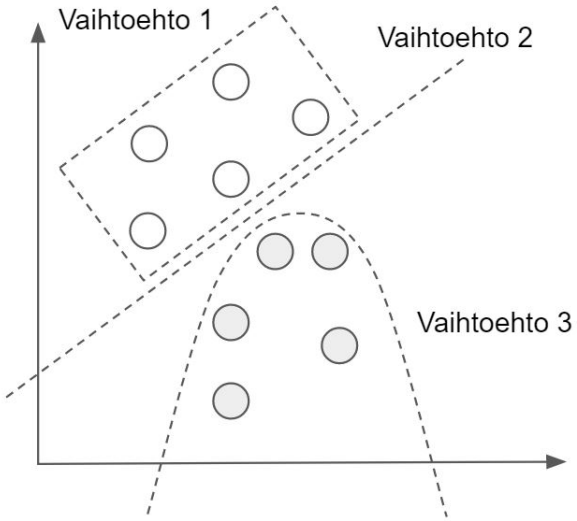
ja α ja β säätämällä tavoitteena on saada sovitettua suora viiva vastaamaan x- ja y-esimerkkidataa mahdollisimman hyvin. Parametrien säätö ei kuitenkaan ole aina näin yksinkertaista. Lecun et al. [40] mukaan tyypillisessä syväoppimisjärjestelmässä voi olla jopa satoja miljoonia parametreja, joita säädetään satojen miljoonien esimerkkien avulla.

Esimerkeistä huolimatta koneoppimisessa ei ole kyse ulkoa opettelusta, koska silloin algoritmi toimisi huonosti, kun se saa syötteenä uutta dataa. Uudella datalla tarkoitetaan dataa, jota ei ole käytetty mallin opettamiseen. Esimerkkien avulla menestyvän oppijan tulisikin tehdä toimivia yleistyksiä (engl. generalization), jota kutsutaan myös induktiiviseksi päättelyksi (engl. inductive reasoning tai inductive inference) [57, s.2]. Esimerkiksi käsin kirjoitettua tekstiä tunnistavan ohjelman tulisi siis osata tulkita kaikkia käsialoja eikä vain koulutusdatassa olevia käsialoja. Yleistyksen onnistumista mitataan harjoittelun jälkeen erillisellä testidatalla, jota ei ole käytetty mallin opettamiseen [40].

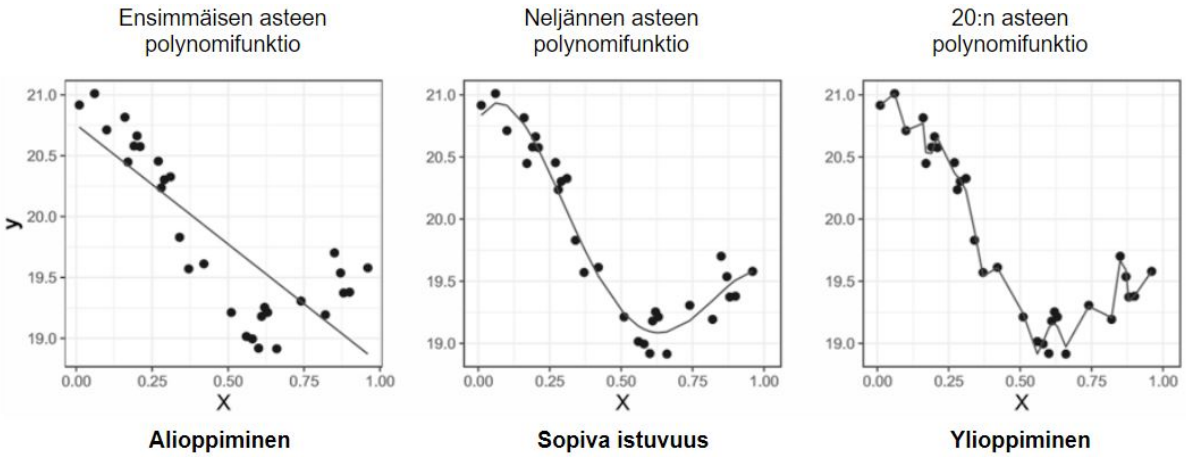
Usein koulutusdata sisältää vain pienen osan kaikista mahdollista variaatioista, sekä usein sopivia ratkaisuja voi olla monia. Jotta oppiminen olisi mahdollista ja yksiselitteinen ratkaisu saataisiin aikaan, on tehtävä tiettyjä olettamuksia. Näitä olettamuksia kutsutaan algoritmin induktiivisiksi harhoiksi (engl. inductive bias) [4, s.40]. Esimerkiksi kuvan 2.3 luokittelutehtävässä, jossa halutaan erotella tummat pisteen valkoisista, voi eri rajausvaihtoehtoja olla monia. Pisteet voidaan jakaa esimerkiksi ensimmäisen asteen polynomifunktiolla (suora viiva) tai toisen asteen polynomifunktiolla (kaareutuva viiva). Funktion astelukuun liittyvä valinta on yksi esimerkki induktiivisesta harhasta.

Mikäli valittu malli on liian yksinkertainen ja algoritmi vastaa huonosti koulutusdataa, niin puhutaan alioppimisesta (engl. underfitting). Mikäli malli on taas liian monimutkainen, ongelmana voi olla ylioppiminen (engl. overfitting). Ylioppimisessa algoritmi vastaa hyvin koulutusdataa mutta huonosti uutta dataa. Ylioppimista voidaan ehkäistä kasvattamalla koulutusdatan määrää, mutta tämä auttaa vain tiettyyn pisteeseen saakka. Ylioppimisessa haasteena voi olla myös koulutusdatassa olevat häiriöt (engl. noise), jotka malli väistämättä oppii. Molemmissa tapauksissa (ali- tai ylioppiminen) lopputuloksena on huonosti toimiva algoritmi. [4, ss.40-42].

Kuvassa 2.4 on havainnollistettu ali- ja ylioppimista. Pisteen kuvaavat koulutusdataa ja viiva koulutuksen jälkeen muodostunutta algoritmia. Ensimmäisessä kuvassa valittu malli (ensimmäisen asteen polynomifunktio) on liian yksinkertainen



Kuva 2.3: Eri vaihtoehtoja erotella tummat ja valkoiset pisteet.



Kuva 2.4: Ali- ja ylioppimista havainnollistava kuva. Lainattu lähteestä [9]

sovitettavaan dataan nähden, jolloin syntyy alioppimista. Viimeisessä kuvassa valittu malli (20:n asteen polynomifunktio) on taas liian monimutkainen käytössä olevaan koulutusdataan nähden, mistä syntyy ylioppimista. Keskellä on esitetty optimaalisempi malli (neljännen asteen polynomifunktio), joka todennäköisesti vastaa parhaiten uutta dataa.

2.4 Koneoppimismenetelmät

Oppiminen on laaja alue, minkä vuoksi koneoppiminen on jakautunut lukuisiin alikokonaisuuksiin, joissa käsitellään erilaisia koneoppimistehtäviä [57, s.4]. Käyttötarkoituksen mukaan jaoteltuna koneoppimismenetelmät voivat olla ennustavia, kuvaileva tai molempia [4, s.3]. Ennustavien menetelmien tarkoitus on antaa vastauksia tietyllä todennäköisyydellä (esim. ”tässä kuvassa on kissa 85 % todennäköisyydellä”), kun taas kuvailevien menetelmien tarkoitus on antaa jotain tietoa datasta (esim. ”tästä datasta on tunnistettavissa kolme eri ryhmää”). Toinen yleinen tapa on jakaa koneoppiminen seuraaviin neljään ryhmään: ohjattu oppiminen (engl. supervised learning), ohjaamaton oppiminen (engl. unsupervised learning), puoliohjattu oppiminen (engl. semisupervised learning) ja vahvistusoppiminen (engl. reinforcement learning) [45, s.xxi]. Edellä mainituista menetelmistä ohjattu oppiminen on yleisin koneoppimisen muoto [40].

2.4.1 Ohjattu oppiminen

Ohjatussa oppimisessa tarkoituksena on ennustaa jotakin käyttäen hyväksi etukäteen luokiteltua (engl. labelled) dataa [9]. Luokitellulla datalla viitataan syötteisiin, joilla on olemassa jokin kategoria tai numeerinen arvo. Esimerkiksi syötteenä voi olla valokuvia, jotka ovat valmiiksi luokiteltu joko kissa- tai koirakuviksi. Ohjattu oppiminen koostuu kahdesta eri ryhmästä: luokittelusta ja regressiosta [45, s.xxi]. Luokittelun tarkoituksena on ryhmitellä dataa kategorisiin luokkiin. Roskaposteja tunnistava algoritmi on yksi esimerkki luokittelusta. Harjoittelun aikana mallille syötetään tuhansia esimerkkejä erilaisista sähköposteista. Jokaisen esimerkin kohdalla mallille kerrotaan, onko kyseessä roskaposti vai tavallinen sähköposti. Näiden esimerkkien avulla malli oppii tunnistamaan roskapostien ominaispiirteet (engl. patterns), minkä jälkeen sitä voidaan käyttää uusien sähköpostien lajitteluun. Regressio tapahtuu vastaavalla tavalla, mutta siinä tuloksena on jokin juokseva luku, kuten

lämpötila tai henkilön ikä. Regression avulla voitaisiin esimerkiksi ennustaa henkilön pituutta kengännumeron perusteella.

Ohjatun oppimisen puolella käytettyjä koneoppimismenetelmiä ovat muun muassa tukivektorikone (engl. support vector machine), logistinen regressio (engl. logistic regression), naiivi bayesin luokitin (engl. Naïve Bayes), päätöspuu (engl. decision tree) ja neuroverkkopohjaiset menetelmät [12]. Buckley et al. [11] suorittama tutkimus on yksi esimerkki ohjatun oppimisen käytöstä. Tutkimuksessa onnistuttiin satunnaismetsä-menetelmän (engl. random forest) ja kehon liikedatan avulla ennustamaan onko ihminen juostessa väsynyt vai ei. Sampedro et al. [56] tutkimuksessa taas neuroverkkojen avulla onnistuttiin tunnistamaan sähköpylväitä sähkölinjoja pitkin kuvatuista videoista. Tutkimuksen tarkoituksena oli vähentää kuntotarkastuksien manuaalista työtä ja tarkastuksiin kuluvaa aikaa.

2.4.2 Ohjaamaton oppiminen

Ohjaamattoman oppimisen tarkoituksena on tutkia dataa, jota ei ole luokiteltu etukäteen millään tavalla [9]. Datan luokittelun puute voi johtua monesta syystä, muun muassa riittämättömistä resursseista tai datan luonteesta [45, s.9]. Ohjaamattomia oppimismenetelmiä käytetään muun muassa datan ryhmittelyyn (engl. clustering) ja dimensioiden vähentämiseen (engl. dimensional reduction) [9]. Ryhmittelyssä voidaan esimerkiksi jakaa ruokakaupan asiakkaita ryhmiin tehtyjen ruokaostosten perusteella, mikä taas mahdollistaa asiakkaiden kohdennetun mainonnan. Dimensioiden vähentämisen tarkoituksena voi olla taas turhien muuttujien vähentäminen datasta, tarvittavan mallin yksinkertaistaminen tai datan visualisoinnin helpottaminen [4, s.118]. Ostosdatasta voitaisiin esimerkiksi poistaa joitain muuttujia, kuten kellonaika, jotta asiakkaiden ryhmittely olisi helpompaa.

Ohjaamattoman oppimisen menetelmiä ovat muun muassa k:n keskiarvon klusterointi (engl. k-means clustering), hierarkkinen klusterointi (engl. hierarchical clustering) ja pääkomponenttianalyysi (engl. principal component analysis) [3]. Myös syväoppimista käytetään ohjaamattomassa oppimisessa [40]. Esimerkiksi Bhatia et al. [10, s.118] tutkimuksessa käytettiin onnistuneesti neuroverkkopohjaista autoenkooderia (engl. autoencoder) tunnistamaan palvelunestohyökkäyksiä (engl. Distributed Denial of Service, DDoS) reaaliaikaisesti IoT-laitteiden tuottamasta verkkoliikenteestä.

2.4.3 Puoliohjattu oppiminen

Puoliohjattu oppiminen on koneoppimisen muoto, jonka pyrkimyksenä on tyypillisesti parantaa joko ohjatun tai ohjaamattoman oppimisen algoritmeja käyttämällä hyödyksi sekä luokiteltua että luokittelematonta dataa. Esimerkiksi luokittelutehtävässä luokitellun datan lisäksi voidaan käyttää hyväksi luokittelematonta dataa. Vastaavasti ryhmittelytehtävässä voidaan käyttää luokittelemattoman datan lisäksi luokiteltua dataa. Puoliohjatun oppimisen tutkimus keskittyy kuitenkin enimmäkseen luokittelutehtävien ympärille, kuten koneoppiminen yleisestikin. Luokittelutehtävissä puoliohjatun oppimisen käyttö on erityisen relevanttia silloin, kun luokiteltua dataa on vähän saatavilla. Puoliohjatun oppimisen ongelmana voi kuitenkin olla algoritmin suorituskyvyn heikkeneminen verrattuna siitä, että tehtävässä käytettäisiin ohjatun oppimisen algoritmia. [62]

Puoliohjatussa oppimisessa käytetään muun muassa ohjatun oppimisen menetelmiä ja niiden laajennuksia sekä graafeihin perustuvia (engl. graph-based) menetelmiä [62]. Puoliohjattuja oppimismenetelmiä on kokeiltu muun muassa lunnasohjelmien tunnistamiseen [48] sekä valokuvien kategorisoimiseen [25] positiivisin tuloksin.

2.4.4 Vahvistusoppiminen

Vahvistusoppimisessa ei ole dataa käytössä [45, s.7], kuten muilla koneoppimismenetelmillä, vaan oppiminen tapahtuu ympäristöä tutkimalla [30]. Vahvistusoppimisessa tiettyä ongelmaa ratkaiseva toimija (engl. agent) kokeilee eri toimenpiteitä ympäristössä saaden niistä palautetta joko palkinnon tai rangaistuksen muodossa [4, s.563]. Saadun palautteen perusteella toimija optimoi käytöstään tavoitteenaan maksimoida saadut palkinnot ja minimoida rangaistukset [39, s.331]. Tietyn harjoittelumäärän jälkeen toimijan tulisi oppia paras toimintatapa, joka johtaa suurimpaan kumulatiiviseen palkintoon [4, s.563]. Haastavimmissa oppimistehtävissä toimijan valitsema yksittäinen toimenpide ei välttämättä vaikuta vain välittömään palkintoon, vaan myös sitä seuraaviin tilanteisiin ja palkintoihin [59, s.1].

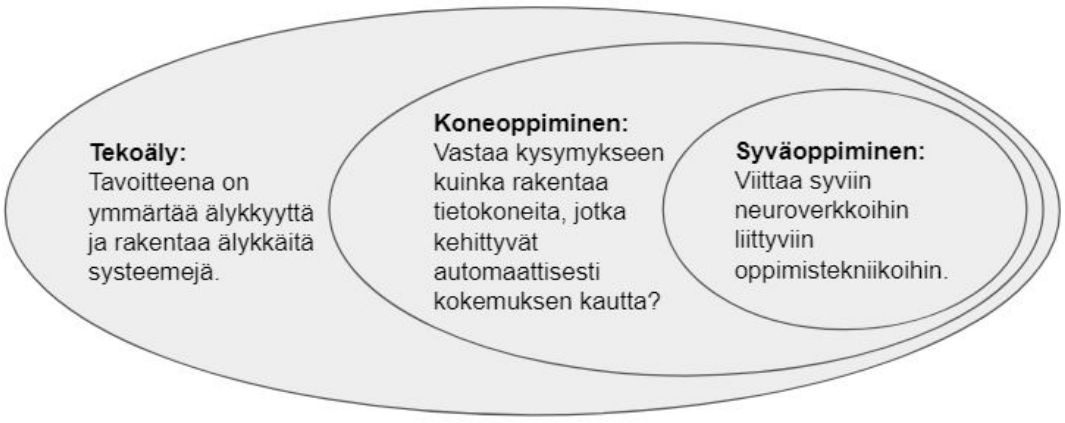
Vahvistusoppimista voidaan havainnollistaa Alpaydinin [4, s.563] shakkipeliesimerkin avulla. Siinä pelilauta on ympäristö, toimija on pelaaja ja toimenpide on yksittäinen siirto. Harjoitellessaan toimija tekee siirtoja ja saa täydestä sarjasta siirtoja palautetta (voitit tai hävisit). Pelissä yksittäiset siirrot eivät ole merkityksellisiä vaan oleellista on peräkkäisten siirtojen kautta voittaa peli. Harjoittelun aikana pe-

laaja oppii eri pelitilanteiden ja yksittäisten siirtojen merkityksen voiton suhteen, ja sen perusteella optimoi algoritmia.

Vahvistusoppimismenetelmien ryhmään kuuluvat muun muassa Monte Carlo, Q-oppiminen (engl. Q-learning), SARSA ja dynaaminen ohjelmointi (engl. dynamic programming) [46]. Vahvistusoppimista voidaan hyödyntää muun muassa terveydenhuoltoalalla, roboteissa, autonomisessa ohjauksessa, peleissä ja luonnollisen kielen käsittelyssä (engl. natural language processing) [46]. Vahvistusoppimista hyödyntävä tutkimus on koskenut esimerkiksi itseohjautuvia autoja [75], yksilöllisten hoito-ohjelmien kehittämistä sairauksia vastaan [73] sekä Go-lautapelin pelaamista [58].

2.5 Yhteenveto

Tässä luvussa käytiin koneoppimiseen liittyvää teoriaa kattavasti läpi, koska tutkielman empiirinen osa käsittelee koneoppimista ja sen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Luvussa esiteltiin muun muassa, että koneoppimisella tarkoitetaan tietokoneiden ohjelmointia siten, että tietokone kykenee optimoimaan suoritustaan esimerkkidatan tai aiemmin kokemansa avulla. Koneoppimista voidaan käyttää esimerkiksi silloin, kun tietokoneen ohjelmointi ei ole mahdollista perinteisellä tavalla eli kirjoittamalla tietokoneelle tarkat ohjeet, miten suorittaa jokin tehtävä. Tällainen tilanne voi syntyä esimerkiksi silloin, kun analysoitavaa dataa on liikaa tai se on hyvin monimutkaista.



Kuva 2.5: Tekoäly [21, s.vii], koneoppiminen [4, s.3] ja syväoppiminen [28].

Koneoppiminen on tekoälyn osa-alue, jota ilman tekoälyn kehittäminen ei ehkä

olisi mahdollista. Toisin kuin tekoäly, koneoppiminen ei pyri rakentamaan jäljitelmää älykkäästä käyttäytymisestä, vaan pyrkii täydentämään ihmisen älykkyyttä. Koneoppimisen historia ulottuu kymmenien vuosien päähän, mutta varsinainen läpimurto tapahtui vasta 2000-luvun alussa. Läpimurron mahdollisti kolme rinnakkaista trendiä: ”Big data”, tietokoneiden muistin ja rinnakkaislaskennan halventuminen sekä syväoppimiseen liittyvä kehitys. Syväoppimisella viitataan syviin neuroverkkoihin liittyviin koneoppimistekniikoihin. Syväoppimisen, koneoppimisen ja tekoälyn keskinäinen suhde ja määritykset ovat esitetty kuvassa 2.5.

Yleinen kiinnostus koneoppimista kohtaan on tällä hetkellä suurta ja erilaisia koneoppimista hyödyntäviä sovelluksia syntyy jatkuvasti. Koneoppimista hyödynnetään muun muassa itseohjautuvissa autoissa ja puheentunnistuksessa. Loistavista tuloksista huolimatta koneoppimiseen liittyy monia haasteita, kuten sopivan koneoppimismallin valinta ja laadukkaan koulutusdatan puute.

3 Systemaattinen kirjallisuuskartoitus

Tässä luvussa esitellään systemaattinen kirjallisuuskartoitus -tutkimusmenetelmä. Luku alkaa johdatuksella, jossa käydään läpi systemaattisen kirjallisuuskartoituksen taustoja ja päämääriä. Tämän jälkeen esitellään kaksi systemaattiseen kirjallisuuskartoitukseen vahvasti liittyvää käsitettä, joihin lukijan on hyvä tutustua: näyttöön perustuva paradigma ja systemaattinen kirjallisuuskatsaus. Luvun lopussa esitellään, kuinka systemaattinen kirjallisuuskatsaus voidaan suorittaa. Luvun luettuaan lukijalla tulisi olla hyvä ymmärrys muun muassa systemaattisten menetelmien tärkeydestä, historiasta sekä suoritustavoista.

3.1 Johdatus systemaattiseen kirjallisuuskartoitukseen

Kun tutkimusala kypsyy, saatavilla olevien julkaisujen ja tulosten määrä lisääntyy usein jyrkästi, jolloin aiheesta on tarpeellista tehdä yhteenvetoja ja tarjota yleiskatsauksia [50]. Ilman perusteellista ja objektiivista tutkimusmenetelmää katsauksilla on kuitenkin vain vähän tieteellistä arvoa [34]. Esimerkiksi Kitchenhamin et al. [33] mielestä on selvää, että tuloksiin vaikuttavia lähteitä voi puuttua, mikäli aineiston etsintä- ja valintaprosessi ei ole selkeästi jäsennelty.

Eräs perusteellinen ja läpinäkyvä menetelmä, jolla kirjallisuutta voidaan kartoittaa, on systemaattinen kirjallisuuskartoitus (engl. systematic mapping study tai scoping study) [7]. Kyseessä on sekundäärinen eli toissijainen tutkimusmenetelmä, mikä tarkoittaa sitä, että tutkimuksessa ei tuoteta uutta dataa suorilla mittauksilla vaan käytetään hyväksi primääri- eli alkuperäisiä tutkimuksia [37, s.xxxii]. Aineisto voi kuitenkin pohjautua myös toisiin sekundäärisiin tutkimuksiin, kuten tehtiin esimerkiksi Petersen et al. [52] systemaattisessa kirjallisuuskartoituksessa, jossa kartoitettiin eri systemaattisia kirjallisuuskartoitusmenetelmiä.

Systemaattisen kirjallisuuskartoituksen päämääränä on tunnistaa mitä näyttöä /todistusaineistoa (engl. evidence) on saatavilla tietystä aiheesta [34] sekä luokitella tätä aineistoa [35]. Arksey ja O'Malley [7] mukaan systemaattisen kirjallisuuskartoituksen tarkempi tavoite riippuu tutkimuksesta itsestään, mutta ainakin seuraavat neljä syytä on tunnistettu:

1. Sen avulla voidaan tarkastella tutkimustoiminnan laajuutta, vaihtelevuutta ja luonnetta. Tällöin kyseessä on nopea katsaus, jossa ei välttämättä mennä yksityiskohtiin.

2. Sen avulla voidaan selvittää kuinka järkevää systemaattisen kirjallisuuskatsauksen tekeminen olisi. Voidaan esimerkiksi selvittää löytyisikö riittävästi aineistoa tai onko aiheesta tehty jo systemaattinen kirjallisuuskatsaus.

3. Se tarjoaa menetelmän vetää yhteen ja jakaa tutkimustuloksia. Tällöin kyseessä saattaa olla yksityiskohtaisempi tutkimus, jossa tuloksia jaetaan esimerkiksi päätöksentekijöille.

4. Sen avulla voidaan tunnistaa kirjallisuudessa olevia tutkimusaukkoja, ottamatta kuitenkaan kantaa tutkimuksien laatuun.

Vielä 2000-luvun alussa systemaattinen kirjallisuuskartoitus oli saanut tutkimusmenetelmänä vähän huomiota ja sen suorittamiseksi oli olemassa niukasti tietoa [7]. Petersenin et al. [52] mukaan 2015 tilanne oli jo kuitenkin toinen — systemaattisten kartoitustutkimusten määrä ja kiinnostus kasvoi jatkuvasti eri aloilla. Vuonna 2014 julkaistussa systemaattisessa kirjallisuuskartoituksessa [53] tunnistettiin yhteensä 344 systemaattisista kirjallisuuskartoitusta vuosien 1999 ja 2012 väliltä. Tutkimuksen mukaan jopa lähes kolme neljäsosaa kartoituksista koski terveydenhoitoalaa. Muita merkittäviä sovellusalueita olivat ohjelmistotuotanto (n. 12 %), opetusala (n. 4,5 %) ja yhteiskuntatieteet (n. 4 %). Ohjelmistoalalla systemaattisia kirjallisuuskartoituksia on tehty lukuisista ei aiheista, muun muassa startup-yritysten ohjelmistokehityksestä [49], mikropalveluarkkitehtuureista [6], teknisestä velasta ja sen hallitsemisesta [42], DevOps:n määritelmästä ja käytännöistä [27] ja koneoppimisen soveltamisesta ohjelmistotestaukseen [17].

3.2 Näyttöön perustuva paradigma

Systemaattisen kirjallisuuskartoituksen juuret tulevat näyttöön perustuvasta paradigmasta. Näyttöön perustuva paradigma (engl. evidence-based paradigm) tarkoittaa *"empiiristä tutkimusta, jossa tutkija pyrkii löytämään parasta mahdollista tutkimusnäyttöä ja yhdistämään sitä toimialueen asiantuntemuksen kanssa, tarjotakseen tietoa ammattilaisille ja päätöksentekijöille"* [37, s.xxvi]. Se on kehitetty terveydenhoitoalalla, jossa potilaita hoitavien lääkärien täytyi kerätä ja yhdistää lääketieteellistä aineistoa

potilaiden hoitopäätösten tueksi [37, s.7]. David Sackettin [55] mukaan ilman parhaan mahdollisen hoitoon liittyvän ulkopuolisen tutkimusaineiston käyttöä hoitokäytännöt voisivat nopeasti vanhentua ja aiheuttaa potilaille vahinkoa. Terveydenhoitoalalla paradigman soveltamisesta alettiin käyttämään nimitystä näyttöön perustuva lääketiede (engl. Evidence-Based Medicine tai lyh. EBM), jolla tarkoitetaan *"yksilöllisen kliinisen asiantuntemuksen ja parhaimman mahdollisen ulkopuolisen kliinisen todistusaineiston, joka on hankittu systemaattisin menetelmin, yhdistämistä"* [55]. Monet lääketieteen tutkijat omaksuivat paradigman käytön 90-luvulta lähtien ja alan tutkimus muuttui sen myötä radikaalisti [38]. Muun muassa hoitojen ja toimenpiteiden tehokkuudesta kertovien systemaattisten kirjallisuuskatsausten määrä rupesi kasvamaan ja ohjeita katsauksien suosittamiseksi tuli saataville [7].

Ohjelmistoalalla näyttöön perustuva tutkimus näyttää saaneen alkunsa vuonna 2004, kun Kitchenham et al. [38] esittelivät EMB:n pohjalta näyttöön perustuvan ohjelmistokehityksen (engl. Evidence-Based Software Engineering, EBSE). Heidän mukaansa EBSE:n tavoite oli *"tarjota keinot, joilla nykyhetken paras mahdollinen todistusaineisto voidaan yhdistää päätöksentekoprosessissa käytännön osaamisen ja ihmisarvojen kanssa, koskien ohjelmistojen kehittämistä ja ylläpitoa"*. Tavoitteen lisäksi he esittelivät EMB:stä johdetut viisi askelta EBSE:n harjoittamiseksi [38] .

1. Muutetaan tiedon tarve (esim. koskien ohjelmistonkehitys- tai ylläpitomenetelmiä) kysymysmuotoon.

2. Etsitään parasta näyttöä, joilla kysymykseen voidaan vastata.

3. Arvioidaan kriittisesti aineiston pätevyyttä, vaikuttavuutta sekä soveltuvuutta.

4. Yhdistetään edellä syntynyt arvio ohjelmistotekniikan asiantuntemuksemme ja sidosryhmien arvojen ja olosuhteiden kanssa.

5. Arvioidaan edellisten vaiheiden vaikuttavuutta ja tehokkuutta, ja pyritään kehittämään niitä seuraavaa tutkimusta varten.

Kitchenham et al. [38] artikkelissa pohjustettiin EBSE:n merkitystä sillä, että ohjelmistoilla alkaa olla niin keskeinen rooli ihmisten jokapäiväisessä elämässä, ja huonoilla teknologiavalinnoilla negatiiviset seuraukset voivat olla merkittävät. Esimerkiksi kannettavat lääkintälaitteet voivat uhata yksilön henkeä ja pieleen mennyt julkinen hankinta voi tulla yhteiskunnalle kalliiksi. Dyba et al. [18] mukaan ohjelmis-

toyritykset saattavatkin tehdä huonoja teknologiapäätöksiä, mikäli valintojen tueksi ei ole puolueetonta aineistoa koskien teknologian sopivuutta, rajoituksia, laatua, hintaa tai riskejä. Esimerkiksi markkinoiden tai yrityksen johdon paineiden alla saatetaan ottaa käyttöön epäkypsiä teknologioita. Tämän vuoksi he suosittelevatkin yrityksiä tutustumaan EBSE:n käyttöön. Vaikka harva kaupallinen ohjelmistokehityshanke noudattaa vielä näyttöön perustuvaa paradigmaa, alkaa se olla jo laajasti tunnistettu ja arvostettu ohjelmistoalalla [37, s.xix].

3.3 Systemaattinen kirjallisuuskatsaus

Systemaattinen kirjallisuuskatsaus on näyttöön perustuvan tutkimuksen keskeisin työkalu [37, s.xix], jonka avulla voidaan muun muassa minimoida tutkijan subjektiivisuudesta johtuvia vääristymiä [38]. Systemaattinen kirjallisuuskatsauksen (engl. systematic literature review tai pelkkä systematic review) tarkka määritelmä on *sekundäärinen tutkimusmenetelmä, joka pyrkii tarjoamaan objektiivisen ja ennakkoluulottoman tavan löytää asiaankuuluvia primääritutkimuksia, sekä poimia, kerätä ja syntetisoida niistä tietoa* [37, s.xxxiii]. Yleisin syy menetelmän käyttöön on halu tehdä yhteenvetoja todistusaineistoista, tunnistaa tutkimusaukkoja ja pohjustaa tulevia tutkimuksia [34]. Ohjelmistoalalla ensimmäisen ohjeen systemaattisten kirjallisuuskatsauksen suorittamiseksi julkaisivat Kitchenham ja Charters [34] vuonna 2007. Ohjeen päävaiheet ovat esitelty tämän tutkielman liitteissä A.

Systemaattisessa kirjallisuuskatsauksessa on paljon päällekkäisyyttä systemaattisen kirjallisuuskartoituksen kanssa. Systemaattinen kirjallisuuskartoitus voidaankin nähdä joko systemaattisen kirjallisuuskatsauksen muotona [35] tai sen alivaiheena [65]. Ne eivät ole kuitenkaan sama asia, sillä niiden päämäärät ovat hyvin erilaiset. Petersen et al. [52] mukaan systemaattisen kirjallisuuskatsauksen tavoitteena on syntetisoida todistusaineistoa, kun taas systemaattisen kirjallisuuskartoitus pyrkii ensisijaisesti tutkimusalueen jäsentämiseen. Päämäärän lisäksi myös niiden sisältö ja aineiston analysointitapa ovat erilaiset [65]. Taulukossa 3.1 on esitetty menetelmien keskeisimpiä eroavaisuuksia niin tavoitteiden, tutkimuskysymyksen, etsintäprosessin, aineiston laajuuden, etsintästrategian, aineiston laadun arvioinnin kuin tulosten näkökulmasta. Taulukon perusteella erot näiden kahden menetelmän välillä ovat ilmeisiä yhtäläisyyksistä huolimatta.

Taulukko 3.1: Systemaattisen kirjallisuuskartoituksen ja systemaattisen kirjallisuuskatsauksen eroja [36].

	Systemaattinen kirjallisuuskartoitus	Systemaattinen kirjallisuuskatsaus
Tavoitteet	Kirjallisuuden luokittelu ja temaattinen analyysi.	Parhaiden käytäntöjen tunnistaminen yhdistämällä tietoja eri tutkimuksista.
Tutkimuskysymys	Yleinen - koskee tutkimustrendejä.	Tarkka - koskee empiiristen tutkimusten tuloksia.
Etsintäprosessi	Aihealueen mukaan määritelty.	Tutkimuskysymyksen mukaan määritelty.
Laajuus	Laaja - kaikki aihealueeseen liittyvät tutkimukset otetaan mukaan. Voi sisältää sekä teoreettisia että empiirisiä tutkimuksia.	Rajattu - vain tutkimuskysymykseen liittyvät empiiriset tutkimukset ovat mukana.
Etsintästrategian vaatimukset	Useimmiten ei niin tiukat. Voidaan esim. valita mukaan vain kaksi digitaalista tietokantaa.	Erittäin tiukka. Kaikki asiaankuuluvat tutkimukset tulisi löytää.
Aineiston laadun arviointi	Ei oleellista.	Oleellista. On tärkeää varmistua siitä, että primääritutkimukset ovat laadukkaita.
Tulokset	Moniulotteinen luokittelu primääritutkimusten lukumäärien perusteella.	Tutkimuskysymykseen vastaaminen vetämällä yhteen primääritutkimuksia.

3.4 Systemaattinen kirjallisuuskartoitusprosessi

Systemaattisia kirjallisuuskartoituksia suoritetaan monilla eri tutkimusaloilla hyödyntäen erilaisia ohjeita ja menetelmiä [52]. Levacin et al. [41] mukaan ensimmäisen viitekehyksen systemaattisen kirjallisuuskartoituksen suorittamiseksi julkaisivat Arksey & O'Malley [7] vuonna 2005. Kyseinen viitekehys sisältää kuusi eri vaihetta:

1. Tutkimuskysymyksen muodostaminen. Ensimmäinen vaihe on hyvin tärkeä, koska sen perusteella valitaan tutkimukselle sopiva kartoitusstrategia.

2. Relevanttien primääritutkimuksien kerääminen mahdollisimman kattavasti.

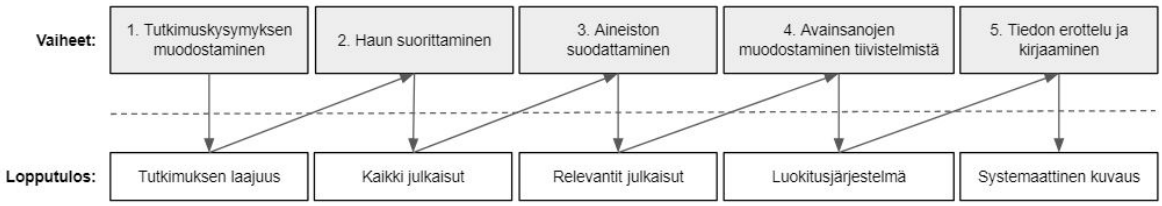
3. Relevanttien primääritutkimuksien suodattaminen sisällyttämis- ja poissulkemiskriteerien avulla. Tarkoituksena on ottaa mukaan ne tutkimukset, jotka vastaavat ensimmäisessä vaiheessa muodostettuun tutkimuskysymykseen.

4. Primääritutkimuksissa olevien tietojen kerääminen ja kirjaaminen taulukkomuotoon. Tietoihin voi kuulua esimerkiksi tekijä, julkaisuvuosi, tutkimuksen tarkoitus, menetelmä sekä oleelliset tutkimustulokset.

5. Tulosten kokoaminen yhteen, yhteenvedon tekeminen sekä tulosten raportointi.

6. Valinnainen vaihe: Sidosryhmien konsultoiminen, minkä tarkoituksena on parantaa tutkimusta. Sidosryhmät voivat muun muassa auttaa mahdollisten primääritutkimuksien löytämisessä tai jakaa omia näkemyksiään.

Vuonna 2015 Petersenin et al. [52] suorittamassa systemaattisessa kirjallisuuskartoituksessa tunnistettiin yhteensä kymmenen ohjelmistoalalla käytettyä ohjetta, mukaan lukien edellä mainittu Arkseyn & O'Malleyn ohje. Monessa tutkimuksessa sovellettiin samaan aikaan useita eri ohjeita, mikä tutkimuksen mukaan saattaa viitata ohjeiden puutteellisuuteen. Selkeästi suosituin ohje oli Petersenin et al. [50] ohje vuodelta 2008. Toiseksi suosituin oli Kitchenhamin [31] ohje vuodelta 2004 sekä sen päivitetty versio [34] vuodelta 2007. Kitchenhamin ohjeet keskittyvät kuitenkin systemaattisen kirjallisuuskatsauksen suorittamiseen, eivätkä ota juuri kantaa systemaattiseen kirjallisuuskartoitukseen. Petersenin et al. [50] ohje sen sijaan koskee juuri systemaattisen kirjallisuuskartoituksen suorittamista.



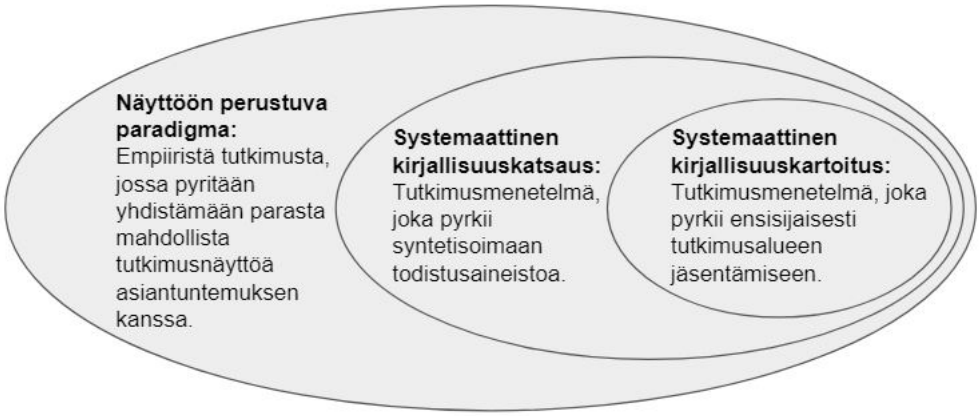
Kuva 3.1: Systemaattinen kirjallisuuskartoitus -prosessi [50].

Petersenin et al. [50] esittämä prosessi, joka vastaa hyvin paljon Arkseyn & O'Malleyn [7] prosessia, koostuu viidestä peräkkäisestä vaiheesta. Ensimmäisessä vaiheessa muodostetaan tutkimuskysymykset, joiden perusteella tulisi selvitä tutkimuksen laajuus. Toisessa vaiheessa suoritetaan aineiston haku, jonka tavoitteena on tunnistaa kaikki tutkimukseen liittyvät julkaisut. Kolmannessa vaiheessa julkaisut suodatetaan valittujen kriteerien mukaan, minkä lopputuloksena ovat relevantit julkaisut. Neljännessä vaiheessa suodatetusta aineistosta poimitaan avainsanoja, joiden pohjalta muodostetaan tutkimuksen luokitusjärjestelmä. Viimeisessä vaiheessa julkaisut käydään uudelleen läpi ja niistä tehdään yksilölliset kirjaukset luokitusjärjestelmään. Kirjauksista syntyneestä datasta muodostetaan lopulta varsinaiset tulokset eli niin kutsuttu systemaattinen kuvaus. Edellä kuvattu prosessi on esitetty kuvassa 3.1. Prosessin eri vaiheita käydään läpi vielä yksityiskohtaisemmin seuraavassa luvussa, koska ohjetta päätettiin soveltaa tämän tutkielman empiirisessä vaiheessa.

3.5 Yhteenveto

Tässä luvussa käytiin systemaattiseen kirjallisuuskartoitukseen liittyvää teoriaa kattavasti läpi. Tämä tehtiin sen vuoksi, että tutkielman empiirisessä vaiheessa suoritetaan systemaattinen kirjallisuuskartoitus, joka koskee koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Luvussa esiteltiin muun muassa, että kun tutkimusala kypsyy ja saatavilla olevien julkaisujen määrä kasvaa, on usein tarpeellista tehdä yhteenvetoja ja tarjota yleiskatsauksia. Mikäli katsauksia tehdään ilman perusteellista ja objektiivista tutkimusmenetelmää, on tutkimuksella vain vähän tieteellistä arvoa. Yksi menetelmä, jolla kirjallisuutta voidaan kartoittaa perusteellisesti ja läpinäkyvästi on systemaattinen kirjallisuuskartoitus. Systemaattisen kirjallisuuskartoituksen tarkempi tavoite on tutkimuskohtainen, mutta ensisijaises-

ti sitä käytetään tutkimusalueen jäsentämiseen.



Kuva 3.2: Systemaattinen kirjallisuuskartoitus, -katsaus [52] ja näyttöön perustuva paradigma [37, s.xxvi].

Systemaattisen kirjallisuuskartoituksen juuret tulevat näyttöön perustuvasta paradigmasta. Näyttöön perustuva paradigma on empiiristä tutkimusta, jossa tutkija pyrkii löytämään parasta mahdollista tutkimusnäyttöä ja yhdistämään sitä toimialueen asiantuntemuksen kanssa. Näyttöön perustuvalla tutkimuksella tulisi olla merkittävä rooli ohjelmistoalalla, sillä ohjelmistoilla on erittäin tärkeä ja jopa kriittinen rooli ihmisten jokapäiväisessä elämässä. Näyttöön perustuvan tutkimuksen keskeisin työkalu on systemaattinen kirjallisuuskatsaus -tutkimusmenetelmä, joka pyrkii syntetisoimaan todistusaineistoa. Kyseessä on hyvin samantyyppinen menetelmä kuin systemaattinen kirjallisuuskartoitus, mutta ne eivät ole sama asia. Systemaattinen kirjallisuuskartoitus voidaan nähdä joko systemaattisen kirjallisuuskatsauksen muotona tai sen alivaiheena. Kuvassa 3.2 on esitetty systemaattisen kirjallisuuskartoituksen, -katsauksen ja näyttöön perustuvan paradigman määritelmät sekä niiden keskinäiset suhteet. Systemaattisia kirjallisuuskartoituksia suoritetaan monilla eri aloilla hyödyntäen useita eri ohjeita. Yksi suosittu menetelmä on tässäkin tutkimuksessa hyödynnetty Petersenin et al. [50] ohje.

4 Tutkimuksen toteutus

Tässä luvussa esitellään systemaattinen kirjallisuuskartoitus, joka koskee koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Tutkimuksen päätarkoituksena on tarjota vesitoimialan organisaatioille tietoa tutkimustoiminnan laajuudesta, suunnasta, luonteesta ja sisällöstä. Luvussa käydään läpi tutkimuksessa sovellettu prosessi ja sen eri vaiheiden suoritus. Luvun luettuaan lukijalla tulisi olla hyvä ymmärrys siitä, kuinka tutkimus on toteutettu. Eri vaiheiden yksityiskohtaisen dokumentoinnin tarkoituksena on helpottaa tutkimuksen toteutuksen arvioimista sekä toistettavuutta.

4.1 Tutkimuksen noudattama prosessi

Tutkimuksessa sovelletaan Petersenin et al. [50] systemaattisen kirjallisuuskartoituksen prosessia, sekä siihen vuonna 2015 tulleita Petersenin et al. [52] päivityksiä. Merkittävin muutos aikaisemmin esitetyn kuvan 3.1 prosessiin on kuudennen vaiheen lisääminen. Tässä vaiheessa arvioidaan tutkimuksen validiteettia ja toistettavuutta koskevia uhkia. Tutkimuksen noudattaman prosessin eri vaiheet ja niiden keskeisin sisältö on esitetty kuvassa 4.1. Tutkimuksessa lähdetään liikkeelle vaiheesta yksi, jonka lopputulos on syöte seuraavalle vaiheelle. Vaiheen kaksi lopputulos on taas syöte kolmannelle vaiheelle ja niin edelleen. Vaiheen viisi lopputulos on tutkimuksen lopullinen tulos eli tutkittavan aiheen systemaattinen kuvaus. Tutkimus päättyy vaiheeseen kuusi eli tutkimuksen validiteettia ja toistettavuutta koskevien uhkien arvioimiseen.

4.2 Vaihe 1: tutkimuskysymyksen muodostaminen

Petersenin et al. [50] prosessin ensimmäisen vaiheen tarkoituksena on määritellä tutkimuksen tavoitteet ja niiden pohjalta varsinaiset tutkimuskysymykset. Petersenin et al. päivityksessä [52] nostetaan esille myös systemaattisen kirjallisuuskartoituksen tarpeen pohjustamisista. Tämän tutkimuksen päätavoitteena on tarjota vesitoimialan organisaatioille tietoa tutkimustoiminnan laajuudesta, suunnasta,

Kuva 4.1: Tutkimuksen noudattama prosessi ja sen keskeisin sisältö.

luonteesta ja sisällöstä koskien koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Systemaattinen kirjallisuuskartoitus menetelmänä sopii tutkielman tarkoitukseen hyvin, sillä se pyrkii juuri tutkimusalueen jäsentelyyn.

Samasta aiheesta ei ole tehtyä aiemmin systemaattista kirjallisuuskartoitusta tai -katsausta, tai ainakaan sellaista ei löydetty. Aihetta sivuavia sekundäärisiä tutkimuksia tunnistettiin kuitenkin muutama kappale. Esimerkiksi Wun & Liun [67] vuoden 2017 katsauksessa esitellään dataan perustuvia menetelmiä, joilla voidaan tunnistaa putkivuotoja. Vuotta myöhemmin Chan et al. [14] julkaisivat artikkelin, jossa esiteltiin älykkäitä menetelmiä vuotojen havaitsemiseksi. Edellä mainittujen lisäksi Zaman et al. [71], El-Zahabin & Zayedin [19] Guptan & Gulatin [26], Adedejin et al. [2], Xun et al. [68] ja Puustin et al. [54] katsauksissa on esitelty erilaisia vuotojen hallintaan liittyviä menetelmiä. Ainoa systemaattisin menetelmin suoritettu tutkimus oli Tariqin et al. [60] systemaattinen kirjallisuuskatsaus, joka koski sähkömekaanisten mikrosysteemien (engl. micro electro mechanical systems) käyttöä vesijohtoverkoston vuotojen havaitsemisessa ja paikantamisessa.

Tämä systemaattinen kirjallisuuskartoitus pyrkii vastaamaan kysymykseen ”Kuinka paljon ja minkälaista tutkimusta on tehty koneoppimisen käytöstä vesijohto-

verkostojen vuotojen hallinnassa?”. Tutkimuksen tarkemmat tutkimuskysymykset ovat seuraavat:

1. Kuinka aktiivista aiheen tutkiminen on ollut viime vuosina?
2. Millä foorumeilla tutkimuksia on julkaistu?
3. Minkä tyyppisiä tutkimukset ovat olleet?
4. Mitä koneoppimismenetelmiä tutkimuksissa on käytetty?
5. Mihin tarkoitukseen koneoppimismenetelmiä on käytetty?
6. Mitä dataa koneoppimismenetelmät ovat käyttäneet?

Ensimmäisen tutkimuskysymyksen sanalla aktiivisuus tarkoitetaan julkaisujen vuosittaista lukumäärää ja kehitystä. Toisen tutkimuskysymyksen sanalla foorumit tarkoitetaan julkaisupaikkatyyppejä (esim. konferenssit) ja tutkimusten julkaisijoita (esim. tietty vertaisarvioitu lehti). Kolmannen tutkimuskysymyksen sanalla tyyppi viitataan erilaisiin tutkimustyyppeihin, joita ovat esimerkiksi arviointitutkimus tai filosofinen julkaisu. Neljännen tutkimuskysymyksen sanalla koneoppimismenetelmät tarkoitetaan vain niitä menetelmiä, joita on käytetty vuotojen hallintaan vesijohtoverkostossa. Vuotojen hallinta on laaja käsite, minkä vuoksi viidennen tutkimuskysymyksen pyrkimyksenä on kartoittaa koneoppimismenetelmien tarkempaa käyttötarkoitusta. Kuudennen tutkimuskysymyksen sanalla data tarkoitetaan koneoppimismallin koulutuksessa sekä algoritmin testaamisessa käytettyä dataa.

4.3 Vaihe 2: Haun suorittaminen

Petersenin et al. [50] prosessin toisen vaiheen tarkoituksena on etsiä primääritutkimuksia, jotka käsittelevät tutkimuksen tutkimuskysymyksiä, tieteellisistä tietokannoista tai etsimällä artikkeleita manuaalisesti. Tietokanta- ja manuaalisen haun lisäksi on olemassa lumipallostrategia (engl. snowballing) [52]. Tässä strategiassa tutkimuskysymyksiä käsittelevien julkaisujen lähteitä läpikäymällä pyritään löytämään lisää julkaisuja, joiden lähteitä läpikäymällä taas yritetään löytää lisää julkaisuja ja niin edelleen. Petersenin et al. [52] mukaan yhtä oikeaa strategiaa ei ole olemassa, mutta aikataulusyistä monien strategioiden valitseminen yhteen tutkimukseen ei ole suositeltavaa.

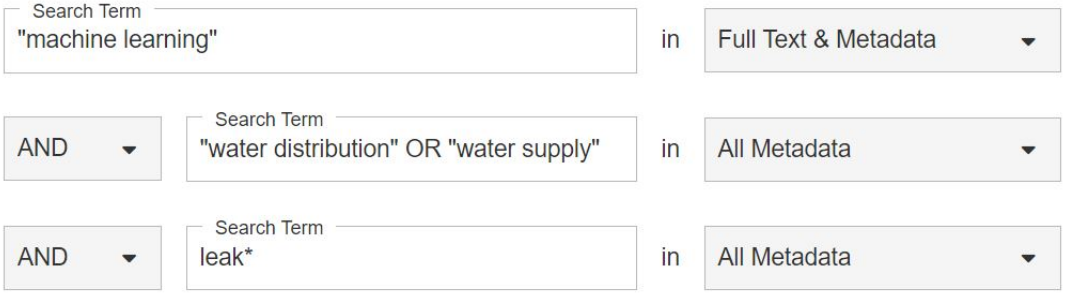
Tässä tutkimuksessa artikkelien etsimisstrategiaksi valittiin tietokantahaku, joka oli selkeästi suosituin etsimisstrategia ohjelmistoalalla Petersenin et al. [52] suorittaman tutkimuksen mukaan. Tutkimukseen päätettiin valita yhteensä neljä tutkimustietokantaa (IEEE Xplore, ACM Digital Library, ScienceDirect ja Springer) sekä kaksi tieteellisten julkaisujen haku- ja indeksointipalvelua (Web of Science ja Scopus). Valittujen tietokantojen lukumäärä on kattava, sillä se ylittää Petersenin et al. [52] suosituksen (IEEE ja ACM sekä kaksi indeksointipalvelua).

Artikkeleita tulee Petersen et al. [50] mukaan etsiä tietokannoissa käyttämällä tutkimuskysymysten pohjalta muodostettuja hakusanavariaatioita, joilla saataisiin tutkimusalueesta mahdollisimman laaja ja vääristymätön kuva. Tässä tutkielmassa hakusanavariaatiot muodostettiin tutkimuskysymyksen pohjalta ja suorittamalla testihakuja valittuihin tietokantoihin ja hakupalveluihin. Sopivia sanoja kartoitettiin myös tutkimalla muita samasta aiheesta tehtyjä kartoituksia, kuten Kitchenham et al. [32] suositteli. Myös Petersen et al. [52] suosittelemaa PICO-menetelmää kokeiltiin. PICO-menetelmässä avainsanoja muodostetaan tutkimuksen kohderyhmän (engl. population), intervention (engl. intervention), vertailumenetelmän (engl. comparison) ja tuloksien (engl. outcome) mukaan. Menetelmän ei kuitenkaan koettu tuovan lisähyötyä hakusanojen muodostamiseen, minkä vuoksi se jätettiin tutkimuksesta pois.

Testihakujen jälkeen avainsanoiksi muodostuivat lopulta ”machine learning” ja ”water distribution/water supply” ja ”leak/leakage”, jotka toistuivat lähes kaikissa tutkimuskysymyksiä käsittelevissä artikkeleissa. Poikkeuksen muodosti sana ”machine learning”, jota ei aina käytetty. Tämän sanan sijaan saatettiin puhua esimerkiksi syväoppimisesta (engl. deep learning) tai tukivektorikoneista (engl. support vector machines), jotka molemmat ovat kuitenkin koneoppimismenetelmiä. Jotta nämä tutkimukset olisi saatu mukaan, olisi hakusanavariaatioiden joukkoon pitänyt lisätä kyseiset sanat. Tämä olisi kuitenkin saattanut vääristää tutkielman tuloksia koskien käytettyjä koneoppimismenetelmiä. Vääristymien välttämiseksi päädyttiin käyttämään pelkästään sanaa ”machine learning”.

Hakutulosten määrään vaikutti paljon se, mistä osasta artikkelia hakusanoja etsittiin. Kun sanoja etsittiin kaikista osista artikkelia, oli hakutuloksia tuhansia, joista suurin osa ei vastannut tämän tutkielman tutkimuskysymyksiin. Petersen et al. [52] ohjeistivat, että mikäli hakutulokset sisältävät paljon epärelevanttia aineistoa, tulisi haun tarkempaa rajausta miettiä. Tämän vuoksi sanoja ”water distribution/water supply” ja ”leak/leakage” päätettiin hakea vain artikkelin metatiedoista (ml. tiivis-

telmä) ja sanaa ”machine learning” kaikista tekstiosioista, pois lukien lähteet. IEEE Xplore, ACM Digital Library, ScienceDirect ja Springer tarjosivat kuitenkin kaikki hieman erilaiset hakumahdollisuudet, minkä vuoksi edellä mainitun mukaisia hakuja ei voitu suorittaa kaikissa tietokannoissa. Esimerkiksi Springerissä hakusanoja pystyi kohdistamaan ainoastaan artikkelin kaikkiin osiin tai otsikkoon. IEEE Xplore taas mahdollisti paljon monipuolisemmat hakuvariaatiot. Esimerkki IEEE Xploren tarjoamasta hakukentästä on esitetty kuvassa 4.2. Lisäksi haku- ja indeksointipalvelut Scopus ja Web of Science toimivat siten, että ne etsivät sanoja vain artikkeleiden metatiedoista (ml. tiivistelmä). Tämän vuoksi sanaa ”machine learning” ei voitu etsiä kaikista tekstin osista. Myös näiden palveluiden hakukentissä oli eroja. Esimerkiksi Scopus käytti hyväkseen artikkeleiden lähteitä, minkä vuoksi haku piti rajata vain otsikkoon, abstraktiin ja avainsanoihin.



Kuva 4.2: Esimerkkinä IEEE Xploren tarjoamat hakukentät.

Taulukossa 4.1 on esitetty yhteenveto käytetyistä hakukentistä ja -sanoista eri tietokannoissa yhteneväisessä muodossa sekä hakutulosten määrä. Aineiston haku suoritettiin 25-27.5.2022 aikana, ja sen lopputuloksena oli yhteensä 270 artikkelia.

4.4 Vaihe 3: aineiston suodattaminen

Petersenin et al. [50] ohjeen kolmannessa vaiheessa määritellään sisällyttämis- ja poissulkemiskriteerit, joiden avulla rajataan pois ne primääritutkimukset, jotka eivät ole tutkimuskysymysten kannalta oleellisia. Petersenin et al. [52] mukaan kriteereihin voi kuulua esimerkiksi julkaisupaikka (engl. venue), tietty aikarajaus ja arviointimenetelmä (esim. vain vertaisarvioidut).

Taulukko 4.1: Käytetyt tietokannat ja hakutermit.

Tietokanta	Hakutermit	Tulokset
IEEE Xplore	(full text + metadata:"machine learning") AND ((all metadata:"water distribution") OR (all metadata:"water supply")) AND (all metadata:"leak*")	62
ACM Digital Library	(all: "machine learning") AND ((abs: "water distribution") OR (abs: "water supply")) AND (abs: "leak*")	22
ScienceDirect	(all: "machine learning") AND ((title-abs-key:"water distribution") OR (title-abs-key:"water supply")) AND ((title-abs-key:"leak") OR (title-abs-key:"leakage"))	45
Springer	(all:"machine learning") AND ((all:"water distribution") OR (all:"water supply")) AND (all:"leak*")	29
Web of Science	(all:"machine learning") AND ((all:"water distribution") OR (all:"water supply")) AND ((all:leak) OR (all:leakage))	41
Scopus	(title-abs-key: "machine learning") AND ((title-abs-key:"water distribution") OR (title-abs-key:"water supply")) AND (title-abs-key:"leak*")	71
Yhteensä:		270

Tutkimuksen sisällyttämiskriteerit olivat seuraavat:

- Kyseessä on vertaisarvioitu tieteellinen artikkeli.
- Artikkeli on julkaistu vuoden 2012 aikana tai sen jälkeen.
- Kyseessä on primääritutkimus.
- Artikkeli on julkaistu kokonaan englannin kielellä.
- Artikkeli on saatavilla ilmaiseksi ja kokonaisena Jyväskylän yliopiston opiskelijana.
- Artikkelista on saatavilla kaikki luokittelujärjestelmän kannalta oleelliset metatiedot eli otsikko, tekijät, julkaisuvuosi ja julkaisupaikka.
- Artikkeli on tämän tutkimuksen tutkimuskysymysten kannalta relevantti. Artikkelin tutkimuksen tulee käsitellä koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa.

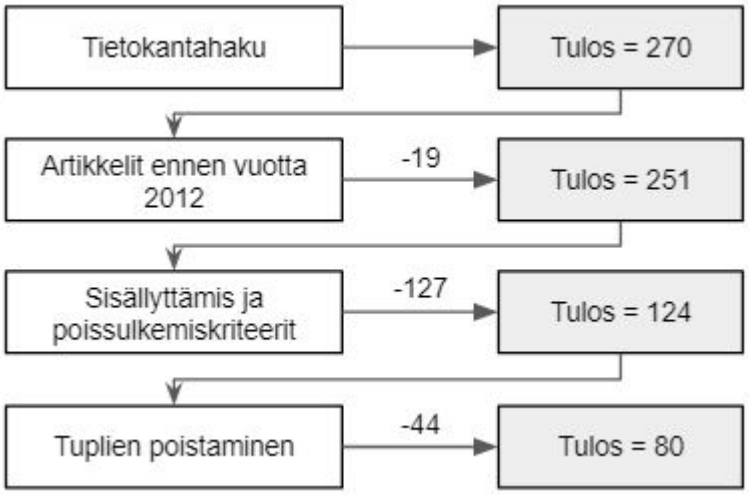
Tutkimuksen poissulkemiskriteerit olivat seuraavat:

- Kyseessä ei ole vertaisarvioitu tieteellinen artikkeli vaan esimerkiksi oppikirja, väitöskirja tai opinnäytetyö.
- Artikkeli on julkaistu ennen vuotta 2012.
- Kyseessä ei ole primääritutkimus, vaan esimerkiksi sekundäärinen tutkimus (mm. erilaiset katselmukset).
- Artikkelia ei ole julkaistu kokonaan englannin kielellä.
- Artikkeli ei ole saatavilla ilmaiseksi tai kokonaisena Jyväskylän yliopiston opiskelijana, vaan esimerkiksi pelkkä abstrakti on saatavilla.
- Artikkelista ei ole saatavilla luokittelujärjestelmän kannalta oleellisia metatietoja, kuten otsikko, tekijät, julkaisuvuosi ja julkaisupaikka.
- Artikkeli ei ole tämän tutkimuksen tutkimuskysymysten kannalta relevantti. Artikkelin tutkimus ei käsittele koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Esimerkiksi tutkimukset, jotka koskevat kastelujärjestelmiä tai vuotoja rakennusten sisällä rajataan pois.

Yläpuolella luetellut kriteerit muodostuivat osittain iteratiivisesti. Esimerkiksi ainoastaan vertaisarvioidut artikkelit päätettiin ottaa mukaan kesken suodattamisprosessin, koska hakutuloksissa oli mukana epäluotettavan oloisia julkaisuja. Tämä oli vastoin Petersenin et al. [52] suositusta välttää arviointimenetelmän mukaan tehtävää rajausta, jotta uusimmatkin artikkelit ja trendit saataisiin tutkimukseen mukaan. Toisaalta he tuovat myös esille, että artikkeleiden laadullista arvioimista ja rajausta voidaan tehdä, mutta vaatimuksien tulee olla vaatimattomat. Tässä tutkielmassa aineiston laatuun otettiin kantaa epäsuorasti valitsemalla mukaan vain vertaisarvioidut artikkelit. Artikkelin sisällön perusteella tehtävää laadullista arviointia ei kuitenkaan tehty, koska se ei yleisesti ole osa systemaattista kirjallisuuskartoitusta.

Koska tutkielmassa haluttiin selvittää erityisesti nykytutkimuksen tilaa ja suuntaa, ei yli kymmenen vuotta vanhojen artikkeleiden mukaan ottamista nähty järkevänä. Tämän vuoksi ennen vuotta 2012 julkaistut artikkelit päätettiin suodattaa pois. Systemaattinen kirjallisuuskartoitus tehdään aina primääritutkimuksista, ellei tutkimuksessa haluta tutkia esimerkiksi juuri sekundäärisiä tutkimuksia. Tämän vuoksi primääritutkimukset otettiin yhdeksi valintakriteeriksi. Englannin kieli valittiin valintakriteeriksi, koska se on yleisesti käytetty tutkimusten julkaisukieli. Yhtään artikkelia ei kuitenkaan tarvinnut suodattaa pois kielen perusteella. Sen sijaan useita artikkeleita jäi pois koska ne eivät olleet täysin saatavilla. Tämä oli oleellinen valintakriteeri, koska pelkkä otsikko tai abstrakti ei olisi aina riittänyt artikkelin tarkistamiseen eikä myöhemmin suoritettavan luokitusjärjestelmän tekemiseen. Haasteellisimman kriteerin asetti artikkelin relevanttius tutkielmalle eli sen tuli käsitellä koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Haasteellisuus johtui siitä, että tutkimuksien tavoitteet eivät aina käyneet selkeästi ilmi otsikosta tai abstraktista, minkä vuoksi koko teksti piti käydä läpi.

Aineiston suodatusprosessi on esitetty kuvassa 4.3. Ensimmäisenä tuloksista suodatettiin pois ennen vuotta 2012 julkaistut artikkelit, joita oli vain 19 kappaletta. Tämän jälkeen artikkelit suodatettiin käyttäen loppuja sisällyttämis- ja poissulkemiskriteerejä, jolloin noin puolet artikkeleista karsiutuivat pois. Yleisimpänä syynä artikkelien karsiutumiseen oli se, että ne eivät olleet täysin saatavilla tai eivät koskeneet koneoppimista. Viimeisessä vaiheessa eri tietokantojen tulokset yhdistettiin ja kahdesti esiintyneet artikkelit poistettiin. Suodattamisprosessin lopputuloksena oli yhteensä 80 artikkelia.



Kuva 4.3: Aineiston suodatusprosessi.

4.5 Vaihe 4: avainsanojen muodostaminen tiivistelmistä

Petersenin et al. [50] ohjeen neljännessä vaiheessa muodostetaan primääritutkimuksille luokittelujärjestelmä kaksivaiheisella avainsanojen tunnistamisprosessilla. Ensimmäisessä vaiheessa luetaan primääritutkimuksien tiivistelmät, joista etsitään tutkimuksen avainsanoja ja käsitteitä, jotka kuvaavat tutkimuksen myötävaikutuksia (engl. contribution). Mikäli tiivistelmä ei tarjoa tarvittavia tietoja, myös johdanto ja johtopäätökset voidaan lukea. Toisessa vaiheessa löydökset ryhmitellään ja ryhmien pohjalta muodostetaan tutkimuksen luokittelujärjestelmä.

Tässä tutkimuksessa avainsanoja ei etsitty vapaamuotoisesti vaan tutkimuskysymysten pohjalta. Esimerkiksi tutkimuskysymystä ”Mitä dataa koneoppimismenetelmät ovat käyttäneet?” koskien artikkeleista yritettiin poimia erilaisia dataan viittaavia sanoja tutkimuksen otsikosta ja tiivistelmästä. Mikäli näistä osioista tarvittavia avainsanoja ei löytynyt, luettiin artikkelin johtopäätökset. Mikäli johtopäätöksistä ei löydetty avainsanoja, käytiin artikkelin loput osiot läpi. Avainsanojen rajauksesta johtuen artikkelien kaikkiin osioihin jouduttiin turvautumaan hyvin usein. Kirjauksia ylläpidettiin Excel-taulukkolaskentaohjelmassa, jossa yksi rivi vastasi yhtä artikkelia ja tätä koskevia kirjauksia. Kun kaikki artikkelit olivat käyty läpi, avainsanoista muodostettiin ryhmiä ja ryhmät nimettiin. Ryhmän vaatimuksena oli, että siihen kuuluu vähintään kaksi avainsanaa. Tämän vuoksi esimerkiksi tietty koneoppimismenetelmä, joka mainittiin vain yhdessä tutkimuksessa jäi tulosten ulkopuolelle. On myös mainittava, että satunnaismetsä-koneoppimismenetelmä pidettiin

omana ryhmänä, vaikka se kuuluukin teoriassa ensemble luokittajat-koneoppimisryhmään. Käytetyn datan osalta joissain tutkimuksissa hyödynnettiin verkoston kiinteää dataa (esim. putkien paksuus), mutta tietojen hajanaisuudesta johtuen näitä tietoja ei päätetty ottaa mukaan luokitusjärjestelmään.

Seuraavassa on listattu tutkimuskysymykset 4-6 ja niitä vastaavat ryhmät:

- **TK 4: Mitä koneoppimismenetelmiä tutkimuksissa on käytetty?** Muodostetut ryhmät olivat seuraavat: 1) neuroverkkopohjaiset (engl. artificial neural networks) menetelmät, 2) tukivektorikone (engl. support vector machine), 3) *k*:n lähimmän naapurin menetelmä (engl. *k*-nearest neighbors), 4) satunnaismetsä (engl. random forest), 5) naiivi bayesin luokitin (engl. naive bayes classifier), 6) *k*:n keskiarvon klusterointi (engl. *k*-means clustering), 7) päätöspuu (engl. decision tree), 8) pääkomponenttianalyysi (engl. principal component analysis), 9) ensemble luokittelija (engl. ensemble classifier), 10) lineaarinen erotteluanalyysi (engl. linear discriminant analysis), 11) logistinen regressio (engl. logistic regression), 12) sumean *c*:n keskiarvon klusterointi (engl. fuzzy *c*-means clustering) ja 13) gaussin prosessiregressio (engl. gaussian process regression).

- **TK 5: Mihin tarkoitukseen koneoppimismenetelmiä on käytetty?** Muodostetut ryhmät olivat seuraavat: 1) vuodon tunnistaminen, 2) vuodon paikantaminen ja 3) kunnon arvioiminen.

- **TK 6: Mitä dataa koneoppimismenetelmät ovat käyttäneet?** Muodostetut ryhmät olivat seuraavat: 1) paine, virtaama, värinä ja ääni.

Kaksivaiheisen avainsanojen tunnistamisprosessin lisäksi tutkimuksessa käytettiin tutkimuksista riippumattomia luokkia, jotka vastasivat tutkimuskysymyksiä kaksi: ”millä foorumeilla tutkimuksia on julkaistu?” ja kolme: ”minkä tyyppisiä tutkimukset ovat olleet?”. Petersen et al. [52] suosittelivat riippumattomia luokkia muun muassa sen takia, että samasta aiheesta tehtyjen systemaattisten kirjallisuuskartoitusten tuloksien vertailua olisi helpompaa. Päivityksessä ehdotettuja luokkia olivat julkaisupaikka (engl. venue), tutkimuksen tyyppi ja tutkimusmenetelmä. Tässä tutkimuksessa päädyttiin ottamaan mukaan julkaisupaikka ja tutkimuksen tyyppi. Ryhmittely tutkimusmenetelmän mukaan jätettiin pois, koska tutkimukset saattoivat sisältää monia tutkimusmenetelmiä ja kaikkien niiden tunnistaminen luotettavasti olisi ollut liian työläs tehtävä tälle tutkimukselle. Julkaisupaikkojen luokitte-

lussa Petersen et al. [52] ehdottivat käytettäväksi Suomen opetusministeriön vuonna 2010 julkaisemaa julkaisutyyppiluokittelua [1]:

- Vertaisarvioidut tieteelliset artikkelit
- Vertaisarvioimattomat tieteelliset kirjoitukset
- Tieteelliset kirjat (monografiat)
- Ammattiyhteisölle suunnatut julkaisut
- Suurelle yleisölle suunnatut julkaisut
- Julkinen taiteellinen ja taideteollinen toiminta
- Opinnäytteet
- Patentit ja keksintöilmoitukset
- Audiovisuaaliset aineistot ja tieto- ja viestintätekniset ohjelmat

Koska tässä tutkimuksessa sisällyttämiskriteerinä oli vertaisarvioidut tutkimukset, jäi julkaisupaikoista jäljelle ainoastaan vertaisarvioidut tieteelliset artikkelit. Tämä ryhmä pitää sisällään artikkelit tieteellisissä lehdissä, katsausartikkelit tieteellisissä lehdissä, kirjat tai muun kokoomateoksen osat ja artikkelit konferenssijulkaisuissa. Koska poissulkemiskriteereillä katsaukset ja kirjat rajattiin pois, jäi jäljelle vain alkuperäisartikkelit tieteellisissä lehdissä sekä artikkelit konferenssijulkaisuissa. Näiden kahden ryhmän rinnalle valittiin vielä symposiumi, eli yhden päivän mittainen tieteellinen keskustelutilaisuus tai kokous, koska kyseinen ryhmä ei varsinaisesti kuulu kahtaan edellä mainittuun ryhmään. Seuraavassa on yhteenveto tutkimuksessa käytetyistä ryhmistä julkaisupaikan mukaan:

- Artikkeli tieteellisessä lehdessä
- Artikkeli konferenssijulkaisussa
- Artikkeli symposiumijulkaisussa

Tutkimustyypin osalta Petersen et al. [52] suosittelee käytettäväksi Wieringan et al. [64] ehdottamia ryhmiä. Nämä ryhmät, joita tässäkin tutkimuksessa päätettiin käyttää, ovat seuraavat:

- Arviointitutkimus (engl. Evaluation paper).
- Ratkaisuehdotus (engl. Solution proposal)
- Validointitutkimus (engl. Validation research)
- Filosofinen julkaisu (engl. Philosophical paper)
- Mielipidejulkaisu (engl. Opinion paper)
- Kokemusperäinen julkaisu (engl. Experience paper)

Taulukko 4.2: Tutkimustyypin valintaa helpottava taulukko.

Tutkimustyypit (avattu yläpuolella):	A	R	K	V	F	M
Ehdot:						
Sovelletaan käytännössä	P	-	P	E	E	E
Uusi ratkaisu	-	P	E	-	E	E
Empiirinen arviointi	P	E	E	P	E	E
Käsitteellinen viitekehys	-	-	-	-	P	E
Mielipide jostakin asiasta	E	E	E	E	E	P
Kirjoittajan mielipide	-	-	P	-	E	E

Edellä mainittujen ryhmien erottelun helpottamiseksi muodostettiin taulukko 4.2, joka pohjautuu Petersenin et al. [52] esittämään päätöspuuhun. Taulukossa käytetyt lyhenteet ovat seuraavat: A = Arviointitutkimus, R =Ratkaisuehdotus, K = Kokemusperäinen julkaisu, V = Validointitutkimus, F = Filosofinen julkaisu, M = Mielipidejulkaisu, P = Pätee, E = Ei päde ja Tyhjä = epäolennainen tai ei sovellettavissa.

Petersenin et al. [52] mukaan eniten vaikeuksia aiheuttaa ero arviointitutkimuksen ja validointitutkimuksen välillä. Heidän mukaansa oleellista näiden ryhmien välillä on se, että arviointitutkimus suoritetaan teollisessa ympäristössä, kun taas validointitutkimusta ei suorita käytännössä, vaan laboratoriossa. Tämä rajaus ei kuitenkaan osoittautunut riittävän selkeäksi, sillä osa tutkimuksesta saatettiin tehdä teollisessa ympäristössä, mutta valtaosa laboratoriossa tai siihen rinnastettavassa tilassa. Tämän vuoksi tässä tutkimuksessa määrityksiä tiukennettiin siten, että pelkkä datan keräys teollisesta ympäristöstä tai oikean vesijohtoverkon mallintaminen ei riittänyt täyttämään arviointitutkimuksen määritelmää.

4.6 Vaihe 5: tiedon erottelu ja kirjaaminen

Petersenin et al. [50] viidennessä vaiheessa valittujen primääritutkimuksien sisältö lajitellaan perusteluineen edellisessä vaiheessa muodostettuun luokittelujärjestelmään. Näin tehtiin myös tässä tutkimuksessa. Kaikki artikkelit käytiin uudelleen läpi ja jokaisen artikkelin osalta luokittelujärjestelmään merkittiin kaikki ne ryhmät, joihin artikkeli kuului. Petersenin et al. [50] mukaan tässä vaiheessa luokittelujärjestelmää on vielä mahdollista hioa, kuten lisätä tai yhdistää ryhmiä. Tässä tutkimuksessa tälle ei kuitenkaan syntynyt tarvetta, sillä lajittelussa ei havaittu puutteita, esteitä tai epäloogisuutta.

Kun kirjaukset luokitusjärjestelmään on tehty, muodostetaan niistä helposti luettavia ja visuaalisia kaavioita, joissa esitetään lukumääriä ja prosentuaalisia osuuksia [50]. Tämän vaiheen lopputulosta kutsutaan systemaattiseksi kuvaukseksi. Petersenin et al. [52] artikkelissa ehdotetaan esimerkiksi kupla-, pylväs- ja piiraskaavioiden käyttöä. Tämän tutkimuksen luokitusjärjestelmän mukaiset visuaaliset esitykset eli systemaattinen kuvaus selityksineen löytyvät luvusta 5.

4.7 Vaihe 6: validiteettia koskevat uhat ja tutkimuksen toistettavuus

Petersenin et al. [52] mukaan tutkimuksen validiteettiin (=pätevyyden suhde tavoitteisiin) vaikuttavista uhista tulisi keskustella tutkimuksessa. Artikkelin mukaan tutkimuksessa tulee ottaa huomioon kuvaileva validiteetti, teoreettinen validiteetti, yleistettävyys ja tulkinnallinen pätevyys. Joseph Maxwell [43] avaa edellä mainitut termit seuraavalla tavalla:

- Kuvailevalla validiteetilla tarkoitetaan tutkimushavaintojen kuvaamisen tarkkuutta ja objektiivisuutta.
- Teoreettinen validiteetti tarkoittaa nimensä mukaisesti teorian soveltuvuutta tutkittuun ilmiöön, ja se on täysin riippumaton itse tutkimustilanteesta. Petersen et al. [52] täsmentää, että teoreettisen pätevyyden määrittää tutkijoiden kyky saavuttaa se mitä he tutkimuksessa pyrkivät saavuttamaan.
- Yleistettävyys viittaa siihen kuinka hyvin tutkimuksen tiettyä tilannetta tai otosta voidaan yleistää koskemaan tutkimuksen sisä- ja ulkopuolelle jääviä asioita, joita ei tutkimuksessa suoraan tutkittu. Eli esimerkiksi kuinka hyvin

tiettyä ryhmää koskevan tutkimuksen tuloksia voitaisiin yleistää koskemaan ryhmään kuuluvia mutta tutkimukseen osallistumattomia yksilöitä, tai kuinka hyvin voimme yleistää tuloksia koskemaan tutkimuksen ulkopuolelle jääviä ryhmiä.

- Tulkinnallinen validiteetti tarkoittaa, miten hyvin asioita tulkitaan tutkimuksessa, esimerkiksi tapahtumia tai ihmisen käytöstä.

Kuvailevaan validiteettiin (1.uhka) liittyvät uhat arvioidaan yleisellä tasolla pieniksi, sillä kyseessä on kvantitatiivinen tutkimus. Tutkimushavaintojen tekeminen rajoittui artikkelien lukemiseen ja avainsanojen kirjaamiseen. Avainsanat kirjattiin luokittelujärjestelmään (Excel-taulukko) sellaisenaan kuin ne tekstissä oli, minkä vuoksi tutkijan objektiivisuus havaintojen kuvaamisessa jäi pieneksi.

Teoreettisen validiteetin (2. uhka) osalta systemaattinen kirjallisuuskartoitus menetelmänä tulisi sopia tutkimuksen tarkoitukseen nähden hyvin, sillä sen päätarkoituksena on jäsennellä tutkimuksia. Tutkimuksessa sovellettiin Petersenin et al. [50] ohjetta, jota on hyödynnetty lukuisissa systemaattisissa kirjallisuuskartoituksissa ohjelmistoalalla [52]. Koska koneoppiminen liittyy ohjelmointiin, tulisi valitun ohjeen sopia tutkimukseen hyvin. Ohjeen soveltamiseen käytännössä liittyi kuitenkin neljä haastetta, joista ensimmäinen oli hakustrategian valinta. Hakustrategiaksi valittiin tietokantahaku, joka valituilla hakusanoilla tuotti 270 osumaa. Jo testihakuja tehdessä huomattiin, että valituilla hakusanoilla kaikkia relevantteja artikkeleita ei saatu mukaan. Tämä johtui siitä, että joissain tutkimuksissa ei käytetty sanaa ”Machine learning”. Toisaalta hakusanojen laajentaminen eri koneoppimismenetelmillä olisi voinut vääristää tutkimuksen tuloksia siten, että hakusanoissa mukana olleet koneoppimismenetelmät olisivat olleet yliedustettuina. Tämän vuoksi osa relevanteista tutkimusta jäi tietoisesti pois tutkimuksesta. Tämä ei kuitenkaan ole uhka tutkimuksen validiteetille, sillä aineiston määrän sijaan oleellisempaa on saada tutkittavasta alueesta hyvä otos [65].

Toisena haasteena oli artikkelien sisällyttämis- ja poissulkemiskriteerien objektiivinen soveltaminen. Selkeiden kriteerien, kuten rajaus tutkimuksen julkaisuvuoden mukaan, kohdalla ongelmia ei ollut. Tässä tutkimuksessa yhtenä kriteerinä oli kuitenkin artikkelin relevanttius tutkimukselle, mikä oli haastavampi kriteeri. Kriteerinä oli, että artikkelin tuli käsitellä koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Tutkimuksia oli kuitenkin monenlaisia, ja muutamien artikkelien kohdalla kriteerin soveltaminen koettiin alttiiksi tutkijan subjektiivisel-

le näkemykselle ja tarkkuudelle. Mikäli yksittäisten artikkelien kohdalla olisi tehty valinta- tai hylkäämisvirhe, ei tällä olisi kuitenkaan ollut suurta vaikutusta tutkimuksen kokonaistuloksiin. Tämän vuoksi tätä haastetta ei nähdä merkittävä uhkana tutkimuksen validiteetille.

Kolmantena haasteena oli luokitusjärjestelmän muodostaminen tutkimuskysymysten ja yleisten luokkien pohjalta. Tutkimustulokset olisivat voineet olla kattavammat, mikäli luokkia olisi valittu enemmän tai avainsanoja olisi kirjattu ylös vain artikkelin myötävaikutusten näkökulmasta, kuten Petersen et al. [50] ohjeisti. Luokkien määrällä ei kuitenkaan ole vaikutusta tutkimuksen validiteettiin. Luokkien määrän sijaan suurimman uhan muodostivat avainsanojen tunnistaminen. Koska avainsanoja (esim. koneoppimiseen liittyvät sanat) jouduttiin etsimään usein koko artikkelista, on mahdollista, että kaikkia avainsanoja ei tunnistettu. Käytännössä avainsanojen puuttuminen saattaisi näkyä tuloksissa puuttuvina ryhmänä (esim. tietty koneoppismenetelmä). Tämä ei ole kuitenkaan uhkaa tutkimuksen validiteettia, sillä todennäköisesti vain yksittäisiä harvinaisempia avainsanoja, jotka esiintyi vain yhdessä artikkelissa, jäi pois.

Viimeisenä haasteena oli tiedon kirjaaminen luokitusjärjestelmään. Tähän liittyi sama haaste, kuin avainsanojen muodostamisessa eli suuren tekstiaineiston läpikäynti. Esimerkiksi tutkimustyypin tunnistaminen ei ollut aina mahdollista käymättä läpi koko artikkelia. Mitä enemmän aineistoa piti käydä läpi, sitä suuremmaksi kävi riski sille, että jotain oleellista jäi huomaamatta tai tulkittiin väärin. Yleisesti ottaen tiedon kirjaaminen oli kuitenkin hyvin suoraviivaista ja nopeaa. Tutkimuksen validiteetin kannalta tätä haastetta ei nähdä uhkana.

Sisäisen yleistettävyyden (3. uhka) osalta tulisi arvioida kuinka hyvin tulokset vastaisivat artikkeleita, jotka kuuluivat tutkimuksen piiriin mutta jäivät kuitenkin tutkimuksen ulkopuolelle (= otoksen edustettavuus). Mukaan otettujen artikkelien suuren määrän (80 kpl) perusteella voidaan arvioida, että sisäinen yleistettävyys on korkea. Vaikka mukaan olisi otettu enemmän artikkeleita, tulokset tuskin olisivat muuttuneet merkittävästi. Luotettavan arvion saamiseksi otoksen edustettavuutta tulisi kuitenkin laskea statistisin menetelmin (esim. otoskoon laskeminen). Ulkoisen yleistettävyyden osalta voidaan arvioida, että saadut tulokset vastaavat tutkimuksia, joissa on keskitytty paineistettuihin putkiverkostoihin, koneoppimiseen ja vuotoihin. Esimerkiksi koneoppimismenetelmien näkökulmasta kaukolämpöverkostojen vuodot tuskin eroavat merkittävästi vesijohtoverkostojen vuodoista. Eroja on kuitenkin tietysti esimerkiksi käytetyn datan osalta.

Tulkinnallinen pätevyys (4. uhka) koskee tässä tutkimuksessa lähinnä tulosten tulkintaa. Tulkinnalliseen pätevyyteen liittyvät uhat ovat matalat sillä tutkimuksen tulokset (mm. luokat, ryhmät ja lukumäärät) ovat esitetty sellaisenaan tekstimuodossa ja yksinkertaisina visuaalisina esityksiä. Tuloksien pohjalta tehdyt johtopäätökset ja pohdinta on eriytetty lukuun 6.

Petersenin & Gencelin [51] mukaan edellä mainitut neljä uhkaa vaikuttavat yhdessä siihen, kuinka hyvin tutkimus on toistettavissa. Toistettavuutta parantaa myös tutkimuksen eri vaiheiden huolellinen dokumentointi, jota Petersen et al. [52] mukaan tulisi myös arvioida. Tässä tutkimuksessa dokumentointiin on kiinnitetty erityistä huomiota. Kaikki tarvittava tieto tutkimuksen arvioimiseksi tai toistamiseksi löytyy luvuista 4 (sis. sovellettu prosessia ja eri vaiheiden suoritus), 5 (sis. tulokset) sekä liitteestä B (sis. luokitusjärjestelmä).

5 Tulokset

Tässä luvussa esitellään systemaattisen kirjallisuuskartoituksen tulokset eli systemaattinen kuvaus helposti tulkittavassa ja visuaalisessa muodossa. Tutkijan subjektiivisuuden välttämiseksi tulokset on pyritty esittämään sellaisenaan ilman syvempää analysointia. Tuloksien pohjalta tehdyn johtopäätökset ja pohdinnat ovat jätetty lukuun 6. Luokittelujärjestelmä, johon tulokset perustuvat, löytyy tutkielman liitteistä B.

5.1 Tutkimuksen aktiivisuus

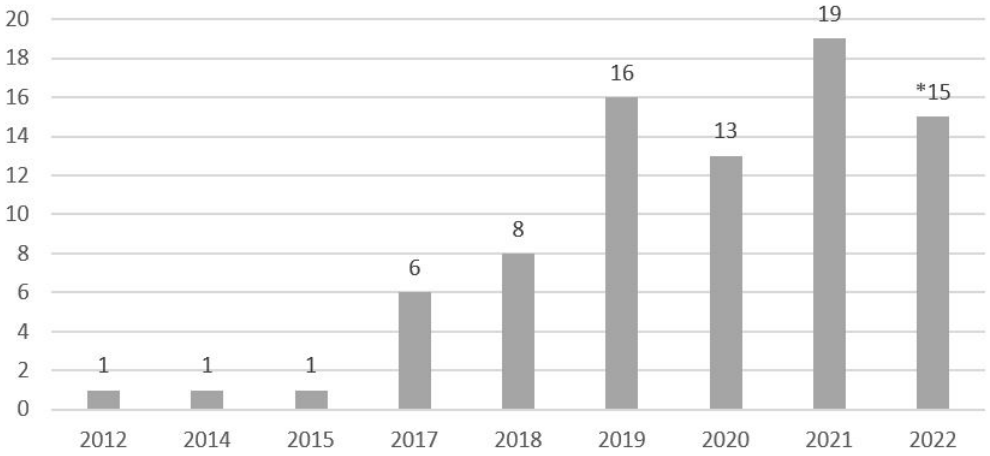
TK 1: Kuinka aktiivista aiheen tutkiminen on ollut viime vuosina?

Vuosien 2012-2022 (toukokuuhun asti) aikana julkaistiin yhteensä 80 tutkimusta. Kuten kuvasta 5.1 nähdään, vuosien 2012-2016 aikana julkaisutahti oli vähäistä. Kyseisten viiden vuoden aikana julkaistiin ainoastaan kolme tutkimusta. Ensimmäisen tutkimuksen julkaisivat Arsene et al. [8] vuonna 2012. Tämä tutkimus koski neuroverkkoihin ja verkkoteoriaan perustuvaa päätöksentekojärjestelmää vesijohtoverkoston vuotojen tunnistamiseksi. 2017 lähtien julkaisujen määrä on kuitenkin kasvanut vuosi vuodelta vuoden 2020 pientä notkahdusta lukuun ottamatta. Julkaisujen määrä per kuukausi on ollut korkein vuonna 2022 (3/kk).

5.2 Julkaisufoorumit

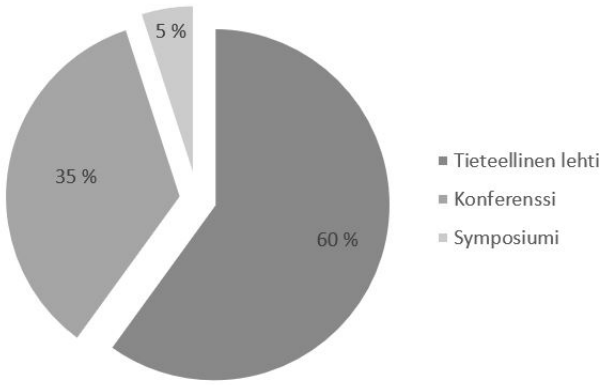
TK 2: Millä foorumeilla tutkimuksia on julkaistu?

Kuvasta 5.2 nähdään, että yleisimmät julkaisutyypit olivat tieteelliset lehdet (60 %, 48 kpl) sekä konferenssit (35 %, 28 kpl). Myös symposiumeja oli mukana neljä kappaletta (5 %). Muita julkaisutyyppejä ei sisällyttämis- ja poissulkemiskriteerien johdosta ollut mukana. Eri julkaisupaikkoja oli yhteensä 56 kappaletta, mikä tarkoittaa keskimäärin 1,4 tutkimusta per julkaisupaikka. Eniten tutkimuksia julkaisi Water-lehti (7 kpl) ja IFAC-PapersOnLine (5 kpl), mikä näkyy kuvassa 5.3. 43 jul-

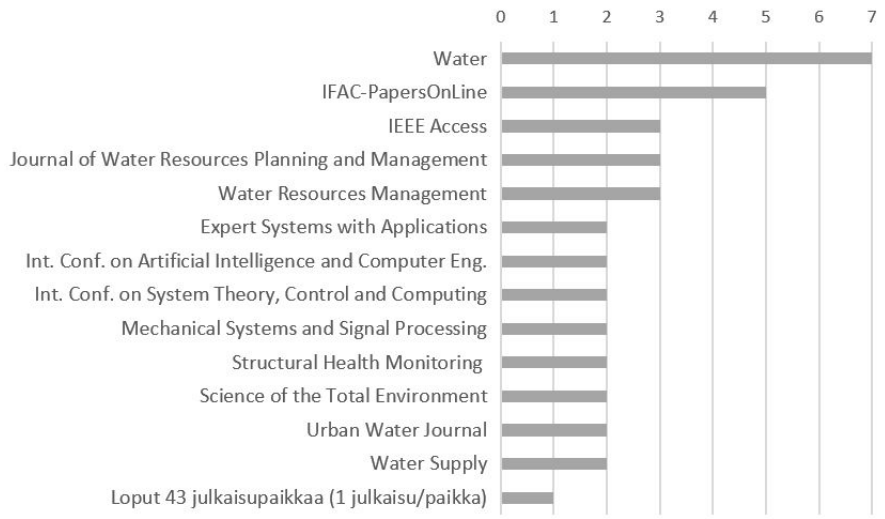


Kuva 5.1: Tutkimusten vuosittainen lukumäärä. *Mukana 5/2022 mennessä julkaistut artikkelit.

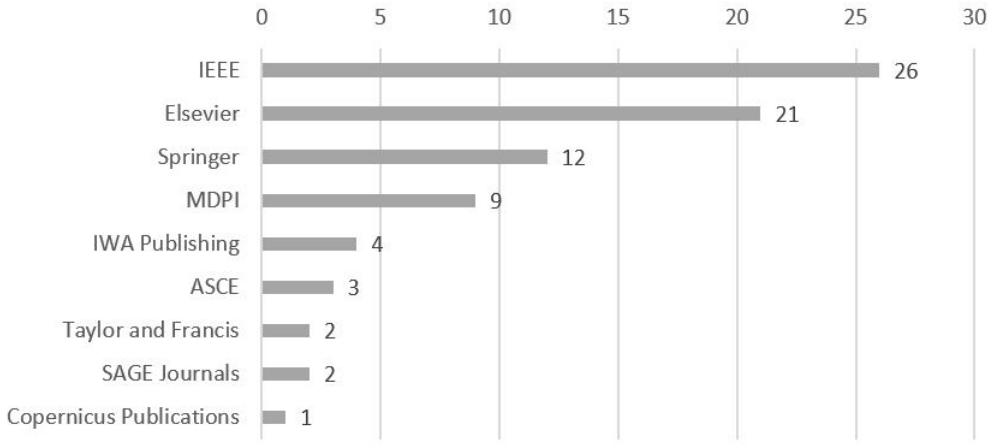
kaisupaikassa oli julkaistu vain yksi tutkimus. Tutkimusten julkaisijoita verkossa oli yhteensä yhdeksän kappaletta, mikä näkyy kuvassa 5.4. Pelkästään IEEE (26 kpl) ja Elsevier (21 kpl) yhdessä julkaisivat yli 50 prosenttia tutkimuksista verkossa. Kolmanneksi eniten tutkimuksia julkaisi MDPI (12 kpl).



Kuva 5.2: Julkaisutyypit sekä niiden prosentuaaliset osuudet.



Kuva 5.3: Tutkimusten julkaisupaikat sekä niiden lukumäärät.

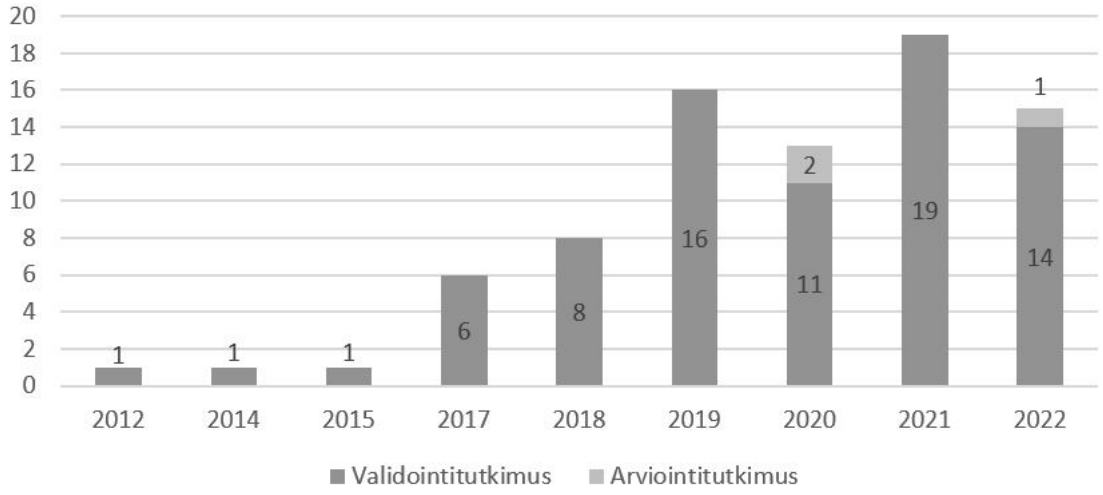


Kuva 5.4: Tutkimusten julkaisijat verkossa sekä niiden lukumäärät.

5.3 Tutkimusten tyyppi

TK 3: Minkä tyyppisiä tutkimukset ovat olleet?

Kuvasta 5.5 nähdään, että lähes kaikki tutkimukset olivat validointitutkimuksia. Validointitutkimuksien lisäksi tunnistettiin ainoastaan kolme arviointitutkimusta eli tutkimusta, jossa ehdotettua menetelmää testattiin teollisessa ympäristössä. Nämä kolme tutkimusta oli julkaistu vuosina 2020 ja 2022. Muita tutkimustyyppejä ei tutkimuksessa tunnistettu. Tyypillisessä validointitutkimuksessa mallinnettiin ensin oikea tai keksitty vesijohtoverkko jollakin simulointiohjelmistolla (useimmiten Epanetilla). Veden virtauksen simulointiin käytettiin joko keksittyä tai oikeaa dataa. Tämän jälkeen eri koneoppimismalleja opetettiin tunnistamaan vuotoja ja niiden sijainteja simulaatiomallin avulla. Lopuksi koneoppimisalgoritmien toimintaa testattiin ja vertailtiin keskenään.

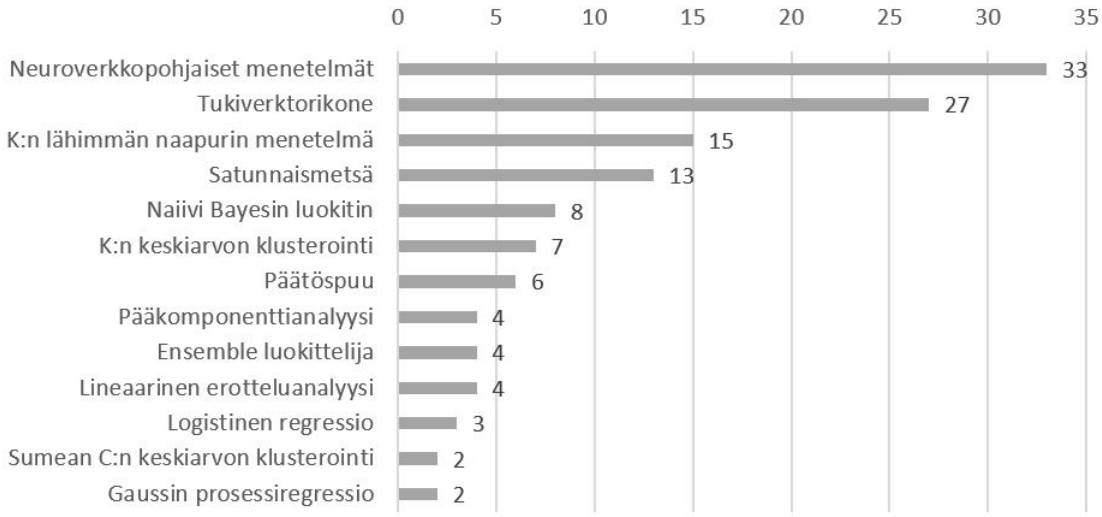


Kuva 5.5: Tutkimustyypit sekä niiden lukumäärä vuosittain.

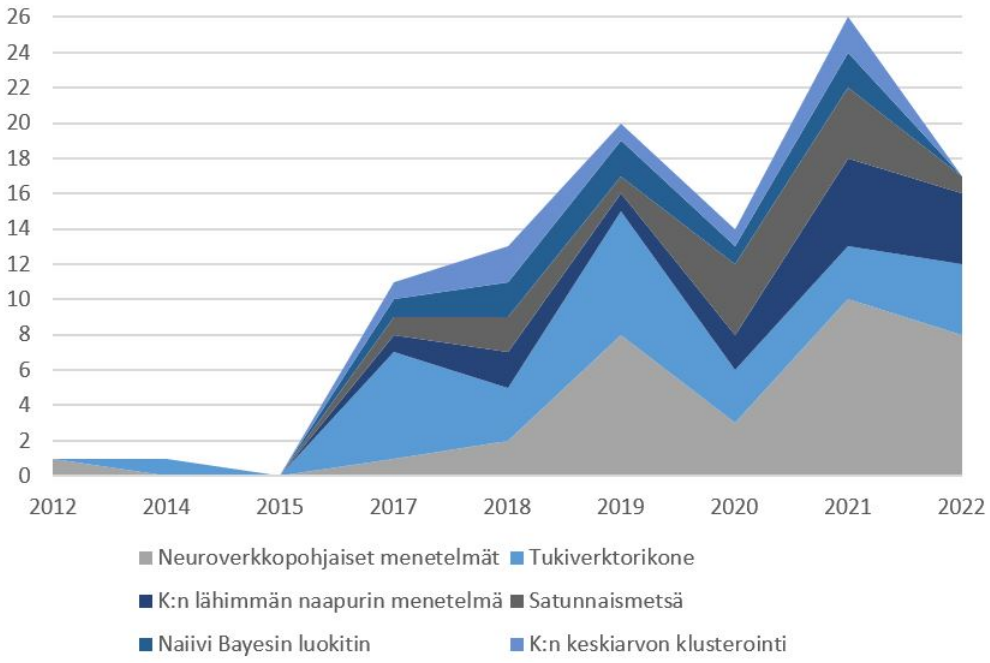
5.4 Koneoppimismenetelmät

TK 4: Mitä koneoppimismenetelmiä tutkimuksissa on käytetty?

Tutkimuksessa tunnistettiin yhteensä kolmetoista koneoppimismenetelmää. Tuloksia lukiessa on hyvä huomioida, että yhdessä tutkimuksessa saattoi olla käytössä monta eri menetelmää. Tutkimuksessa saatettiin esimerkiksi vertailla eri menetelmien toimivuutta. Tämän vuoksi menetelmien lukumäärä kuvassa 5.6 ylittää artikkelien lukumäärän. Neuroverkkopohjaiset menetelmät sekä tukivektorikoneet olivat selkeästi suosituimpia koneoppimismenetelmiä. Ne kattoivat 47 % kaikista käytetyistä menetelmistä. 70 % (9 kpl) menetelmistä oli ohjattua koneoppimista (6 kpl luokittamista ja 3 kpl regressiota). Myös neljä ohjaamattoman koneoppimisen menetelmää tunnistettiin: *K*:n keskiarvon klusterointi, sumean C:n keskiarvon klusterointi, pääkomponenttianalyysi ja lineaarinen erotteluanalyysi. Kahta jälkimmäistä käytetään pääsääntöisesti dimensioiden vähentämiseen. Kuvasta 5.7 nähdään kuuden suosituimman menetelmän vuosittaiset lukumäärät. Lukumäärien perusteella menetelmien käyttö on ollut melko tasaista lukuun ottamatta neuroverkkopohjaiset menetelmiä, joiden käyttö on selkeästi kasvanut. Vuonna 2022 kolme suosituinta menetelmää olivat neuroverkkopohjaiset menetelmät, tukivektorikoneet ja *K*:n lähimmän naapurin menetelmä.



Kuva 5.6: Tutkimuksissa käytetyt koneoppimismenetelmät sekä niiden lukumäärät.

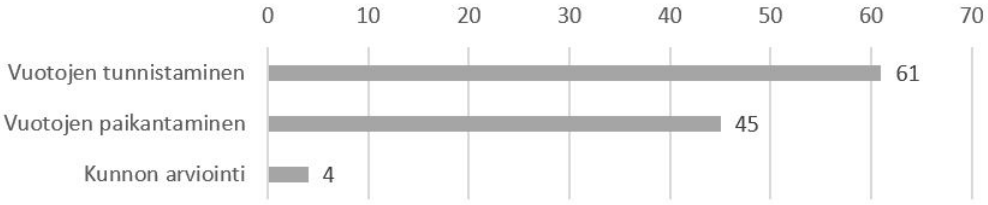


Kuva 5.7: Kuuden suosituimman koneoppimismenetelmän lukumäärä vuosittain.

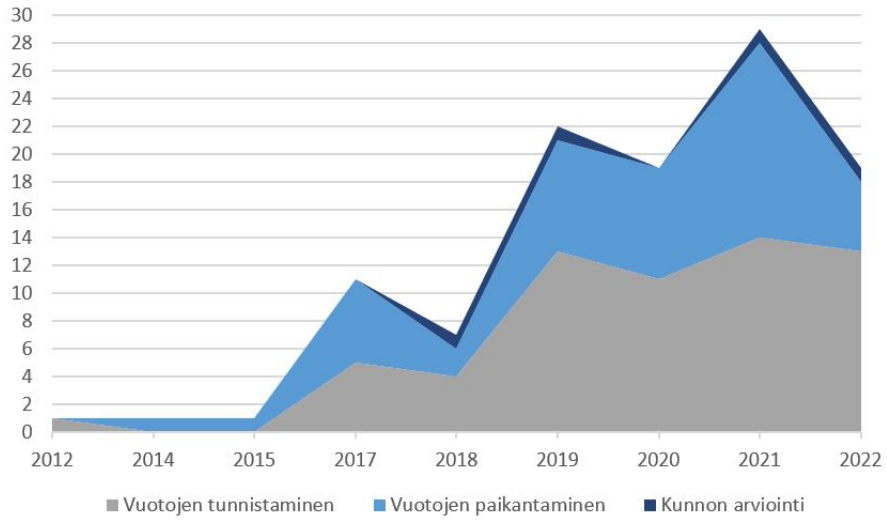
5.5 Menetelmien tarkoitus

TK 5: Mihin tarkoitukseen koneoppimismenetelmiä on käytetty?

Kuvasta 5.8 nähdään, että koneoppimismenetelmiä käytettiin kolmeen eri tarkoitukseen: vuotojen tunnistamiseen, vuotojen paikantamiseen ja verkoston kunnon arvioimiseen. Suosituin tarkoitus oli vuotojen tunnistaminen, jolla tarkoitetaan vuodon olemassaolon tiedostamista. Vuotojen paikantamisella tarkoittaa taas vuotojen paikantamista verkostossa. 33 tutkimuksessa tarkoituksena oli sekä vuotojen tunnistaminen että paikantaminen. Ainoastaan neljä tutkimusta keskittyi verkoston kunnon arviointiin. Kuvassa 5.9 on esitetty käyttötarkoitusten vuosittainen kehitys. Kuvan mukaan vuotojen tunnistaminen on kasvanut suhteellisen tasaisesti vuosi vuodelta. Myös vuotojen paikantaminen on ollut kasvamaan päin vuosta 2015 lähtien.



Kuva 5.8: Koneoppimismenetelmien käyttötarkoitukset sekä niiden lukumäärät.

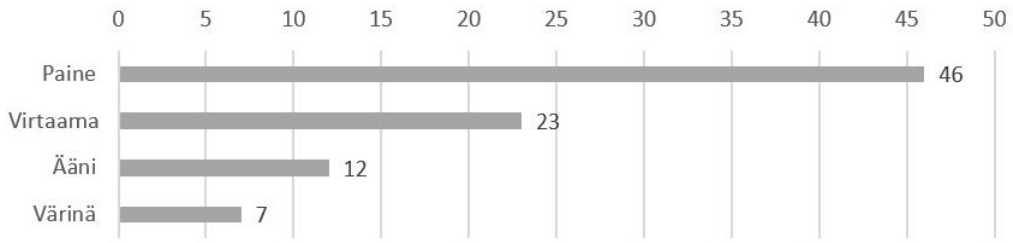


Kuva 5.9: Koneoppimismenetelmien käyttötarkoitusten lukumäärä vuosittain.

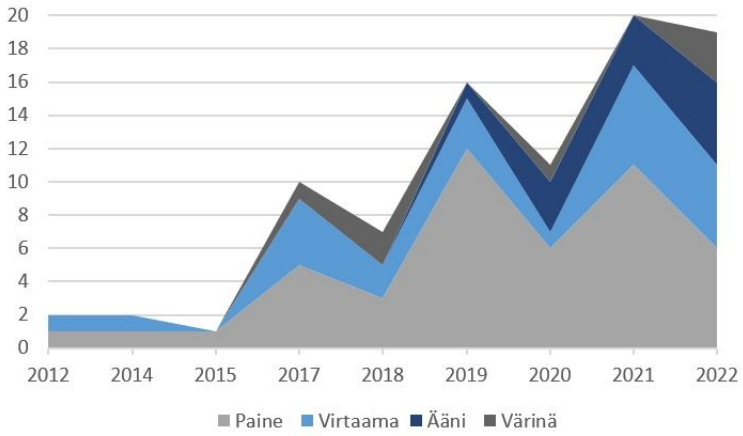
5.6 Menetelmien käyttämä data

TK 6: Mitä dataa koneoppimismenetelmät ovat käyttäneet?

Koneoppimismenetelmät hyödynsivät tutkimuksissa neljää erilaista lähtödataa (paine, virtaama, ääni ja värinä), kuten kuvasta 5.10 käy ilmi. Näiden lähtötietojen käyttö johtui siitä, että vesivuodon sattuessa putken paine-, virtaus-, ääni- ja tärinätiedot muuttuvat. Painedataa käytettiin selvästi eniten (46 tutkimusta). Samassa tutkimuksessa saatettiin hyödyntää myös useampaa dataa. Suosituin yhdistelmä oli paine- ja virtaamadata, jota käytettiin yhteensä 14 tutkimuksessa. Kuvassa 5.11 näkyy menetelmien käyttämä data vuosittain. Kuvan perusteella datojen käytössä on ollut merkittävää vuosittaista vaihtelua. On myös huomionarvoista, että vuonna 2022 virtaama- (5 kpl) ja äänidataa (5 kpl) käytettiin jo lähes yhtä paljon kuin painedataa (6 kpl).



Kuva 5.10: Koneoppimismenetelmien käyttämä data.



Kuva 5.11: Koneoppimismenetelmien käyttämä data vuosittain.

6 Johtopäätökset ja pohdinta

Tässä tutkielmassa suoritettiin systemaattinen kirjallisuuskartoitus koskien koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa. Luvussa 5 esiteltyjen tulosten pohjalta voidaan tehdä seuraavat kuusi johtopäätöstä:

1. Julkaisumäärien perusteella koneoppimisen hyödyntämisen tutkiminen vesijohtoverkostojen vuotojen hallinnassa on ollut viime vuodet aktiivista ja kasvamaan päin. Vuosien 2012-2016 aikana tutkimuksia julkaistiin vain kolme kappaletta, ja maksimissaan yksi per vuosi. Vuosien 2017-2021 aikana tutkimuksia julkaistiin kuitenkin jo keskimäärin 12 kappaletta vuodessa. Nopean kasvun taustalla on todennäköisesti yleisen kiinnostuksen kasvu koskien koneoppimista, mitä kuva 2.2 havainnollistaa. Kartoitus suoritettiin vuoden 2022 toukokuussa, minkä vuoksi kyseinen vuosi jäi vajaaksi. Mikäli julkaisutahti pysyy loppuvuoden samana kuin alkuvuosi, vuonna 2022 julkaistaan arviolta 36 tutkimusta. Mikäli tämä arvio pitää paikkansa, on aiheeseen liittyvä tutkimus aktiivisimmillaan kymmeneen vuoteen.

2. Aihe kiinnostaa tutkijoita laajalla rintamalla. Tämä perustuu siihen, että tutkimukset (80 kpl) julkaistiin jopa 56:ssa eri julkaisupaikassa. Näistä 43 oli julkaissut vain yhden tutkimuksen. Eniten tutkimuksia julkaistiin Water nimisessä tieteellisessä lehdessä (7 kpl).

3. Koneoppimisen tutkiminen vesijohtoverkoston vuotojen hallinnassa on vielä kovin teoreettista. Tämä perustuu siihen, että tutkimuksista 77 kappaletta oli validointitutkimuksia eli tutkimuksia, joita ei sovelleta teollisessa ympäristössä, vaan laboratoriossa tai siihen vastaavassa tilassa. Ainoastaan kolme tutkimusta vuosilta 2020 ja 2022 luokiteltiin arviointitutkimuksiksi. Tutkimustyypin osalta on siis tapahtunut vain vähän muutosta koko tarkastelujakson aikana. Tulos viittaa siihen, että aihetta koskevassa tutkimuskentässä on aukkoja muun tyyppisissä tutkimuksissa kuin validointitutkimuksissa.

4. Tutkijoita kiinnostaa erityisesti neuroverkkopohjaiset koneoppimismenetelmät. Tämä perustuu siihen, että kaikista tutkimuksista yhteensä 33:ssä (41 %) käytettiin neuroverkkopohjaisia menetelmiä. Toisena oli tukivektorikoneet, joita

käytettiin 27:ssä (34 %) tutkimuksessa. Vuosien 2022 ja 2021 aikana neuroverkkopohjaisten menetelmien osuus oli jo 53 %. Tukivektorikoneiden osuudet olivat vastaavina vuosina 27 % ja 16 %).

5. Koneoppimista halutaan käyttää sekä vuotojen tunnistamiseen että paikantamiseen. Tämä perustuu siihen, että tutkimuksista 61 (76 %) koski vuotojen tunnistamista ja 45 (56 %) paikantamista. 33 (41 %) tutkimusta koski sekä tunnistamista ja paikantamista. Ainostaan neljä tutkimusta koski putkien kunnon arviointia.

6. Vuotojen hallintaan käytetään erityisesti putkistojen painetietoa. Tämä perustuu siihen, että tutkimuksista 46 (58 %) käytti hyväkseen painetietoa. Myös virtaama- (29 %), ääni- (15 %) ja värinädataa (9 %) käytettiin tutkimuksissa. Painetiedon käytön yleisyyden syynä saattaa olla se, että paineen lasku putkessa mahdollisesti indikoi todennäköisemmin ongelmasta, kuin esimerkiksi Virtaaman kasvu.

Tutkimuksen tulosten 5 sekä edellä mainittujen johtopäätösten perusteella vesitoimialan organisaatioiden kannattaa pitää koneoppimista erittäin potentiaalisena työkaluna vesijohtoverkoston vuotojen hallinnassa. Koska tutkimustoiminta on vielä melko teoreettisella tasolla, on vesilaitoksilla ja erityisesti alan teknologiayrityksillä juuri nyt sopiva hetki perehtyä aiheeseen. Omien pilottihankkeiden toteuttaminen on yksi hyvä tapa lisätä tietämystä aiheesta ja kasvattaa kyvykkyyttä koneoppimiseen perustuvien vuodonhallintajärjestelmien toteuttamiseen. Erimerkiksi yhteishankkeet yliopistojen kanssa ovat yksi keino toteuttaa pilottihankkeita matalalla kynnyksellä. Myös yliopistojen tulisi olla kiinnostuneita aiheesta, sillä vesijohtoverkoston vuodot ovat merkittävä ongelma ja tutkimuskentästä löytyy aukkoja ainakin tutkimustyyppien (mm. arviointitutkimusten) osalta.

Ratkaisujen soveltaminen käytännössä sisältää kuitenkin monia haasteita. Jotta koneoppimisalgoritmit voivat havaita vuotoja, tarvitaan dataa. Simulaatioympäristössä esimerkiksi verkoston mittauspisteiden lisääminen ei ole ongelma, mutta todellisuudessa sensoreiden lisääminen on kallista. Putkiin asennettavat sensorit saattavat olla myös alttiimpia erilaisille häiriötekijöille kuten liikenteen aiheuttamalla tärinälle. Myös luotettavien verkkoyhteyksien rakentamiseen voi liittyä haasteita, koska putket kulkevat usein maan alla. Edellä mainitut haasteet saattavat olla osa syy siihen, miksi tutkimukset tehtiin pääosin laboratorioympäristössä. Haasteista

huolimatta koneoppimisen mahdollisuuksiin kannattaa tarttua vesitoimialalla. Esimerkiksi ilmastonmuutos ja hupenevat vesivarat lisäävät todennäköisesti uusien ratkaisujen kysyntää.

Tämän tutkimuksen luvussa 4.7 arvioitiin tutkimuksen validiteettia koskevia uhkia ja toistettavuutta. Arvion perusteella merkittäviä validiteettia koskevia uhkia ei tunnistettu, ja yksityiskohtaisen dokumentaation ansiosta tutkimuksen tulisi olla hyvin toistettavissa. Tästä huolimatta täysin indenttisiin tuloksiin vastaava tutkimus tuskin päätyisi. Tämä johtuu muutamasta haasteesta, joita kartoituksen aikana kohdattiin. Ensimmäisenä haasteena oli tutkimusten kattavuus, johon vaikutti erityisesti hakustrategiaksi valittu tietokantahaku. Pelkästään tietokantahaun avulla, jota rajoittavat tietyt hakusanat, kaikkia relevantteja tutkimuksia oli vaikea saada mukaan. Yksi ratkaisu tähän olisi ollut hakustrategian laajentaminen lumipallostrategialla, jossa aineistoa täydennetään käymällä läpi hakutulosten lähteitä. Lumipallostrategialla voitaisiin varmistua myös paremmin haun kattavuudesta, sillä jossain vaiheessa uusia artikkeleita ei enää pitäisi löytyä. Lumipallostrategian mukaan ottaminen olisi ollut kuitenkin liian työläs tehtävä, eikä Petersenin et al. [52] systemaattisen kirjallisuuskartoituksen ohjeessa suositeltu monien strategioiden käyttöä.

Toisena haasteena oli artikkeleiden valintakriteerien objektiivinen soveltaminen. Suurin osa kriteereistä oli hyvin selkeitä, kuten artikkelien rajaaminen julkaisuvuoden mukaan. Yksi kriteeri, jonka mukaan tutkimuksen tuli käsitellä koneoppimisen hyödyntämistä vesijohtoverkostojen vuotojen hallinnassa, jätti kuitenkin tilaa tutkijan omalle tulkinnalle. Petersen et al. [52] ehdotti, että valittujen kriteerien toimivuutta voidaan arvioida tekemällä testirajauksia monen tutkijan toimesta ja vertailemalla tuloksia keskenään. Tämä olisi ollut hyvä ratkaisu tähän ongelmaan, mutta tässä tutkimuksessa monen tutkijan käyttäminen ei ollut mahdollista. Toinen tapa, joka olisi myös ollut mahdollinen, olisi ollut jakaa kriteerit pienempiin ja tarkempiin osiin.

Kolmantena haasteena oli tutkimuskysymysten pohjalta etsittyjen avainsanojen tunnistaminen. Avainsanoja ei aina löydetty artikkelin otsikosta, tiivistelmästä tai johtopäätöksistä, jolloin jouduttiin turvautumaan kaikkiin tekstiosioihin. Avainsanojen etsiminen suuresta tekstimassasta oli työlästä ja lisäsi virheiden todennäköisyyttä. Tutkimuksen toistettavuuden kannalta olisikin parempi jättää tutkimuksen varsinainen tekstiosio kokonaan väliin, kuten Petersen et al. [50] ohjeisti. Lisäksi avainsanoista muodostetun luokitusjärjestelmän tarkastamisessa olisi voitu konsultoida alan asiantuntijaa, mitä Petersen et al. [52] ehdotti. Aikataulusyistä ulkopuoli-

sen asiantuntijan käyttö ei tässä tutkimuksessa ei onnistunut.

Neljäntenä haasteena oli artikkelien tietojen kirjaaminen luokittelujärjestelmään, mikä myös jätti tilaa tutkijan subjektiivisuudelle. Esimerkiksi harvinaisempien koneoppimismenetelmien kirjaaminen oikeisiin ryhmiin oli usein haastavaa ja vaati paikoin lisäselvityksiä. Kirjaamisvaiheeseen Petersen et al. [52] ehdotti kahden tutkijan käyttöä. Kahden tutkijan mallissa toinen tutkija tarkistaa lopputuloksen tai suorittaa koko kirjaamisvaiheen itse. Valitettavasti tässä tutkimuksessa monen tutkijan käyttäminen ei ollut mahdollista, kuten jo aiemmin todettiin.

Edellä mainituista haasteista huolimatta systemaattinen kirjallisuuskartoitus koettiin toimivaksi ja luotettavaksi tutkimusmenetelmäksi. Selkeimpänä tunnistettuna keinona nostaa tutkimuksen validiteettia ja toistettavuutta on käyttää tutkimuksessa vähintään kahta tutkijaa. Toisena selkeänä keinona on ulkopuolisen asiantuntijan hyödyntäminen. Ulkopuolinen asiantuntija voisi muun muassa ottaa kantaa artikkelien kattavuuteen ja sekä luokittelujärjestelmään. Lisäksi on selvää, että mitä enemmän kirjallisuustutkimuksia tutkija on tehnyt, sitä sujuvampaa ja laadukkaampaa tutkimuksen tekeminen on.

Tätä tutkimusta olisi luonnollista jatkaa muun muassa systemaattisella kirjallisuuskatsauksella, joka tekisi tutkimuksen aineistosta (liite B) synteesin. Synteesin muodostaminen ei ole osa systemaattista kirjallisuuskartoitusta, minkä vuoksi menetelmäksi tarvitaan juuri systemaattinen kirjallisuuskatsaus. Synteesi voisi vastata esimerkiksi seuraavaan kysymykseen: millainen koneoppimiseen perustuva järjestelmä pystyy tehokkaimmin tunnistamaan ja paikantamaan vesijohtoverkoston vuodot? Lisäksi, koska tutkimuskentässä havaittiin aukkoja tutkimustyyppien osalta, aiheesta kannattaisi tehdä muitakin kuin validointitutkimuksia.

7 Yhteenveto

Tässä pro gradu -tutkielmassa suoritettiin systemaattinen kirjallisuuskartoitus koskien koneoppimisen hyödyntämistä vesijohtoverkoston vuotojen hallinnassa. Vesijohtoverkostojen vuodot ovat vesilaitoksilla haastava ja kallis ongelma, jonka ratkaisemisessa koneoppiminen voi mahdollisesti auttaa. Koneoppimisella tarkoitetaan tietokoneiden ohjelmointia siten, että tietokone kykenee optimoimaan suoritustaan esimerkkidatan tai aiemmin kokemansa avulla. Koneoppimista voidaan käyttää esimerkiksi silloin, kun tietokoneen ohjelmointi ei ole mahdollista perinteisellä tavalla eli kirjoittamalla tietokoneelle tarkat ohjeet, miten suorittaa jokin tehtävä. Yleinen kiinnostus koneoppimista kohtaan on tällä hetkellä suurta ja erilaisia koneoppimista hyödyntäviä sovelluksia syntyy jatkuvasti.

Tutkimuksen päällimmäisenä tarkoituksena oli jakaa vesitoimialan organisaatioille tietoa aiheeseen liittyvän tutkimustoiminnan laajuudesta, suunnasta, luonteesta ja sisällöstä. Systemaattinen kirjallisuuskartoitus sopi tähän tarkoitukseen hyvin, sillä sen ensisijaisena tarkoituksena on juuri tutkimusalueen jäsentäminen. Tutkimuksessa saatiin selville, että koneoppimisen hyödyntämisestä vesijohtoverkostojen vuotojen hallinnassa on tehty ainakin 80 primääritutkimusta 1/2012-5/2022 välillä. Vuosittaisten julkaisumäärien perusteella aiheen tutkiminen on ollut aktiivista ja kasvanut vuodesta 2017 lähtien. Vuonna 2021 julkaistiin yhteensä 19 tutkimusta, ja vuonna 2022 niitä julkaistaan todennäköisesti vielä enemmän. Tutkimukset julkaistiin yhteensä 56:ssa eri julkaisupaikassa, mikä viittaa siihen, että aihe kiinnostaa tutkijoita laajalla rintamalla. Eniten tutkimuksia julkaistiin Water-nimisessä tieteellisessä lehdessä (7 kpl).

Tutkimuksista 96 % oli validointitutkimuksia eli tutkimuksia, joita ei sovelleta teollisessa ympäristössä. Tulos viittaa siihen, että tutkimustoiminta on teoreettista ja kaukana vesilaitosten operatiivisesta toiminnasta. Tutkimuksessa tunnistettiin yhteensä kolmetoista koneoppimismenetelmää, joista neuroverkkopohjaiset menetelmät olivat kaikkein suosituimpia. Tutkimuksista 41 % käytti neuroverkkopohjaisia menetelmiä. Toiseksi suosituin menetelmä oli tukivektorikone (34 %). Molempia menetelmiä käytetään ensisijaisesti datan luokitteluun. Vuosien 2022 ja 2021 aikana neuroverkkopohjaisten menetelmien osuus oli jo 53 % (tukivektorikoneiden 27 % ja

16 %). Käyttötarkoituksen perusteella koneoppimista hyödynnetään sekä vuotojen tunnistamiseen että paikantamiseen. Vuotojen hallintaan käytetään erityisesti putkistojen painetietoa (58 %), mutta myös virtaama- (29 %), ääni- (15 %) ja värinädataa (9 %).

Tutkimuksen tulosten perusteella vesitoimialan organisaatioiden kannattaa pitää koneoppimista erittäin potentiaalisena työkaluna vesijohtoverkoston vuotojen hallinnassa. Tutkimuksen tulokset sekä liitteessä B listatut tutkimusartikkelit yhdessä antavat erinomaisen tausta-aineiston toteuttaa tutkimuksia tai pilottihankkeita aiheeseen liittyen. Artikkeleihin tutustumalla saa esimerkiksi tietoa siitä, kuinka simulaatioympäristöjä voi rakentaa, millä tavalla koneoppimismalleja voi opettaa ja millaista dataa kannattaa hyödyntää.

Lähteet

[1] AALTO-YLIOPISTO. Aalto-yliopiston versio opetusministeriön julkaisutyyppiluokittelosta. URL https://aaltodoc.aalto.fi/doc_public/ohjeet/julkaisuluokitus2010.pdf, viitattu 8.4.2022.

[2] ADEDEJI, K. B., HAMAM, Y., ABE, B. T., JA ABU-MAHFOUZ, A. M. Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview. *IEEE Access 5* (2017), 20272–20285.

[3] ALLOGHANI, M., AL-JUMEILY OBE, D., MUSTAFINA, J., HUSSAIN, A., JA AL-JAAF, A. *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. Springer, Cham, 2020, ss. 3–21.

[4] ALPAYDIN, E. *Introduction to Machine Learning*. MIT press, Cambridge, Massachusetts, 2020.

[5] ALPAYDIN, E. *Machine Learning*. MIT Press, Cambridge, Massachusetts, 2021.

[6] ALSHUQAYRAN, N., ALI, N., JA EVANS, R. A systematic mapping study in microservice architecture. Julkaisusarjassa *IEEE 9th International Conference on Service-Oriented Computing and Applications* (Macau, 11 2016), 44–51.

[7] ARKSEY, H., JA O'MALLEY, L. Scoping studies: Towards a methodological framework. *International journal of social research methodology 8*, 1 (2005), 19–32.

[8] ARSENE, C. T., GABRYS, B., JA AL-DABASS, D. Decision support system for water distribution systems based on neural networks and graphs theory for leakage detection. *Expert Systems with Applications 39*, 18 (2012), 13214–13224.

[9] BADILLO, S., BANFAI, B., BIRZELE, F., DAVYDOV, I. I., HUTCHINSON, L., KAM-THONG, T., SIEBOURG-POLSTER, J., STEIERT, B., JA ZHANG, J. D. An introduction to machine learning. *Clinical Pharmacology & Therapeutics 107*, 4 (2020), 871–885.

[10] BHATIA, R., BENNO, S., ESTEBAN, J., LAKSHMAN, T., JA GROGAN, J. Unsupervised machine learning for network-centric anomaly detection in IoT. Julkaisusarjassa *3rd ACM CoNEXT Workshop on Big Data, Machine Learning and Artificial Intelligence for Data Communication Networks* (Orlando, 12 2019), 42–48.

[11] BUCKLEY, C., O'REILLY, M. A., WHELAN, D., FARRELL, A. V., CLARK, L., LONGO, V., GILCHRIST, M., JA CAULFIELD, B. Binary classification of running fatigue using a single inertial measurement unit. Julkaisusarjassa *IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks* (Eindhoven, 5 2017), 197–201.

[12] CARUANA, R., JA NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. Julkaisusarjassa *23rd International Conference on Machine learning* (Pittsburgh, 6 2006), 161–168.

[13] CHAKRABORTY, S., TOMSETT, R., RAGHAVENDRA, R., HARBORNE, D., ALZANTOT, M., CERUTTI, F., SRIVASTAVA, M., PREECE, A., JULIER, S., RAO, R. M., ET AL. Interpretability of deep learning models: A survey of results. Julkaisusarjassa *IEEE Smart World Congress* (San Francisco, 8 2017), 1–6.

[14] CHAN, T. K., CHIN, C. S., JA ZHONG, X. Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. *IEEE Access 6* (2018), 78846–78867.

[15] DEEPMIND TECHNOLOGIES LTD. Yrityksen kotisivut. URL `https://deepmind.com/research/case-studies/alphago-the-story-so-far`, viitattu 16.6.2022.

[16] DENG, L. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing 3* (2014), e2.

[17] DURELLI, V. H., DURELLI, R. S., BORGES, S. S., ENDO, A. T., ELER, M. M., DIAS, D. R., JA GUIMARAES, M. P. Machine learning applied to software testing: A systematic mapping study. *IEEE Transactions on Reliability 68*, 3 (2019), 1189–1212.

[18] DYBA, T., KITCHENHAM, B. A., JA JORGENSEN, M. Evidence-based software engineering for practitioners. *IEEE software 22*, 1 (2005), 58–65.

[19] EL-ZAHAB, S., JA ZAYED, T. Leak detection in water distribution networks: An introductory overview. *Smart Water 4*, 1 (2019), 1–23.

[20] ELGENDY, N., JA ELRAGAL, A. Big data analytics: A literature review paper. Julkaisusarjassa *Industrial Conference on Data Mining* (St. Petersburg, 7 2014), 214–227.

[21] ERTEL, W. *Introduction to Artificial Intelligence*. Springer, Cham, 2017.

[22] FRADKOV, A. L. Early history of machine learning. *IFAC-PapersOnLine 53*, 2 (2020), 1385–1390.

[23] GOODFELLOW, I., BENGIO, Y., JA COURVILLE, A. *Deep Learning*. MIT Press, Cambridge, Massachusetts, 2016.

[24] GOOGLE LLC. Googlen verkkosivusto, joka analysoi Google-hakuja. URL `https://trends.google.com/`, viitattu 1.3.2022.

[25] GUILLAUMIN, M., VERBEEK, J., JA SCHMID, C. Multimodal semi-supervised learning for image classification. Julkaisusarjassa *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, 6 2010), 902–909.

[26] GUPTA, A., JA KULAT, K. A selective literature review on leak management techniques for water distribution system. *Water Resources Management 32*, 10 (2018), 3247–3269.

[27] JABBARI, R., BIN ALI, N., PETERSEN, K., JA TANVEER, B. What is DevOps? A systematic mapping study on definitions and practices. Julkaisusarjassa *Scientific Workshop Proceedings of XP2016* (Edinburgh, 5 2016), 1–11.

[28] JORDAN, M. I., JA MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science 349*, 6245 (2015), 255–260.

[29] JUSTUS, D., BRENNAN, J., BONNER, S., JA MCGOUGH, A. S. Predicting the computational cost of deep learning models. Julkaisusarjassa *IEEE International Conference on Big Data* (Seattle, 12 2018), 3873–3882.

[30] KAELBLING, L. P., LITTMAN, M. L., JA MOORE, A. W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research 4* (1996), 237–285.

[31] KITCHENHAM, B. *Procedures for performing systematic reviews*. Tekninen raportti TR/SE-0401, Keele University, Keele, 7 2004.

[32] KITCHENHAM, B., BRERETON, P., JA BUDGEN, D. Mapping study completeness and reliability – A case study. Julkaisusarjassa *16th International Conference on Evaluation Assessment in Software Engineering* (Ciudad Real, 5 2012), 126–135.

[33] KITCHENHAM, B., BRERETON, P., LI, Z., BUDGEN, D., JA BURN, A. Repeatability of systematic literature reviews. Julkaisusarjassa *15th Annual Conference on Evaluation & Assessment in Software Engineering* (Durham, 4 2011), 46–55.

[34] KITCHENHAM, B., JA CHARTERS, S. *Guidelines for performing systematic literature reviews in software engineering*. Tekninen raportti EBSE-2007-01, University of Durham, Department of Computer Science, Durham, 7 2007.

[35] KITCHENHAM, B. A., BUDGEN, D., JA BRERETON, O. P. The value of mapping studies – A participant-observer case study. Julkaisusarjassa *14th International Conference on Evaluation and Assessment in Software Engineering* (Swindon, 4 2010), 1–9.

[36] KITCHENHAM, B. A., BUDGEN, D., JA BRERETON, O. P. Using mapping studies as the basis for further research – A participant-observer case study. *Information and Software Technology 53*, 6 (2011), 638–651.

[37] KITCHENHAM, B. A., BUDGEN, D., JA BRERETON, P. *Evidence-Based Software Engineering and Systematic Reviews*. CRC press, Boca Raton, 2015.

[38] KITCHENHAM, B. A., DYBA, T., JA JORGENSEN, M. Evidence-based software engineering. Julkaisusarjassa *26th International Conference on Software Engineering* (Edinburgh, 5 2004), 273–281.

[39] KUBAT, M., JA KUBAT. *An Introduction to Machine Learning*. Springer, Cham, 2017.

[40] LECUN, Y., BENGIO, Y., JA HINTON, G. Deep learning. *Nature 521*, 7553 (2015), 436–444.

[41] LEVAC, D., COLQUHOUN, H., JA O'BRIEN, K. K. Scoping studies: Advancing the methodology. *Implementation Science 5*, 1 (2010), 1–9.

[42] LI, Z., AVGERIOU, P., JA LIANG, P. A systematic mapping study on technical debt and its management. *Journal of Systems and Software 101* (2015), 193–220.

[43] MAXWELL, J. Understanding and validity in qualitative research. *Harvard Educational Review 62*, 3 (1992), 279–301.

[44] MITCHELL, T. *The discipline of machine learning*. Tekninen raportti CMU ML-06 108, Pittsburgh, 7 2006.

[45] MOHAMMED, M., KHAN, M. B., JA BASHIER, E. B. M. *Machine Learning: Algorithms and Applications*. CRC Press, Boca Raton, 2016.

[46] NAEEM, M., RIZVI, S. T. H., JA CORONATO, A. A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access 8* (2020), 209320–209344.

[47] NIELSEN, M. A. *Neural Networks and Deep Learning*. Determination Press, 2015.

[48] NOORBEHBAHANI, F., JA SABERI, M. Ransomware detection with semi-supervised learning. Julkaisusarjassa *10th International Conference on Computer and Knowledge Engineering* (Mashhad, 10 2020), 024–029.

[49] PATERNOSTER, N., GIARDINO, C., UNTERKALMSTEINER, M., GORSCHEK, T., JA ABRAHAMSSON, P. Software development in startup companies: A systematic mapping study. *Information and Software Technology 56*, 10 (2014), 1200–1218.

[50] PETERSEN, K., FELDT, R., MUJTABA, S., JA MATTSSON, M. Systematic mapping studies in software engineering. Julkaisusarjassa *12th International Conference on Evaluation and Assessment in Software Engineering* (Bari, 6 2008), 1–10.

[51] PETERSEN, K., JA GENCEL, C. Worldviews, research methods, and their relationship to validity in empirical software engineering research. Julkaisusarjassa *Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement* (Ankara, 10 2013), 81–89.

[52] PETERSEN, K., VAKKALANKA, S., JA KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology 64* (2015), 1–18.

[53] PHAM, M. T., RAJIĆ, A., GREIG, J. D., SARGEANT, J. M., PAPADOPOULOS, A., JA MCEWEN, S. A. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods 5*, 4 (2014), 371–385.

[54] PUUST, R., KAPELAN, Z., SAVIC, D., JA KOPPEL, T. A review of methods for leakage management in pipe networks. *Urban Water Journal 7*, 1 (2010), 25–45.

[55] SACKETT, D. L. Evidence-based medicine. *Seminars in Perinatology 21*, 1 (1997), 3–5.

[56] SAMPEDRO, C., MARTINEZ, C., CHAUHAN, A., JA CAMPOY, P. A supervised approach to electric tower detection and classification for power line inspection. Julkaisusarjassa *International Joint Conference on Neural Networks* (Peking, 7 2014), 1970–1977.

[57] SHALEV-SHWARTZ, S., JA BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, 2014.

[58] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V., LANCTOT, M., ET AL. Mastering the game of Go with deep neural networks and tree search. *Nature 529*, 7587 (2016), 484–489.

[59] SUTTON, R. S., JA BARTO, A. G. *Reinforcement learning: An introduction*. MIT press, Cambridge, Massachusetts, 2018.

[60] TARIQ, S., HU, Z., JA ZAYED, T. Micro-electromechanical systems-based technologies for leak detection and localization in water supply networks: a bibliometric and systematic review. *Journal of Cleaner Production 289* (2021), 125751.

[61] TESLA, INC. Yrityksen kotisivut. URL `https://www.tesla.com/AI`, viitattu 9.10.2022.

[62] VAN ENGELEN, J. E., JA HOOS, H. H. A survey on semi-supervised learning. *Machine Learning 109*, 2 (2020), 373–440.

[63] VESILAITOSYHDISTYS. Yhdistyksen kotisivut. URL `https://www.vvy.fi/vesihuolto/vesilaitosyhdistyksen-jasenisto/#osio-1-1509534744-2346-1`, viitattu 13.10.2022.

[64] WIERINGA, R., MAIDEN, N., MEAD, N., JA ROLLAND, C. Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Requirements Engineering 11*, 1 (2006), 102–107.

[65] WOHLIN, C., RUNESON, P., NETO, P. A. M. S., ENGSTRÖM, E., DO CARMO MACHADO, I., JA DE ALMEIDA, E. S. On the reliability of mapping studies in software engineering. *Journal of Systems and Software 86*, 10 (2013), 2594–2610.

[66] WU, X., JA ZHANG, C. Leakage Location Method of Water Supply Pipe Network Based on Integrated Neural Network. Julkaisusarjassa *IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers* (Dalian, 4 2022), 670–675.

[67] WU, Y., JA LIU, S. A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal 14*, 9 (2017), 972–983.

[68] XU, Q., LIU, R., CHEN, Q., JA LI, R. Review on water leakage control in distribution networks and the associated environmental benefits. *Journal of Environmental Sciences 26*, 5 (2014), 955–961.

[69] YLEISRADIO OY. Yle Areena-suoratoistopalvelu. URL `https://areena.yle.fi/podcastit/1-61912096`, viitattu 15.7.2022.

[70] YLEISRADIO OY. Yle.fi-uutissivusto. URL `https://yle.fi/uutiset/3-12560944`, viitattu 5.8.2022.

[71] ZAMAN, D., TIWARI, M. K., GUPTA, A. K., JA SEN, D. A review of leakage detection strategies for pressurised pipeline in steady-state. *Engineering Failure Analysis 109* (2020), 104264.

[72] ZHANG, Q.-S., JA ZHU, S.-C. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering 19*, 1 (2018), 27–39.

[73] ZHAO, Y., KOSOROK, M. R., JA ZENG, D. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine 28*, 26 (2009), 3294–3315.

[74] ZHOU, L., PAN, S., WANG, J., JA VASILAKOS, A. V. Machine learning on big data: Opportunities and challenges. *Neurocomputing 237* (2017), 350–361.

[75] ZHU, M., WANG, X., JA WANG, Y. Human-like autonomous car-following model with deep reinforcement learning. *Transportation Research Part C: Emerging Technologies 97* (2018), 348–368.

A Systemaattisen kirjallisuuskatsauksen vaiheet [34]

1. Tutkimuksen suunnittelu:
    - Tarpeen tunnistaminen systemaattiselle kirjallisuuskatsaukselle
    - Hankkeen käynnistäminen organisaatiossa*
    - Tutkimuskysymysten muodostaminen
    - Katsausprotokollan (sis. käytetyt menetelmät) muodostaminen
    - Katsausprotokollan arvioiminen*
2. Tutkimuksen suorittaminen:
    - Primäätitutkimusten tunnistaminen
    - Primääritutkimusten valitseminen
    - Primääritutkimusten laadun arviointi
    - Tietojen kerääminen
    - Synteesin muodostaminen
3. Tutkimuksen suorittaminen:
    - Levitystavan määrittäminen
    - Katsauksen muotoilu sopivaksi eri julkaisupaikkoihin (esim. konferenssipaperi)
    - Katsauksen arvioituttaminen (esim. vertaisarviointi)*

*Ei pakollisia vaiheita

B Tutkimuksen luokitusjärjestelmä

Vuosi	Otsikko	Tekijät	Julkaisutyyppi	Tutkimustyyppi	ML-menetelmät	Tarkoitus	Data
2022	Automation of Water Distribution System by Prediction of Water Consumption and Leakage Detection Using Machine Learning and IoT	Aggarwal ja Sehgal	Konferenssi	Validointitutkimus	ANN	Tunnistaminen	Virtaama
2022	Data-driven and model-based framework for smart water grid anomaly detection and localization	Wu et al.	Journaali	Arviointitutkimus	Muu/ei luokiteltu	Tunnistaminen ja paikantaminen	Paine ja virtaama
2022	Data-driven application of MEMS-based accelerometers for leak detection in water distribution networks	Tariq et al.	Journaali	Validointitutkimus	DT, KNN ja RF	Tunnistaminen	Värinä
2022	EKF-based observers for multi-leak diagnosis in branched pipeline systems	Delgado-Aguinaga et al.	Journaali	Validointitutkimus	KNN	Tunnistaminen ja paikantaminen	Paine ja virtaama
2022	Enabling low-cost automatic water leakage detection: a semi-supervised, autoML-based approach	Muniz Do Nascimento ja Gomes-Jr	Journaali	Validointitutkimus	Muu/ei luokiteltu	Tunnistaminen	Virtaama
2022	Experiments based comparative evaluations of machine learning techniques for leak detection in water distribution systems	Kammoun et al.	Journaali	Validointitutkimus	ANN, KNN, LOR ja SVM	Tunnistaminen	Paine ja virtaama
2022	Gene expression programming based mathematical modeling for leak detection of water distribution networks	Tijani ja Zayed	Journaali	Validointitutkimus	Muu/ei luokiteltu	Tunnistaminen	Ääni
2022	Improving the leak detection efficiency in water distribution networks using noise loggers	Tijani et al.	Journaali	Validointitutkimus	ANN, DT, KNN ja SVM	Tunnistaminen	Ääni
2022	Leak Identification Based on CS-ResNet under different leakage Apertures for Water-Supply Pipeline	Mei et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen	Värinä ja ääni
2022	Leakage Location Method of Water Supply Pipe Network Based on Integrated Neural Network	Wu ja Chen	Konferenssi	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Paine
2022	Multi-Source Information Fusion to Identify Water Supply Pipe Leakage Based on SVM and VMD	Wang et al.	Journaali	Validointitutkimus	SVM	Tunnistaminen	Värinä
2022	Predicting a water infrastructure leakage index via machine learning	Kzloz et al.	Journaali	Validointitutkimus	ANN ja PCA	Kunnon arvioiminen	Muu/ei luokiteltu

Vuosi	Otsikko	Tekijät	Julkaisutyyppi	Tutkimustyyppi	ML-menetelmät	Tarkoitus	Data
2022	Research on Leak Location Method of Water Supply Network based on Deep Neural Network Model	Wu ja Chen	Konferenssi	Validointitutkimus	ANN ja FCM	Paikantaminen	Paine
2022	Robust GMM Least Square Twin K-Class Support Vector Machine for Urban Water Pipe Leak Recognition	Liu et al.	Journaali	Validointitutkimus	SVM	Tunnistaminen	Ääni
2022	Stochastic Resonance Enhancement for Leak Detection in Pipelines Using Fluid Transients and Convolutional Neural Networks	Bohorquez et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Ääni ja paine
2021	A hybrid model-based method for leak detection in large scale water distribution networks	Fereidooni et al.	Journaali	Validointitutkimus	DT, KNN ja RF ja NB	Tunnistaminen ja paikantaminen	Virtaama
2021	An Improved Ensemble Deep Learning Model for Water Leakage Detection	Zhang et al.	Konferenssi	Validointitutkimus	EL	Tunnistaminen ja paikantaminen	Paine
2021	An innovative machine learning based framework for water distribution network leakage detection and localization	Fan ja Yu	Journaali	Validointitutkimus	ANN, KMC ja PCA	Tunnistaminen ja paikantaminen	Paine
2021	An iterative method for leakage zone identification in water distribution networks based on machine learning	Chen et al.	Journaali	Validointitutkimus	KMC ja RF	Tunnistaminen ja paikantaminen	Paine ja virtaama
2021	Appraising the Impact of Pressure Control on Leakage Flow in Water Distribution Networks	Mosetlhe et al.	Journaali	Validointitutkimus	Muu/ei luokiteltu	Muu/ei luokiteltu	Paine
2021	Clustering-Based Partitioning of Water Distribution Networks for Leak Zone Location	Ares-Milian et al.	Konferenssi	Validointitutkimus	SVM	Tunnistaminen ja paikantaminen	Virtaama
2021	Construction and Experimental Research on Leakage Sound Dataset of Urban Water Supply Pipeline	Chen et al.	Konferenssi	Validointitutkimus	ANN	Tunnistaminen	Ääni
2021	Detailed Leak Localization in Water Distribution Networks Using Random Forest Classifier and Pipe Segmentation	Lucin et al.	Journaali	Validointitutkimus	RF	Paikantaminen	Paine
2021	Leak Detection and Localization in Water Distribution Networks by Combining Expert Knowledge and Data-Driven Models	Soldevila et al.	Journaali	Validointitutkimus	KNN	Tunnistaminen ja paikantaminen	Virtaama
2021	Machine learning model and strategy for fast and accurate detection of leaks in water supply network	Fan et al.	Journaali	Validointitutkimus	ANN	Paikantaminen	Paine
2021	Merging Fluid Transient Waves and Artificial Neural Networks for Burst Detection and Identification in Pipelines	Bohorquez et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Paine
2021	Novel leakage detection and water loss management of urban water supply network using multiscale neural networks	Hu et al.	Journaali	Validointitutkimus	ANN, SVM, NB ja KNN	Tunnistaminen ja paikantaminen	Ääni
2021	Pressure Sensor Placement for Leak Localization Using Simulated Annealing with Hyperparameter Optimization	Morales-Gonzalez et al.	Konferenssi	Validointitutkimus	KNN	Paikantaminen	Paine

Vuosi	Otsikko	Tekijät	Julkaisutyyppi	Tutkimustyyppi	ML-menetelmät	Tarkoitus	Data
2021	Robust leak localization in water distribution networks using computational intelligence	Quinones-Grueiro et al.	Journaali	Validointitutkimus	ANN, PCA, LDA ja SVM	Tunnistaminen ja paikantaminen	Paine ja virtaama
2021	Teaching-Learning-Based Optimization of Neural Networks for Water Supply Pipe Condition Prediction	Elshaboury et al.	Journaali	Validointitutkimus	ANN	Kunnon arvioiminen	Muu/ei luokiteltu
2021	Use of Machine Learning for Leak Detection and Localization in Water Distribution Systems	Mashhadi et al.	Journaali	Validointitutkimus	ANN, LOR ja RF	Tunnistaminen ja paikantaminen	Paine ja virtaama
2021	Water leak detection based on convolutional neural network (CNN) using actual leak sounds and the hold-out method	Nam et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen	Ääni
2021	Water Leak Localization Using High-Resolution Pressure Sensors	Levinas et al.	Journaali	Validointitutkimus	KNN	Tunnistaminen ja paikantaminen	Paine
2021	Water pipes leak prediction in QWAT databases	Vaduva ja Honoriu	Konferenssi	Validointitutkimus	ANN	Tunnistaminen	Muu/ei luokiteltu
2020	A Data Driven Approach for Leak Detection with Smart Sensors	Liang et al.	Konferenssi	Arviointitutkimus	EL	Tunnistaminen	Virtaama ja ääni
2020	A Leak Detection in Water Pipelines Using Discrete Wavelet Decomposition and Artificial Neural Network	Chumchu	Symposiumi	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Ääni
2020	Burst Detection for District Metering Areas Based on Soft-Voting Ensemble Model in Water Distribution System	Lei et al.	Konferenssi	Validointitutkimus	EL	Tunnistaminen	Paine
2020	First Results in Leak Localization in Water Distribution Networks using Graph-Based Clustering and Deep Learning	Romero et al.	Konferenssi	Validointitutkimus	ANN	Paikantaminen	Paine
2020	Leak Detection and Topology Identification in Pipelines Using Fluid Transients and Artificial Neural Networks	Bohorquez et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Paine
2020	Leak localization in water distribution networks using classifiers with cosenoidal features	Santos-Ruiz et al.	Konferenssi	Validointitutkimus	KNN, NB, DT ja LDA	Paikantaminen	Paine
2020	Leakage Diagnosis of Water Supply Network Based on ACO-SVM	Yan et al.	Konferenssi	Validointitutkimus	SVM	Tunnistaminen	Muu/ei luokiteltu
2020	Leakage Diagnosis of Water Supply Network by SVM	Yan et al.	Konferenssi	Validointitutkimus	SVM	Tunnistaminen	Muu/ei luokiteltu
2020	Novel Leakage Detection and Localization Method Based on Line Spectrum Pair and Cubic Interpolation Search	Guo et al.	Journaali	Validointitutkimus	RF	Tunnistaminen ja paikantaminen	Ääni
2020	Pipe fault prediction for water transmission mains	Gorenstein et al.	Journaali	Validointitutkimus	RF	Tunnistaminen	Muu/ei luokiteltu
2020	Prelocalization and leak detection in drinking water distribution networks using modeling-based algorithms: A case study for the city of Casablanca (Morocco)	Taghlabi et al.	Journaali	Arviointitutkimus	RF	Tunnistaminen ja paikantaminen	Paine

Vuosi	Otsikko	Tekijät	Julkaisutyyppi	Tutkimustyyppi	ML-menetelmät	Tarkoitus	Data
2020	Sensor Placement for Leak Localization in Water Distribution Networks using Machine Learning	Madbhavi et al.	Konferenssi	Validointitutkimus	SVM, KNN ja RF	Tunnistaminen ja paikantaminen	Paine
2020	Water Leakage Detection in Hilly Region PVC Pipes using Wireless Sensors and Machine Learning	Moulik et al.	Konferenssi	Validointitutkimus	KMC	Tunnistaminen ja paikantaminen	Värinä
2019	A Machine Learning Approach to Water Leak Localization	Shravani et al.	Konferenssi	Validointitutkimus	LOR, NB, DT, SVM ja ANN	Tunnistaminen ja paikantaminen	Virtaama
2019	Cloud-Based Water Leakage Detection and Localization	Shravani et al.	Konferenssi	Validointitutkimus	ANN ja SVM	Tunnistaminen ja paikantaminen	Virtaama
2019	Evaluation of data driven models for pipe burst prediction in urban water distribution systems	Alizadeh et al.	Journaali	Validointitutkimus	SVM, GPR ja ANN	Tunnistaminen	Muu/ei luokiteltu
2019	Leak Detection and Location Based on ISLMD and CNN in a Pipeline	Zhou et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Paine
2019	Leak Localization in Water Distribution Networks using Deep Learning	Javadiha et al.	Konferenssi	Validointitutkimus	ANN	Paikantaminen	Paine
2019	Leak Localization in Water Distribution Networks Using Pressure and Data-Driven Classifier Approach	Sun et al.	Journaali	Validointitutkimus	LDA ja ANN	Tunnistaminen ja paikantaminen	Paine
2019	Leakage Classification Based on Improved Kullback-Leibler Separation in Water Pipelines	Luong ja Kim	Konferenssi	Validointitutkimus	SVM	Tunnistaminen	Ääni
2019	Leakage Identification in Water Distribution Networks with Error Tolerance Capability	Xie et al.	Journaali	Validointitutkimus	KMC	Tunnistaminen	Paine
2019	Machine-Learning-Based Leakage-Event Identification for Smart Water Supply Systems	Zhou et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen	Paine
2019	Monitoring Support for Water Distribution Systems based on Pressure Sensor Data	Geelen et al.	Journaali	Validointitutkimus	Muu/ei luokiteltu	Tunnistaminen	Paine
2019	Multiscale Gaussian process regression-based generalized likelihood ratio test for fault detection in water distribution networks	Fazai et al.	Journaali	Validointitutkimus	GPR	Tunnistaminen	Paine
2019	Novel Leak Location Approach in Water Distribution Networks with Zone Clustering and Classification	Quinones-Grueiro et al.	Konferenssi	Validointitutkimus	RF ja SVM	Tunnistaminen ja paikantaminen	Paine
2019	Optimal sensor placement for leak localization in water distribution networks based on a novel semi-supervised strategy	Li et al.	Journaali	Validointitutkimus	FCM	Paikantaminen	Paine
2019	Pattern Recognition and Clustering of Transient Pressure Signals for Burst Location	Manzi et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen ja paikantaminen	Paine
2019	The Analysis of Water Supply Operating Conditions Systems by Means of Empirical Exponents	Stanczyk ja Bursztа-Adamiak	Journaali	Validointitutkimus	NB, SVM, KNN, LDA ja EL	Kunnon arvioiminen	Paine ja virtaama
2019	Water Pipeline Leakage Detection Based on Machine Learning and Wireless Sensor Networks	Liu et al.	Journaali	Validointitutkimus	SVM ja PCA	Tunnistaminen	Paine

Vuosi	Otsikko	Tekijät	Julkaisutyyppi	Tutkimustyyppi	ML-menetelmät	Tarkoitus	Data
2018	A modern approach for leak detection in water distribution systems	Predescu et al.	Konferenssi	Validointitutkimus	KMC	Tunnistaminen	Paine ja virtaama
2018	An accelerometer-based leak detection system	El-Zahab et al.	Journaali	Validointitutkimus	SVM, DT ja NB	Tunnistaminen	Värinä
2018	Comparison of Classifiers for Leak Location in Water Distribution Networks	Quinones-Grueiro et al.	Symposiumi	Validointitutkimus	KNN, NB, ANN ja SVM	Paikantaminen	Muu/ei luokiteltu
2018	Experimental investigation into techniques to predict leak shapes in water distribution systems using vibration measurements	Butterfield et al.	Journaali	Validointitutkimus	RF	Muu/ei luokiteltu	Värinä
2018	Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems	Brentan et al.	Journaali	Validointitutkimus	KMC	Tunnistaminen ja paikantaminen	Paine
2018	Identification of High Pressure Critical Links in Water Distribution Systems	Jain et al.	Konferenssi	Validointitutkimus	SVM ja KNN	Kunnon arvioiminen	Paine
2018	Identification of urban drinking water supply patterns across 627 cities in China based on supervised and unsupervised statistical learning	De Clercq et al.	Journaali	Validointitutkimus	RF	Muu/ei luokiteltu	Muu/ei luokiteltu
2018	Performance optimization of a leak detection scheme for water distribution networks	Przystaka	Symposiumi	Validointitutkimus	ANN	Tunnistaminen	Virtaama
2017	A Classification Approach for Monitoring and Locating Leakages in a Smart Water Distribution Framework	Porwal et al.	Konferenssi	Validointitutkimus	SVM	Tunnistaminen ja paikantaminen	Paine ja virtaama
2017	Automatic configuration of kernel-based clustering: an optimization approach	Candelieri et al.	Konferenssi	Validointitutkimus	SVM ja KMC	Paikantaminen	Paine ja virtaama
2017	Leak detection in water pipeline by means of pressure measurements for WSN	Ayadi et al.	Konferenssi	Validointitutkimus	KNN, SVM ja NB	Tunnistaminen ja paikantaminen	Paine
2017	Leakage detection and prediction of location in a smart water grid using SVM classification	Porwal et al.	Konferenssi	Validointitutkimus	SVM	Tunnistaminen ja paikantaminen	Paine ja virtaama
2017	Novel Leakage Detection by Ensemble CNN-SVM and Graph-Based Localization in Water Distribution Systems	Kang et al.	Journaali	Validointitutkimus	ANN ja SVM	Tunnistaminen ja paikantaminen	Värinä
2017	Toward An Integrated Approach to Localizing Failures in Community Water Networks	Han et al.	Konferenssi	Validointitutkimus	RF ja SVM	Tunnistaminen ja paikantaminen	Paine ja virtaama
2015	Leak Localization in Water Distribution Networks using Pressure Residuals and Classifiers	Ferrandez-Gamot et al.	Symposiumi	Validointitutkimus	Muu/ei luokiteltu	Paikantaminen	Paine
2014	Analytical Leakages Localization in Water Distribution Networks through Spectral Clustering and Support Vector Machines. The Icewater Approach	Candelieri et al.	Journaali	Validointitutkimus	SVM	Paikantaminen	Paine ja virtaama
2012	Decision Support System for Water Distribution Systems Based on Neural Networks and Graphs Theory for Leakage Detection	Arsene et al.	Journaali	Validointitutkimus	ANN	Tunnistaminen	Paine ja virtaama

Koneoppimismenetelmien (ML) selityksiä:

- ANN = Neuroverkkopohjaiset menetelmät
- SVM = Tukiverktorikone
- KNN = K:n lähimmän naapurin menetelmä
- RF = Satunnaismetsä
- KMC = K:n keskiarvon klusterointi
- DT = Päätöspuu
- NB = Naiivi Bayesin luokitin
- PCA = Pääkomponenttianalyysi
- EL = Ensemble luokittelija
- LDA = Lineaarinen erotteluanalyysi
- LG = Logistinen regressio
- FCM = Sumean C:n keskiarvon klusterointi
- GPR = Gaussin prosessiregressio