Jukka Armas Joutsalainen

Selittävän tekoälyn potentiaalin hyödyntäminen valmistavassa teollisuudessa: Hyödyt, riskit ja parhaat käytännöt

Tietotekniikan kandidaatintutkielma

5. toukokuuta 2023

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Jukka Armas Joutsalainen

Yhteystiedot: jajoutzs@student.jyu.fi

Ohjaaja: Timo Tiihonen

Työn nimi: Selittävän tekoälyn potentiaalin hyödyntäminen valmistavassa teollisuudessa: Hyödyt, riskit ja parhaat käytännöt

Title in English: Realizing the Potential of Explainable Artificial Intelligence (XAI) in the Manufacturing Industry: Advantages, Challenges, and Best Practices

Työ: Kandidaatintutkielma

Opintosuunta: Kaikki opintosuunnat

Sivumäärä: 39+0

Tiivistelmä:

Tutkimuksen taustalla on pyrkimys selvittää selitettävän tekoälyn (XAI) vaikutusta laadunvalvontaan ja tuotannon suunnitteluun valmistavassa teollisuudessa. Tutkimus tunnisti tekijöitä, jotka vaikuttavat XAI:n toteutuksen tehokkuuteen ja sen mahdollisiin hyötyihin. Nämä tekijät puoltavat sitä, että XAI-menetelmien käyttöönotto teollisuudessa voi parantaa laatua ja tuotannon suunnitteluprosesseja tarjoamalla reaaliaikaista data-analyysiä ja päätöksentekoa.

Erityisesti XAI:n on osoitettu lisäävän tehokkuutta, vähentävän virheitä ja epäjohdonmukaisuuksia, parantavan tuotteiden laatua ja tehostavan toimitusketjun hallintaa, mikä vähentää tuotantokustannuksia ja lisää yleistä tehokkuutta. XAI:n tehokkuuteen vaikuttavia tekijöitä valmistavan teollisuuden laadunvalvonta- ja tuotannonsuunnitteluprosesseissa ovat tiedon saatavuus, valmistusprosessin monimutkaisuus, tehokkuus ja automaatiotaso.

Näiden havaintojen perusteella voidaan päätellä, että XAI:n käyttöönotto valmistavassa teollisuudessa voi tuottaa merkittäviä etuja, jotka ylittävät siihen liittyvät riskit, mikäli riskit tunnistetaan ja niitä hallitaan tehokkaasti. Käyttötarkoitukseen optimoidut ratkaisut voivat helpottaa XAI:n onnistunutta käyttöönottoa valmistavan teollisuuden prosesseissa.

Avainsanat: Selitettävä tekoäly (XAI), valmistava teollisuus, systemaattinen kirjallisuuskatsaus, hyödyt, riskit, parhaat käytännöt, tapaustutkimukset, koneoppiminen, syväoppiminen, optimointi, päätöksenteko

Abstract: This research aims to investigate the impact of explainable artificial intelligence (XAI) on quality control and production planning in the manufacturing industry. Through a systematic literature review, the study identified the factors that affect the efficiency of XAI implementation and its potential benefits. The literature review suggests that introducing XAI methods in the manufacturing industry can improve quality and production planning processes by providing real-time data analysis and decision support. Specifically, XAI has been shown to increase efficiency, reduce errors and inconsistencies, improve product quality, and enhance supply chain management, thereby reducing production costs and increasing overall efficiency. Factors affecting the effectiveness of XAI in manufacturing industry quality control and production planning processes include the availability of information, the complexities of the manufacturing process, efficiency, and level of automation.

Based on these findings, it can be concluded that XAI implementation in the manufacturing industry can result in significant benefits that exceed the associated risks, provided that the risks are identified and managed effectively. Solutions optimized for the intended use can facilitate the successful implementation of XAI in manufacturing industry processes.

Keywords: Explainable artificial intelligence (XAI), manufacturing industry, systematic literature review, benefits, risks, best practises, case studies, machine learning, deep learning, optimization, decision making

Termiluettelo

Black-box mallit	Musta-laatikko malli, (Black-box model) on koneoppimismallin tyyppi, jossa sisäiset työskentely- ja päätöksentekoprosessit eivät ole käyttäjän käytettävissä tai näkyvissä. Tämä tarkoittaa, että mallin käyttäytymistä ei voida helposti ymmärtää tai tulkita, jolloin on vaikea tunnistaa mahdollisia virhepäätelmiä tai selittää näiden päätösten taustalla olevia syitä. Black-box mallien läpinäkyvyyden puute voi asettaa haasteita niiden käytölle tietyissä sovelluksissa, kuten sellaisissa, jotka edellyttävät vastuullisuutta tai tulkintaa (Bishop 2006).
Glass-box mallit	Lasi-laatikko malli (Glass-box model) koneoppimismallin tyyppi, jonka sisäiset työskentely- ja päätöksentekoprosessit ovat läpinäkyviä, mikä helpottaa mallin käyttäytymisen ymmärtämistä sekä mahdollisten mallin käyttäytymisen vääristymien tunnistamista (Bishop 2006).
Ihminen silmukassa	Ihminen silmukassa (Human-In-The-Loop, HITL) liittyy laskentamalliin, jossa ihmiset ja tekoälyjärjestelmät työskentelevät yhdessä jonkin tietyn tehtävän suorittamiseksi. Yleensä ihmisen ohjaa prosessia, antaa palautetta ja hallitsee tilannetta tarvittaessa. Silmukassa kone oppii ihmisen palautteesta ja ihminen koneen tuottamista tuloksista (Veleso, Nanni ja Lippiello 2017).
Konenäkö	Konenäkö (Machine Vision), myös tietokonenäkö, tarkoittaa tietokonelähtöistä kykyä tulkita, analysoida ja ymmärtää visuaalista tietoa ympäröivästä maailmasta. Tämä sisältää kuvien ja videoiden käsittelyn ja analysoinnin sekä päätösten tai ennusteiden tekemisen tuotetun analyysin perusteella (Forsyth ja Ponce 2003).
Koneoppiminen	Koneoppiminen (Machine learning, ML) on tekoälyn alasuunta, joka keskittyy rakentamaan algoritmejä ja tilastollisia mal-

	leja, jotka voivat parantaa tarkkuuttaan suorittaessaan tiettyä tehtävää kokemusperäisen oppimisen kautta, ilman asiaan erikoistunutta ohjelmointia (Bishop 2006).
Laadunvalvonta	Laadunvalvonta on prosessi, jota organisaatiot käyttävät varmistaakseen, että tuotetut tuotteet tai palvelut täyttävät halutut laatustandardit. Siihen kuuluu mm. raaka-aineiden laadun, tuotantoprosessin ja valmiin tuotteen tarkastus (Juran ja Gryna 1988).
Läpinäkyvä tekoäly	Läpinäkyvä tekoäly (Transparent AI) on tekoälyjärjestelmän tyyppi, missä päätöksentekoprosessit ja niihin liittyvöt algoritmit ovat helposti ihmisten tulkittavissa (Mitchell 2019).
Neuroverkot	Neuroverkot (Neural Networks) ovat algoritmien luokka, joka ovat saaneet vaikutteita ihmisaivojen rakenteesta ja toiminnasta. Ne koostuvat useista kerroksista toisiinsa kytkettyjä solmuja, jotka prosessoivat ja muuntavat syöttötiedot tuottamaan saantitietoa. Jokainen solmu suorittaa syötteelleen yksinkertaisen laskutoimituksen ja tulos välitetään seuraavan kerroksen solmuille. Neuroverkkoja voidaan käyttää sekä ohjattuihin että ohjaamattomiin oppimistehtäviin (LeCun, Bengio ja Hinton 2015).
Post-hoc selitteet	Post-hoc-selitteet (post-hoc-explanations) viittaa prosessiin, jossa selitetään koneoppimismallin tuloksia tai päätöksiä sen jälkeen, kun malli on jo koulutettu ja sovellettu. Post-hoc-selitteet voivat antaa käsityksen mallin sisäisestä toiminnasta, mukaan lukien sen, mitkä ominaisuudet ovat tärkeimmässä asemassa tietyn ennusteen tekemisessä tai miksi tietty päätös tehtiin (C. Molnar 2019).
Päätöspuut	Päätöspuu (decision tree) on puurakenteinen päätöksien malli, jossa jokainen puun sisäinen solmu esittää attribuutille suoritettua testiä ja jokainen haara testin lopputulosta. Lehtisolmu taas esittää kaikkien attribuuttien laskennan jälkeistä päätöstä

	(class label) (Duda ja Hart 2000).
Reaaliaikainen data-analyysi	Reaaliaikainen data-analyysi tarkoittaa prosessia, jossa dataa analysoidaan sitä mukaan, kun sitä tuotetaan. Tämän tyyppisiä analyysejä käytetään tilanteissa, joissa täytyy tehdä välittömiä, dataan perustuvia päätöksiä (Sallam ym. 2019).
Satunnaiset metsät	Satunnainen metsät (Random Forests) ovat valvottuja oppimisalgoritmeja, jotka yhdistävät useita päätöspuita tarkemman ja kestävämmän mallin luomiseksi. Algoritmit valitsevat satunnaisesti datan ja ominaisuuksien osajoukot kunkin päätöspuun kouluttamiseksi ja kokoavat sitten ennusteet kaikista puista lopullisen ennusteen luomiseksi (Breiman 2001).
SHAP	SHAP (SHapley Additive ExPlanations) on menetelmä koneoppimismallien yksittäisten ennusteiden selittämiseen. Koneoppimisen yhteydessä SHAP-arvot tarjoavat tavan kvantifioida kunkin syöteominaisuuden merkityksen mallin saantotietoon, tietyssä ennusteessa (Lundberg ja Lee 2017).
Tukivektorikoneet	Tukivektorikoneet (Support Vector Machines, SVMs) ovat valvottuja oppimisalgoritmeja, joita voidaan käyttää luokittelu- tai regressiotehtäviin. Tukivektorikoneet yrittävät löytää parhaan hypertason, joka erottaa datapisteet eri luokkiin ja maksimoi samalla luokkien välisen marginaalin. Algoritmi käyttää ydinfunktiota datan muuntamiseen korkeampiulotteiseen tilaan, jossa hypertaso on helppo tunnistaa. (Cortes ja Vapnik 1995).
Tulkittavuus	Tulkittavuus tarkoittaa tasoa, jolla ihminen on kykenevä ymmärtämään syitä tekoälyjärjestelmän päätöksen takana (Doshi-Velez ja Kim 2017a).
Tulkittavuus malli-tasolla	Tulkittavuus malli-tasolla viittaa kykyyn ymmärtää, kuinka malli toimii kokonaisuudessaan, sen sijaan, että ymmerrättäisiin vain päätökset, joita malli tuottaa liittyen yksittäisiin ilmentymiin (instances). Mallin oletettavuus vastaa tasoa, jolla ihmi-

	nen ymmärtää mallin tuottamia syyseuraussuhteita ja päätöksiä (Ribeiro, Singh ja Guestrin 2016).
Vastuullinen tekoäly	Vastuullinen tekoäly (accountable AI) viittaa tekoälyn muotoon, joka on vastuullinen ja läpinäkyvä päätöksenteossaan, ja joka kantaa vastuun päätöstensä seurauksista. Sen avulla voidaan edistää eettistä ja vastuullista käyttöä tekoälyjärjestelmissä (Løken ja Bültmann 2021).
XAI	XAI (Explainable Artificial Intelligence) on tekoälyn osa-alue, jonka tavoitteena on luoda malleja, jotka ovat läpinäkyviä ja ihmisen tulkittavissa. Sen avulla voidaan edistää luottamusta ja parantaa päätöksentekoa (Molnar 2019).

Kuviot

Sisällys

1 Johdanto

Tekoäly (AI) ja koneoppiminen (ML) muuttavat monia toimialoja, kuten valmistavaa teollisuutta, tarjoamalla tarkkoja päätöksiä ja ennusteita päätöksiä vaativille prosesseille. Monien tekoälyjärjestelmien kohdalla käyttäjän voi olla kuitenkin vaikea ymmärtää järjestelmän tuottamien päätösten ja ennusteiden taustalla olevia perusteluja. Tästä seuraten monet tekoälyjärjestelmät mielletään "mustiksi laatikoiksi". Tämä läpinäkyvyyden ja tulkittavuuden puute on johtanut kasvavaan huoleen älykkäiden järjestelmien vastuullisuudesta. Erityisesti tilanteissa, joissa niiden päätöksillä on merkittävä vaikutus yksilöihin tai yhteisöihin. Näiden ongelmien ratkaisemiseksi Explainable Artificial Intelligence (XAI) pyrkii kehittämään tekoälyjärjestelmiä, jotka eivät ainoastaan tee tarkkoja ennusteita, vaan tarjoavat myös ymmärrettäviä, läpinäkyviä selityksiä tärkeille päätöksille ja ennustuksille.

Sovellusalueen kirjallisuudesta on löydettävissä kattavasti tietoa XAI-sovellusten tapaustutkimuksista. Näihin esimerkkeihin tukeutuen, tässä työssä käydään systemaattisesti läpi, millä tavoin järjestelmien läpinäkyvyys vaikuttaa valmistavan teollisuuden laadunvalvontaan ja tuotannon suunnitteluun. Pyrkimyksenä tuottaa perusteellinen käsitys riskeistä, rajoituksista ja onnistuneen käyttöönoton mahdollistamista suotuisista kehitykuluista.

Tämän saavuttamiseksi tarkastellaan selitettävyyden tärkeyttä teollisessa tuotannossa ja perinteisten "black-box"AI -järjestelmien asettamia haasteita. Ajatuksesta pääsemme etsimään ratkaisuja, tutkimalla XAI:n kykyjä tarjota tulkittavia selityksiä ja parantaa valmistavan teollisuuden päätöksentekoprosesseja. Lopuksi, luodaan katsaus XAI:n käyttöönoton rajoituksista ja käytännön haasteista todellisissa tuotantoympäristöissä.

2 Black-box AI:n haasteet valmistusprosesseissa

Jotta voitaisiin pohtia AI:n ja sen alakentän XAI:n roolia valmistavassa teollisuudessa, luodaan ensin hahmotelma siihen liittyvän toiminnaohjauksen pääsuunnista, joita XAI:n avulla pyritään parantamaan. Tässä luvussa käsittelemme black-box AI:n haasteita valmistavassa teollisuudessa. Aluksi annamme yleiskuvan tekoälyn käytöstä toiminnanohjauksessa. Tämän jälkeen tarkastelemme black-box tekoälyn käsitettä ja sen haasteita. Esittelemme myös useita esimerkkejä sen epäonnistumisista valmistavassa teollisuudessa. Lopuksi pohdimme mahdollisia ratkaisuja näihin haasteisiin.

2.1 Valmistavan teollisuuden toiminnanohjauksesta

Ennakoiva ylläpito ja prosessien optimointi ovat kaksi valmistavan teollisuuden kriittistä toiminnanohjauksen näkökantaa (Y. Li ym. 2020). Ennakoiva huoltaminen, joka on ennakoivan ylläpidon muoto, käyttää sensoreista ja muista lähteistä saatuja tietoja ennustaakseen, milloin laite todennäköisesti vikaantuu.

Saatujen tietojen perusteella voidaan ryhtyä ennaltaehkäiseviin toimenpiteisiin tuotannon keskeytysten ehkäisemiseksi tai minimoimiseksi. Koneoppimiseen perustuvia lähestymistapoja on ehdotettu valmistavan teollisuuden ennakoivaan ylläpitoon (X. Liu ym. 2021).

Prosessin optimointi puolestaan sisältää tuotantoprosessien analysoinnin tehottomuuden, "pullonkaulojen"ja parannusmahdollisuuksien tunnistamiseksi. Koneoppimistekniikoiden avulla voidaan tunnistaa tällaiset tehottomuudet ja pullonkaulat (X. Liu ym. 2021; Guo, Wang ja Dong 2021). Prosessin optimointi voi auttaa parantamaan tuotannon tehokkuutta ja alentamaan kustannuksia tunnistamalla ja ratkaisemalla tällaiset ongelmat.

2.2 Yleiskatsaus tekoälyyn valmistavassa teollisuudessa

Tekoälyn käyttö teollisuudessa on kasvanut merkittävästi viime vuosina. Grand View Researchin raportin (Research 2021) mukaan maailmanlaajuisen tekoälyn teollisuuden markkinoiden koon arvoksi arvioitiin 1,18 miljardia dollaria vuonna 2020. Sen ennustetaan kasva-

van 31,2 prosentin yhdistetyllä vuosikasvulla ja saavuttavan 18,82 miljardia dollaria vuoteen 2028 mennessä.

Yksi merkittävimmistä tekoälyn sovelluksista valmistusprosesseissa on ennakoiva ylläpito. Ennakoiva ylläpito käyttää koneoppimisalgoritmeja laitetietojen analysointiin ja huollon tarpeiden ennustamiseen. Tämä lähestymistapa voi vähentää tuotantokatkoksia, sekä huoltokustannuksia ja lisätä laitteiden luotettavuutta ja tehokkuutta. Kuitenkin, koska koneoppimisalgoritmit usein toimivat monimutkaisesti ja käyttävät suurta määrää dataa, on mahdollista, että koneoppimisalgoritmin lopullinen toiminta voi olla vaikeaa tai mahdotonta ennustaa ja ymmärtää. Tekoäly siis toimii Black-box-mallin mukaisesti (Zhao ym. 2021). AI:n läpinäkyvyyden puute ilmenee myös yleiseksikin miellettävällä tasolla, mm. prosessin valvonnassa. Kun tekoäly alkaa tunnistaa poikkeamia, sen toiminta voi olla myös silloin monimutkaista ja vaikeaa ymmärtää, varsinkin jos se käyttää syvällisiä oppimismenetelmiä (García-Sánchez ym. 2021).

Konenäkö on toinen AI-tekniikka, jota käytetään laajalti valmistuksessa. Sen avulla valmistajat voivat tarkastaa tuotteet ja havaita viat suurella tarkkuudella, mikä vähentää ihmislähtöisten tarkastustoimien tarvetta. Tämä tekniikka on erityisen hyödyllinen aloilla, kuten auto- ja elektroniikkateollisuudessa, joilla laadunvalvonta on kriittistä (Z. Li ym. 2020).

Kuitenkin, on tärkeää huomata, että tekoäly koostuu monista eri tekniikoista, kuten jo mainituista koneoppimisesta ja konenäöstä, mutta myös esimerkiksi luonnollisen kielen käsittelystä ja äänentunnistuksesta. Tekoälyn käyttö valmistavassa teollisuudessa voi tarkoittaa hyvien erilaisiakin yhdistelmien ja ratkaisujen kokonaisuuksia, eikä siten ole aina yksiselitteistä.

2.3 Mikä on "Black-box"AI ja mitkä ovat sen haasteet valmistavassa teollisuudessa?

Black-box AI viittaa tekoälyjärjestelmän tyyppiin, jossa sisäiset toiminnot eivät ole läpinäkyviä tai tulkittavissa ihmisille. Toisin sanoen järjestelmä tekee päätöksiä monimutkaisten algoritmien perusteella, joita on vaikea tai mahdoton ymmärtää ilman erikoisosaamista tai työkaluja. Tämä läpinäkyvyyden puute asettaa useita haasteita valmistusteollisuudessa, jossa tekoälyä käytetään yhä enemmän tuotantoprosessien, laadunvalvonnan ja toimitusketjun hallinnan optimointiin (Y. Liu ym. 2021; Arif ym. 2021).

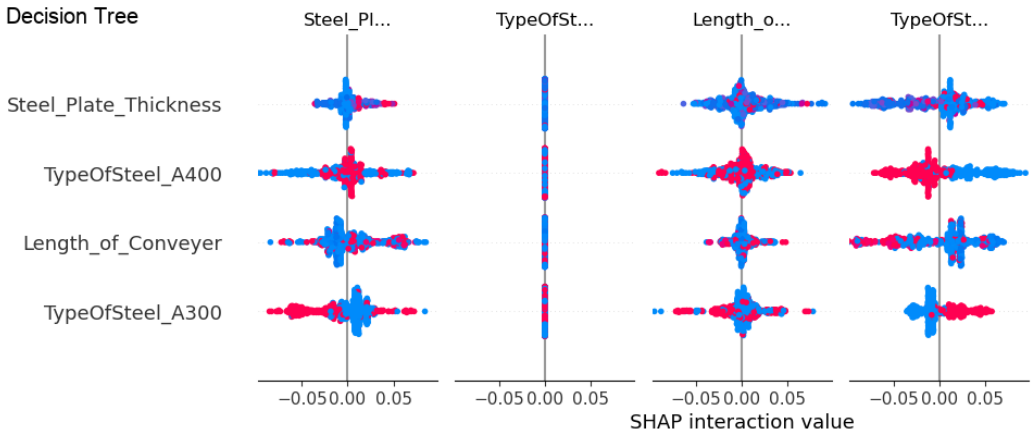
Yksi Black-box tekoälyn suurimmista haasteista on vaikeus selittää, miten tietty päätös tehtiin. Tämä tulkittavuuden puute voi olla ongelmallista tilanteissa, joissa ihmisten on ymmärrettävä tekoälyjärjestelmän tekemän päätöksen perustelut. Jos tekoälyjärjestelmä esimerkiksi havaitsee vian tuotteessa, ihmisen voi olla vaikea ymmärtää, miksi järjestelmä merkitsi kyseisen vian ongelmalliseksi (Yang ja Kim 2020).

Asiaa voidaan havainnollistaa tekolymallien selitettävyyteen pyrkivien SHAP-kuvaajien kautta. Menetelmän toimintaa osoitetaan soveltamalla sitä scikit-learn-kirjaston (Cournapeau ja contributors 2021) Black-box AI:n mukaisiin päätöspuu ja satunnainen-metsä malleihin, jotka ovat koulutettu teräslevyjen valmistukseen liittyvien vikojen osajoukolla (Lichman 2013). Koulutetut mallit esitetään SHAP-kuvaajiksi hyödyntäen Matplotlib-piirtokirjaston (Lundberg ja Erion 2021) SHAP-työkalua.

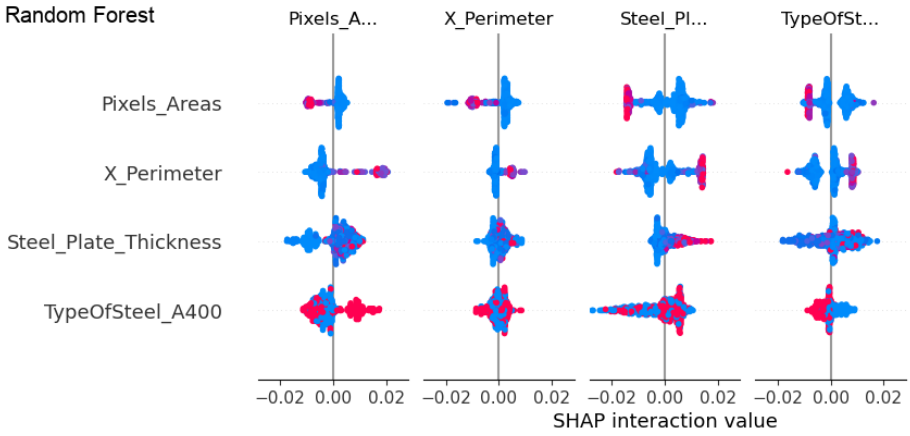
SHAP-kaavioiden pisteet, ts. jokainen tietopiste kuvaa tietojoukon yksittäistä instanssia, jonka sijainti horisontaalisesti akseliin nähden osoittaa, kuinka paljon kyseiseen pisteeseen liittyvä arvo poikkeaa tietojoukon ominaisuuden keskiarvosta. Piste, joka on pystyakselin oikealla puolella, vaikuttaa positiivisesti mallin ennusteeseen, kun taas tietopiste akselin vasemmalla puolella, vaikuttaa negatiivisesti. Pisteiden sijainti vertikaalisesti taas edustaa kohteen vaikutuksen suuruutta, magnitudia.

Tietojoukon ominaisuudet, kuten "X perimeter", on järjestetty tärkeysjärjestykseen siten, että tärkein ominaisuus on vasemmalla ja vähiten tärkeä oikealla. Pisteiden värin merkitys puolestaan on kuvastaa vastaavan ominaisuuden arvoa kyseiselle testijoukon esiintymäl-

le. Sininen edustaa kohteen alhaisia arvoja ja punainen korkeita arvoja, esimerkiksi "Pixel Areas"ominaisuuden kohdalla sininen piste osoittaisi pientä pikselialuetta vastaavassa testijoukon esiintymässä, kun taas punainen piste osoittaisi suurta pikselialuetta.



Kuvio 1. Black-box tekoälyn päätöspuu-malli, esitettynä SHAP-yhteenvetokaaviona.



Kuvio 2. Black-box tekoälyn satunnaisen metsän malli SHAP-yhteenvetokaaviona.

Kuviot osaltaan todentavat läpinäkyvyyden ja tulkittavuuden merkitystä. Huomataan, että suurella tietojoukolla, yksinkertaisempikin päätöspuu-malli voi tuottaa monimutkaisia entiteettejä, joissa syötetietojen merkitys korostuu. Verrattaessa keskenään satunnaisen metsän- ja päätöspuu-mallin pistejoukkojen vaakasuuntaisia etäisyyksiä niitä vastaavien ominaisuuksien pysty-akseleihin, nähdään, että satunnaisen metsän mallin pistejoukot ovat tiiviimpiä. Lisäksi, on tärkeää havaita, että myös vaihteluvälin osoittava asteikko on lyhyempi satunnaisen metsän kohdalla.

Tämä tarkoittaa, että samalla tietojoukolla se tuottaa tarkemman mallin, mutta niin tehdessään toteuttaa huomattavasti monimutkaisempia algoritmeja ja käyttää satunnaisuutta. Lopputuloksen kannalta olennaisten laskutoimituksien ja tehtävien päätöksien määrä siis kasvaa. Näin ollen käyttäjällä ei ole suotuisia lähtökohtia käsittää, kuinka malli toimii, tai miten se tekee päätöksiä.

Toinen "mustan laatikon"tekoälyn haaste on mahdollisuus erehdyksiin tai virheisiin päätöksentekoprosessissa. Koska ihmiset eivät voi helposti ymmärtää järjestelmän käyttämiä algoritmeja, voi olla vaikeaa havaita ja korjata mahdollisia erheitä tai virheitä. Tämä voi olla erityisen ongelmallista tuotannonaloilla, joissa pienillä virheillä tai harhoilla voi olla merkittäviä seurauksia laadunvalvontaan ja tuotannon tehokkuuteen (Arif ym. 2021).

Lisäksi black-box AI voi vaikeuttaa valmistajien säännösten ja standardien noudattamista. Esimerkiksi aloilla, joilla jäljitettävyys on kriittistä, kuten elintarvike- ja lääketeollisuudessa, black-box AI voi vaikeuttaa tuotteiden alkuperän jäljittämistä ja mahdollisten turvallisuusongelmien tunnistamista. Tämä läpinäkyvyyden puute voi myös vaikeuttaa sen varmistamista, että tekoälyjärjestelmät noudattavat eettisiä ohjeita ja periaatteita, kuten oikeudenmukaisuutta ja syrjimättömyyttä (Van Wynsberghe, Robbins ja Sullins 2020).

Kaiken kaikkiaan vaikka black box AI voi tarjota merkittäviä etuja tehokkuuden parantamisessa ja valmistusprosessien optimoinnissa, sen läpinäkyvyyden ja tulkittavuuden puute asettaa alalle merkittäviä haasteita. Näihin haasteisiin vastaamiseksi tarvitaan enemmän avoimuutta ja selitettävyyttä tekoälyjärjestelmissä sekä kehittää uusia standardeja ja määräyksiä, joilla voidaan varmistaa, että tekoälyä käytetään eettisesti ja vastuullisesti.

2.4 Esimerkkejä black-box tekoälyn epäonnistumisista

Vaikka tekoäly tarjoaa monia potentiaalisia hyötyjä, black-box mallien läpinäkymättömyys on johtanut useisiin suuriin epäonnistumisiin eri aloilla (Kirkpatrick ym. 2018). Valmistavassa teollisuudessa Black-box AI:n virheillä voi olla vakavia seurauksia, kuten turvallisuusriskejä, laatuvirheitä ja tuotantokatkoksia (IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems 2016).

Eräs esimerkki Black-box AI:n epäonnistumisesta valmistavassa teollisuudessa liittyy Boeingin 737 MAX lentokoneen tapaukseen. Vuosina 2018 ja 2019 mainitulle lentokonetyypille tapahtui kaksi kohtalokasta lento-onnettomuutta. Syynä oli viallinen lennonohjausjärjestelmä, nimeltään "Maneuvering Characteristics Augmentation System", MCAS (IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems 2016). MCAS-järjestelmä luotti sensoridataan ja black-box algoritmi hallitsi lentokoneen nousua tämän datan perusteella. MCAS-järjestelmästä löydettiin kuitenkin suunnitteluvirhe, joka aiheutti järjestelmän vastaanottaa virheellistä dataa, vain yhdeltä sensorilta mikä aiheutti lentoturmat. Tapaus nostatti huolta liittyen black-box AI:n käyttöön osana korkean merkityksen järjestelmiä, ilman asianmukaista läpinäkyvyyttä.

Toisena esimerkkinä mainiten tapaus liittyen Teslan valmistamaan Model S -sedaniin. Vuonna 2016 Model S:n kuolettava kolari johtui ajoneuvon Autopilot-järjestelmästä, joka käytti black-box-algoritmia ajoneuvon ohjaukseen. Algoritmi hallitsi ohjauksen osa-alueita, kuten renkaiden kääntämistä, ajoneuvon kiihtyvyyttä ja jarrutusta (IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems 2016). Kolari tapahtui, kun Autopilot ei havainnut valkoista kuorma-autoa, joka oli ajoneuvoon nähden asemoituna kirkkasta taivasta vasten. Models S törmäsi kuorma-autoon, tappaen kuljettajan. Tapaus korosti black-box AI:n rajoituksia monimutkaisten ja odottamattomien tilanteiden havaitsemisessa, sekä autonomisten järjestelmien vastuullisuuteen liittyviä eettisiä ja juridisia kysymyksiä.

Nämä esimerkit havainnollistavat potentiaalisia riskejä, jotka voivat aiheutua liiasta luottamisesta black-box tekoälyyn. Erityisesti korostuu black-box AI:n oikeellisen toiminnan syvä riippuvuus relaatiossa oleviin järjestelmiin, päteviin syötetietoihin ja muihin vaikutta-

viin tekijöihin. Tapausten kohdalla, ei siis ole oikein määrittää päävastuuta yksiomaan AI-järjestelmille, mutta järjestelmät epäonnistuivat tulkitsemaan virheellistä tietoa. Toisaalta, asia voidaan nähdä myös algoritmien koulutuksen puutteena, mikä on asia, jonka tunnistamista black-box-mallin mukainen toiminta ei mahdollista. Esimerkit korostavat myös syntynyttä tarvetta entistä läpinäkyvämmille ja tulkittavammille AI-malleille (Wachter, Mittelstadt ja Russell 2019).

3 XAI:n edut tuotannon suunnittelussa ja laadunvalvonnassa

Tässä luvussa käsitellään selitettävää tekoälyä ja sen etuja tuotannon suunnittelussa ja laadunvalvonnassa. Aluksi määritellään selitettävää tekoälyä, seuraten, tulkitsemme sen etuja verrattuna perinteiseen tekoälyyn, kuten parantunutta läpinäkyvyyttä, luotettavuutta sekä säännösten ja standardien mukaisuutta. Luvun lopussa käydään läpi XAI:n sisällyttämistä tuotannon prosesseihin ja esitetään tapaustutkimuksia onnistuneen sisällytyksen tueksi.

3.1 Mitä on selitettävä tekoäly?

Explainable Artificial Intelligence (XAI) on tekoälyn (AI) haara, jonka tavoitteena on tehdä tekoälymalleista ja päätöksentekoprosesseista läpinäkyvämpiä ja ymmärrettävämpiä ihmisille. XAI on erityisen tärkeä aloilla, kuten terveydenhuolto, rahoitus ja valmistus, joissa tekoälymallien tarkkuus ja tulkittavuus ovat ratkaisevan tärkeitä (Miller 2018).

XAI:n tarve johtuu monien tekoälymallien "mustan laatikon"ongelmasta, mikä tekee vaikeaksi ymmärtää, miten tietty päätös tehtiin, tai havaita ja korjata mallissa olevia virheitä tai harhoja. XAI-tekniikat pyrkivät ratkaisemaan tämän ongelman tarjoamalla selityksiä tekoälypäätöksille, mikä voi auttaa rakentamaan luottamusta ja vastuullisuutta, parantamaan päätöksentekoa ja helpottamaan ihmisen ja tekoälyn yhteistyötä (Doshi-Velez ja Kim 2017b).

XAI kattaa laajan valikoiman tekniikoita, mukaan lukien malliagnostiset menetelmät, kuten LIME (Local Interpretable Model-Agnostic Explanations) ja SHAP. Tarkemmin kuvaten, nämä ovat koneoppimisen tulkintatekniikoita, jotka luovat paikallisesti tarkkoja, sekä yksinkertaisia malleja selittämään monimutkaisten tekoälyjärjetelmien ennusteita. Lisäksi yleisiä ovat mallikohtaiset menetelmät, kuten päätöspuut ja sääntöpohjaiset järjestelmät (Samek, Wiegand ja Muller 2017). Näitä tekniikoita voidaan käyttää selitysten luomiseen eri tarkkuustasoilla yksittäisistä ennusteista globaaliin mallin käyttäytymiseen.

Vaikka XAI on vielä suhteellisen uusi ala, kiinnostus XAI-tekniikoita kohtaan on kasvanut viime vuosina, mistä on osoituksena XAI:hen liittyvien tutkimuspapereiden ja konferenssien

kasvava määrä (Arrieta ym. 2020). XAI:lla on potentiaalia alentaa tekoälyn käyttöönoton kynnystä ja tehdä tekoälystä luotettavampi monissa sovelluksissa, mukaan lukien valmistava teollisuus, missä se voi parantaa laadunvalvontaa, vian havaitsemista ja ennakoivaa ylläpitoa (Yang ym. 2020).

3.2 XAI:n edut perinteiseen tekoälyyn verrattuna

Yksi XAI:n tärkeimmistä eduista on parantunut läpinäkyvyys ja tulkittavuus. Toisin kuin perinteinen tekoäly, joka toimii "mustana laatikkona"eikä anna tietoa päätöksentekoprosesseista, XAI tuottaa malleja, jotka ovat läpinäkyviä ja tulkittavia, jolloin ihmiset voivat ymmärtää, kuinka malli tekee päätöksiään (Hofmann ja Tichy 2020). Huomataan, että XAI toteuttaa läpinäkyvämpää glass-box toimintamallia, minkä ansiosta valmistajat voivat tunnistaa mahdolliset vääristymät tiedoissa ja algoritmeissa ja tehdä tarvittavat muutokset päätöksentekoprosessien oikeudenmukaisuuden ja tarkkuuden varmistamiseksi.

Toinen XAI:n etu on lisääntynyt luottamus. Kyky ymmärtää ja tulkita tekoälyn päätöksentekoprosesseja lisää luottamusta teknologiaan, mikä on välttämätöntä sen laajalle käyttöönotolle valmistuksessa. Kun valmistajat voivat luottaa tekoälymalleihin, he voivat tehdä varmempia päätöksiä, parantaa tuotteiden laatua ja alentaa kustannuksia (B. Liu ym. 2021).

XAI voi myös auttaa valmistajia toimimaan määräyksien ja standardien mukaisesti. Terveydenhuollon ja rahoituksen kaltaisilla aloilla, joilla tekoälyn päätöksillä voi olla merkittäviä seurauksia, säädökset edellyttävät, että päätöksentekoprosessit ovat läpinäkyviä ja tulkittavia (Molnar 2019). XAI voi auttaa valmistajia noudattamaan näitä säännöksiä ja standardeja ja välttämään mahdollisia oikeudellisia ja eettisiä ongelmia.

3.3 XAI:n sisällyttämisestä tuotannon prosesseihin

Selitettävän tekoälyn (XAI) käyttöönotto valmistusprosesseissa sisältää joukon vaiheita sen tehokkuuden ja tarkkuuden varmistamiseksi. Yleensä ensimmäisenä toteutetaan on tiedonkeruu ja valmistelevat toimet, joihin kuuluu olennaisten tietolähteiden tunnistaminen ja tietojen kerääminen analysoitavaksi sopivassa muodossa. Tiedot on esikäsiteltävä epäjohdon-

mukaisuuksien, virheiden tai puuttuvien arvojen poistamiseksi, sekä tietojen laadun varmistamiseksi. Tämä on ratkaisevan tärkeää, koska XAI-mallin tarkkuus ja tehokkuus riippuvat tietojen laadusta (Chui, Manyika ja Miremadi 2018).

Seuraava vaihe on valita sopiva XAI-malli ja harjoitella sitä käyttämällä valmisteltujen tietojen osajoukkoa. Saatavilla on useita XAI-malleja, mukaan lukien päätöspuut, satunnaiset metsät, tukivektorikoneet ja neuroverkot. Mallin valinta riippuu valmistusprosessin erityisvaatimuksista ja saatavilla olevan tiedon tyypistä. Malli tulee validoida käyttämällä toista tietojen osajoukkoa sen varmistamiseksi, että se yleistyy hyvin uusiin tietoihin (Gunning 2017).

Kun XAI-malli on kehitetty ja varmennettu, se voidaan integroida valmistusprosessiin kehittämällä käyttöliittymä, joka näyttää mallin tulokset ja selitykset reaaliajassa, tai integroimalla malli olemassa oleviin ohjausjärjestelmiin (Doshi-Velez ja Kim 2017b). Viimeinen vaihe on arvioida XAI-mallin suorituskykyä ja seurata sen suorituskykyä ajan myötä. Tämä voi sisältää XAI-mallin suorituskyvyn vertaamisen olemassa oleviin malleihin tai käyttäjätutkimusten suorittamiseen mallin antamien selitysten tehokkuuden arvioimiseksi. Suorituskykymittareita, kuten tarkkuus ja muistaminen, tulee seurata säännöllisesti sen varmistamiseksi, että XAI-malli toimii odotetulla tavalla (Carvalho, Souza ja Gomes 2019).

4 XAI:n käyttöönoton rajoitukset ja käytännön haasteet tuotannon prosesseissa

Tämä luku käsittelee XAI:n käyttöönoton rajoituksia ja käytännön haasteita tuotannon prosesseissa. Luvussa tarkastellaan erityyppisiä prosesseja, joissa XAI:n käyttöönotolle voi ilmetä merkittäviä vaikeuksia.

4.1 Laajan vaihtelun prosessit

XAI-järjestelmät luottavat suuriin tietokokonaisuuksiin tehdäkseen tarkkoja ennusteita ja päätöksiä. Hyvin vaihtelevat prosessit, kuten kemian- ja lääketeollisuuden prosessit, voivat olla XAI-järjestelmille haastavia oppia, mikä johtaa epätarkkoihin ennusteisiin (Zhou, Liu ja Yu 2020). Tällaisissa tapauksissa perinteiset tilastolliset menetelmät voivat olla sopivampia kuin XAI.

4.2 Erittäin monimutkaiset, suuren tieto -ja työmäärän prosessit

Myös XAI-järjestelmillä voi olla vaikeuksia ymmärtää monimutkaisia prosesseja, joissa on useita vuorovaikutuksessa olevia tekijöitä. Esimerkiksi ilmailuteollisuudessa lentokoneen useiden komponenttien ja järjestelmien välinen monimutkainen vuorovaikutus voi olla vaikea toteuttaa XAI-mallissa (Chen, Zhou ja Li 2021). Tällaisissa tapauksissa tarkkojen ennusteiden ja päätösten tekemiseen voidaan tarvita alan tietoa ja ihmisten asiantuntemusta.

4.3 Prosessit, joihin liittyy kriittisiä turvallisuustekijöitä

Valmistusprosesseissa, joihin liittyy kriittisiä turvallisuustekijöitä, kuten ydin- tai kemianteollisuudessa, virheiden tai epäonnistumisten seuraukset voivat olla vakavia. Näin ollen, yksi keskeisistä haasteista on tarve varmistaa, että XAI-järjestelmät ovat tarkkoja ja luotettavia. Tämä edellyttää XAI-mallien huolellista validointia ja todentamista, mikä voi olla aikaa ja resursseja vievää (Lee ja Kwon 2020). Tällaisissa tapauksissa perinteiset, avoimemmat

sääntöpohjaiset järjestelmät voivat olla tarkoituksenmukaisempia.

4.4 Eettisten näkökohtien prosessit

Joissakin valmistusprosesseissa eettiset näkökohdat voivat olla ensiarvoisen tärkeitä. Esimerkiksi elintarvike- ja maatalousteollisuudessa XAI:n käyttö eläinten hoitoon tai torjunta-aineiden käyttöön liittyvän päätöksenteon automatisoimiseen saattaa herättää eettisiä huolenaiheita (Han ja Taylor 2020). Tällaisissa tapauksissa saattaa olla perusteltua luottaa ihmisen harkintaan ja asiantuntemukseen eettisten periaatteiden ja ohjeiden ohjaamana.

4.5 Prosessit, joissa tietojen saatavuus on rajoitettu

XAI-järjestelmät vaativat suuria määriä korkealaatuista dataa tarkkojen ennusteiden ja päätösten tekemiseen. Valmistusprosesseissa, joissa tietoa on niukasti tai vaikea saada, kuten kehittyvillä tai erikoistuneilla teollisuudenaloilla, XAI ei ehkä ole paras lähestymistapa (Wei ym. 2020). Vastaavissa tapauksissa perinteiset tilastomenetelmät tai sääntöihin perustuvat järjestelmät voivat olla sopivampia.

5 Näkemyksiä ja parhaita käytäntöjä valmistavan teollisuuden XAI-sovellusten tapaustutkimuksista

Tapaustutkimuksista on tullut yhä suositumpi työkalu XAI:n sovellusten tutkimiseen valmistavassa teollisuudessa. Tutkimalla todellisia esimerkkejä XAI:n toteutuksista, tutkijat voivat saada arvokkaita näkemyksiä tämän kehittyvän teknologian eduista, haasteista ja parhaista käytännöistä. Tässä luvussa keskustelemme tapaustutkimusten arvosta XAI:n tukemissa valmistusprosesseissa ja tutkimme joitain aiemmista tutkimuksista nousseita avainteemoja.

5.1 Johdatus tapaustutkimuksiin ja niiden arvo XAI:n tukemissa valmistusprosesseissa

Yksi tapaustutkimusten tärkeimmistä eduista on niiden kyky tarjota yksityiskohtainen ja kuvaava käsitys ympäristöstä, jossa XAI:ta toteutetaan (Gandomi ja Haider 2015). Tarkastelemalla konkreettisia käyttötapauksia, tutkijat ja alan ammattilaiset voivat saada tietoa erilaisiin valmistusprosesseihin liittyvistä ainutlaatuisista haasteista ja mahdollisuuksista (Kusiak ja Verma 2018). Esimerkiksi tapaustutkimus XAI:n toteutuksesta puolijohdevalmistuslaitoksessa saattaa paljastaa, että tekniikka on erityisen tehokas optimoimaan koneiden huoltoaikatauluja, kun taas tapaustutkimus elintarviketeollisuudessa saattaa korostaa selitettävyyden ja läpinäkyvyyden merkitystä päätöksenteossa.

Sen lisäksi, että tapaustutkimukset antavat oivalluksia XAI-toteutuksen erityiseen kontekstiin, ne voivat myös auttaa tunnistamaan laajempia malleja ja teemoja, jotka ulottuvat useisiin käyttötapauksiin (Yadav ja Srinivasan 2021). Esimerkiksi useiden tapaustutkimusten tarkastelu saattaa paljastaa, että onnistunut XAI-käyttöönotto edellyttää vahvaa yhteistyötä datatieteilijöiden, toimialueen asiantuntijoiden ja loppukäyttäjien välillä, tai sen, että tehokas viestintä ja koulutus ovat ratkaisevan tärkeitä käyttäjien omaksumisen ja sisäänoston varmistamiseksi (Kusiak ja Verma 2018).

Taustutkimukset voivat myös toimia inspiraation ja motivaation lähteenä organisaatioille, jotka harkitsevat XAI:n käyttöönottoa. Antamalla konkreettisia esimerkkejä XAI:han liit-

tyvistä eduista ja sijoitetun pääoman tuottoprosentista, tapaustutkimukset voivat edesauttaa rakentamaan liiketoimintaa (Gandomi ja Haider 2015).

5.2 Katsaus tunnettuihin tapaustutkimuksiin, ongelmien kuvauksiaa, toteutettuja ratkaisuja ja saavutettuja tuloksia

Viime vuosina on tehty lukuisia tapaustutkimuksia XAI:n soveltamisesta valmistusteollisuudessa. Nämä tutkimukset tarjoavat arvokasta tietoa valmistajien kohtaamista erityisongelmista, XAI:n avulla toteutetuista ratkaisuista ja saavutetuista tuloksista. Tässä alaluvussa käymme läpi joitakin merkittävimmistä tapaustutkimuksista, korostaen ongelman kuvauksia, toteutettuja ratkaisuja ja niistä saatavia etuja.

5.2.1 Ennakoiva ylläpito

Yksi lupaavimmista XAI:n sovelluksista valmistusteollisuudessa on ennakoiva ylläpito, jossa käytetään dataa ja koneoppimisalgoritmeja laitteiden vikojen ennustamiseen ennen niiden ilmenemistä (Wang ym. 2020). General Electricin tekemässä tapaustutkimuksessa XAI:ta käytettiin ennustamaan kaasuturbiinien jäljellä oleva käyttöikä, sensoreista kerättyjen reaaliaikaisten tietojen perusteella (Rodr'ıguez ym. 2018). XAI-järjestelmä pystyi havaitsemaan poikkeamat tiedoissa ja ennustamaan vikoja jopa 30 päivää etukäteen, mikä mahdollisti huollon ennakoivan ajoituksen ja lyhensi tuotantokatkoja jopa 20 prosenttia.

5.2.2 Laadunvalvonta

XAI:ta voidaan käyttää myös valmistusprosessin laadunvalvontaan. Siemensin tekemässä tapaustutkimuksessa XAI:ta käytettiin parantamaan yhtiön Saksan tehtaalla tuotetun teräksen laatua (Wunderlich ym. 2019). XAI-järjestelmä analysoi sensoreista ja kameroista saatuja tietoja teräksen vikojen havaitsemiseksi, ja tulokset näytettiin kojelautaan, jonka avulla käyttäjät saattoivat ryhtyä korjaaviin toimiin reaaliajassa. Tämän seurauksena vikojen määrä väheni 20 prosenttia ja tuotantonopeus nousi 5 prosenttia.

5.2.3 Prosessin optimointi

AI on osoittautunut arvokkaaksi työkaluksi valmistusprosessien optimoinnissa. Esimerkiksi Zhengin uutta selitettävissä olevaa tekoälymallia käytettiin älykkään valmistuksen tuotannon ajoituksen optimointiin. Malli analysoi reaaliajassa sensoreista ja koneista saatuja tietoja optimaalisten asetusten määrittämiseksi tuotantoparametreille, kuten tuotantonopeudelle, koneiden asennusajoille ja huoltoaikatauluille. XAI-järjestelmä pystyi antamaan suosituksia tuotannon ajoittamisesta, mikä paransi tuotannon yleistä tehokkuutta ja lyhensi seisokkeja (Zheng ym. 2021). Tämä tapaustutkimus osoittaa XAI:n mahdollisuudet optimoida valmistusprosesseja ja parantaa yleistä tuottavuutta.

5.2.4 Toimitusketjun hallinta

XAI:ta voidaan käyttää myös valmistusteollisuuden toimitusketjun hallintaan. Boschin tekemässä tapaustutkimuksessa XAI:ta käytettiin optimoimaan tavaroiden toimitus varastojen ja tuotantolaitosten välillä (Hentze, Wieker ja Rettkowski 2020). XAI-järjestelmä analysoi GPS-seurantalaitteiden ja liikenneantureiden tiedot ennakoidakseen toimitusajat ja tunnistaakseen optimaaliset reitit. Tämän seurauksena toimitusajat lyhenivät 15 prosenttia ja kuljetuskustannukset 10 prosenttia.

Nämä tapaustutkimukset osoittavat XAI:n potentiaalin valmistusteollisuudessa ja hyödyt, joita voidaan saavuttaa XAI-ratkaisujen käyttöönotolla. Ennakoimalla laitevikoja, parantamalla laadunvalvontaa, optimoimalla prosesseja ja hallitsemalla toimitusketjuja, XAI voi auttaa valmistajia vähentämään tuotantokatkoja, parantamaan tehokkuutta ja lisäämään kannattavuutta.

5.3 Keskustelua parhaista käytännöistä ja tapaustutkimuksista saaduista kokemuksista

Tapaustutkimusten tarkastelun kautta voimme tunnistaa joitain parhaita käytäntöjä ja oppeja XAI:n soveltamisesta valmistusteollisuudessa. Yksi tärkeimmistä XAI:n menestyksen tekijöistä valmistavassa teollisuudessa on tiedon laatu ja saatavuus. Kuten tapaustutkimuk-

set ovat osoittaneet, XAI luottaa suuresti tietoihin ennakoidakseen laitevikoja, parantaakseen laadunvalvontaa, optimoidakseen prosesseja, sekä hallitakseen toimitusketjuja. Siksi on erittäin tärkeää varmistaa, että tiedot ovat tarkkoja, jopa täydellisiä ja helposti saatavilla (L. Liu ym. 2019). Tämä edellyttää vankkaa tiedonkeruu- ja hallintajärjestelmää sekä työympäristön kulttuuria, joka arvostaa datalähtöistä päätöksentekoa.

Toinen kriittinen tekijä XAI:n menestykselle valmistavassa teollisuudessa on läpinäkyvyys ja tulkittavuus. XAI-järjestelmät tulee suunnitella antamaan selkeät selitykset päätöksistään, erityisesti tapauksissa, joissa näillä päätöksillä voi olla merkittäviä vaikutuksia valmistusprosessiin (Patel ym. 2019). Tämä ei ainoastaan auta rakentamaan luottamusta järjestelmään, vaan myös antaa käyttäjien ymmärtää ongelmien perimmäiset syyt ja ryhtyä korjaaviin toimiin.

XAI:n onnistunut käyttöönotto valmistusteollisuudessa edellyttää myös yhteistyötä ja integraatiota eri tiimien ja osastojen välillä. Esimerkiksi ennakoivien ylläpitojärjestelmien käyttöönotto saattaa edellyttää tiivistä yhteistyötä ylläpitotiimin, käyttötiimin ja data-analytiikkatiimin välillä (Gondzio ja Małachowski 2018). Vastaavasti toimitusketjun hallintajärjestelmien käyttöönotto saattaa edellyttää yhteistyötä logistiikkatiimin, hankintatiimin ja data-analytiikkatiimin välillä.

Lopuksi, XAI osana valmistavaa teollisuutta tulisi nähdä jatkuvan kehityskulun prosessina. XAI-järjestelmien käyttöönoton yhteydessä olisi hyödyllistä toteuttaa säännöllistä seurantaa ja arviointia, parannettavien alueiden tunnistamiseksi (Gandomi ja Haider 2015). Tämä edellyttää jatkuvan oppimisen ja parantamisen kulttuuria, sekä halukkuutta mukauttaa ja kehittää XAI-järjestelmiä tarpeen mukaan.

6 Tulevaisuuden suunnat ja vaikutukset XAI:lle valmistavassa teollisuudessa

Tässä luvussa tarkastellaan XAI:n tulevia suuntauksia ja vaikutuksia valmistavassa teollisuudessa. Näiksi luetaan XAI-teknologioiden kehitysaskeleet ja niiden mahdolliset vaikutukset, eettiset näkökohdat ja laajalle levinneen käyttöönoton sosiaaliset vaikutukset. Lisäksi käsitellään organisaation valmiudet ja muutoksenhallintastrategiat, sekä mahdolliset riskit ja ideat riskien vähentämiseksi.

6.1 Kehitysaskelia XAI-teknologioissa ja niiden mahdolliset vaikutukset

Valmistusteollisuudessa on käynnissä merkittävä muutos XAI-teknologioiden kehityksen myötä, ja on tärkeää ymmärtää niiden mahdolliset vaikutukset toimialaan. Yksi merkittävistä XAI-teknologioiden kehityksestä on syväoppimisalgoritmien lisääntyvä käyttö. Syväoppiminen on koneoppimisen tyyppi, joka mallintaa ja ratkaisee monimutkaisia ongelmia keinotekoisten hermoverkkojen avulla (Goodfellow, Bengio ja Courville 2016). Syväoppimisen käyttö XAI:ssa voi parantaa ennakoivaa ylläpitoa ja laadunvalvontaa, optimoida tuotantoprosesseja ja virtaviivaistaa toimitusketjun hallintaa (Xin Liu ym. 2019).

Toinen tärkeä kehitys XAI:ssa on luonnollisen kielen käsittelyn (Natural Language Processing, NLP) ja ihmisen ja tietokoneen välisen vuorovaikutuksen (Human-Computer Interaction, HCI) tekniikoiden integrointi. Näiden tekniikoiden avulla käyttäjät voivat olla vuorovaikutuksessa XAI-järjestelmien kanssa luonnollisella kielellä, mikä helpottaa koneoppimisalgoritmien tulosten ymmärtämistä ja tulkitsemista (Patel ja Mohapatra 2019).

Kasvava riippuvuus XAI-teknologioista herättää kuitenkin myös huolta mahdollisista vaikutuksista työllisyyteen. Kun koneet pystyvät paremmin suorittamaan monimutkaisia tehtäviä, on olemassa vaara, että jotkut työnkuvat voivat vanhentua tai tulla tarpeettomiksi (Brynjolfsson ja McAfee 2014). Siksi on ratkaisevan tärkeää pohtia, kuinka XAI-teknologiat voidaan integroida olemassa oleviin valmistusprosesseihin samalla, kun varmistetaan, että työntekijät

eivät jää jälkeen.

Tämän huolen ratkaisemiseksi jotkut tutkijat ehdottavat, että XAI-teknologioiden kehittämisessä tulisi keskittyä ihmisten kykyjen lisäämiseen sen sijaan, että niitä korvattaisiin. Tämä lähestymistapa, joka tunnetaan nimellä ihmiskeskeinen tekoäly, korostaa ihmisten ja koneiden välistä yhteistyötä parempien tulosten saavuttamiseksi (Horvitz ym. 2019). Yhdistämällä inhimillinen osaaminen XAI-teknologioihin, valmistusyritykset voivat luoda uusia työtehtäviä ja taitovaatimuksia, jotka parantavat heidän työvoimansa kykyjä.

6.2 XAI:n laajan käyttöönoton eettiset näkökohdat ja yhteiskunnalliset vaikutukset valmistavassa teollisuudessa

XAI-teknologioiden laaja omaksuminen valmistusteollisuudessa tuo mukanaan useita eettisiä näkökohtia ja mahdollisia sosiaalisia vaikutuksia. Tekoälyjärjestelmien kehittyessä on tärkeää ottaa huomioon niiden käytön vaikutukset ja varmistaa, että niitä kehitetään ja toteutetaan eettisesti ja vastuullisesti.

Eräs olennainen näkökanta liittyy puolueellisuuden ehkäisemiseen. Koneoppimisalgoritmit käyttävät suuria tietojoukkoja malliensa kouluttamisessa ja jos nämä tietojoukot ovat jollain tavalla vinoutuneet, myös tuloksena olevat mallit ovat puolueellisia (Barocas, Hardt ja Narayanan 2018). XAI:n etu black-box tekoälyyn on tarjota mahdollisuus puolueellisen toiminnan havaitsemiseen. Ihmislähtöisestä ajattelumallista puolueellisuus voisi tarkoittaa esimerkiksi mallin kouluttamista perustuen tietyn tasakoosteisen ihmisryhmän piirteisiin, joten mallin uudelleenkäyttö yleisesti jollekin toisella ihmisjoukolle voisi olla kyseenalaista.

XAI:n laajalla käyttöönotolla valmistavassa teollisuudessa voi myös olla merkittäviä sosiaalisia vaikutuksia. Esimerkiksi robottien ja muiden automatisoitujen järjestelmien lisääntyvä käyttö voi johtaa työpaikkojen menetykseen tietyillä teollisuuden aloilla (Brynjolfsson ja McAfee 2014). Tällä voi olla erityisen merkittävä vaikutus matalaa ammattitaitoa vaativien virkojen työntekijöihin, joille kouluttautuminen uudelleen korkeampaa ammattitaitoa vaativiin tehtäviin voi olla haasteellista. Samalla XAI-teknologioiden avulla saavutettava tehokkuuden ja tuottavuuden kasvu voi luoda uusia työmahdollisuuksia myös muilla teollisuuden aloilla.

Sidosryhmien, kuten työntekijöiden, yritysten ja yhteiskunnan, on selviydyttävä mahdollisista haasteista ja mahdollisuuksista, joita XAI:n laaja käyttöönotto valmistavassa teollisuudessa aiheuttaa. Työntekijöille tämä voi tarkoittaa uusien taitojen kehittämistä ja sopeutumista työpaikan muutoksiin. Yritysten on pohdittava, miten XAI-teknologiaa voidaan parhaiten integroida olemassa oleviin prosesseihinsa ja varmistaa, että niitä käytetään eettisesti ja vastuullisesti. Yhteiskunnan on pohdittava XAI:n mahdollisia vaikutuksia työllisyyteen, eriarvoisuuteen ja muihin sosiaalisiin kysymyksiin.

6.3 Organisaatioiden valmius ja muutoksenhallintastrategiat XAI:n käyttöönottamiseksi

Kun valmistusteollisuus jatkaa XAI:n käyttöönottoa, organisaatioiden valmiudet ja muutosten hallinta tulevat yhä tärkeämmiksi onnistuneen toteutuksen varmistamiseksi. Organisaation valmiudella tarkoitetaan organisaation kykyä mukautua ja ottaa käyttöön uusia teknologioita, kuten XAI, minimoiden samalla toimintahäiriöt (Coghlan ym. 2021). Muutoshallintaan sisältyy systemaattinen lähestymistapa, jolla valmistetaan, tuetaan ja autetaan yksilöitä, ryhmiä ja organisaatioita siirtymään onnistuneesti nykyisestä tilasta halutunlaiseen, tulevaan tilaan (Hiatt 2018). Tässä osiossa tutkimme organisaation valmiuden ja muutoksenhallinnan periaatteita XAI:n käyttöönoton yhteydessä valmistavassa teollisuudessa.

Organisaation valmius XAI:n käyttöönottamiseksi tuotannossa riippuu useista tekijöistä, kuten organisaation kulttuurista, johtajuudesta, teknologiainfrastruktuurista ja työvoiman taidoista (Chauhan, Khurana ja Jain 2020). Sen varmistamiseksi, että organisaatio on valmis XAI:n käyttöönottoon, on erittäin tärkeää ymmärtää nämä tekijät selkeästi ja korjata mahdolliset aukot organisaation valmiudessa (Coghlan ym. 2021).

Muutoksenhallintastrategiat voivat auttaa tuotanto-organisaatioita valmistautumaan XAI:n käyttöönottoon lisäämällä tietoisuutta, luomalla kiireellisyyden tunnetta, kehittämällä visiota ja viestimällä siitä, vahvistamalla muita ja luomalla lyhyen aikavälin tavoitteita (Hiatt 2018). Näillä strategioilla pyritään helpottamaan organisaation muutosta ottamalla sidosryhmät mukaan muutosprosessiin, tarjoamalla heille tarvittavat tiedot ja resurssit sekä vahvistamalla heidän sitoutumistaan muutokseen (Coghlan ym. 2021).

Muutoksenhallintastrategioita voidaan käyttää helpottamaan XAI:n käyttöönottoa kehittämällä kattava toteutussuunnitelma, joka sisältää tavoitteet, aikataulut ja virstanpylväät (Hiatt 2018). Suunnitelman tulee sisältää myös riskienhallintastrategia mahdollisten riskien tunnistamiseksi ja lieventämisstrategioiden kehittämiseksi niiden käsittelemiseksi (Coghlan ym. 2021).

Toimivan muutoksenhallintastrategian edellytyksenä on usein työvoiman uudelleenkoulutus ja osaamisen parantaminen sen varmistamiseksi, että työntekijöillä on tarvittavat taidot työskennellä XAI-tekniikoiden kanssa (Chauhan, Khurana ja Jain 2020). Tämä vaatii merkittäviä investointeja työntekijöiden koulutus- ja kehitysohjelmiin, mikä voi olla kallista ja aikaa vievää.

7 Johtopäätös ja yhteenveto

Tämä tutkielma keskittyi black-box AI:n toteuttamisen haasteisiin ja mahdollisuuksiin valmistavassa teollisuudessa. Työssä tutkittiin erityisesti selitettävän tekoälyn etuja tuotannon suunnittelussa ja laadunvalvonnassa sekä XAI:n toteutuksen rajoituksia ja käytännön haasteita valmistusprosesseissa.

Tulokset viittaavat siihen, että vaikka black-box AI voi tuoda merkittäviä etuja valmistukseen, se asettaa myös useita haasteita, joihin on puututtava. Näitä haasteita ovat muun muassa avoimuuden ja tulkittavuuden puute, ennakkoluulojen ja syrjinnän riski, eettiset huolenaiheet sekä tietojen saatavuuden ja laadun rajoitukset.

Näiden haasteiden voittamiseksi tutkielmassa ehdotettiin XAI:n käyttöönottoa, joka tarjoaa enemmän läpinäkyvyyttä, tulkittavuutta ja vastuullisuutta tekoälyjärjestelmissä. Tutkielma sisälsi myös tapaustutkimuksia onnistuneesta XAI:n toteutuksesta valmistusprosesseissa, jotka osoittivat XAI:n arvon ennakoivassa kunnossapidossa, laadunhallinnassa, prosessien optimoinnissa ja toimitusketjun hallinnassa.

Pohdittaessa XAI:n soveltuvuutta johonkin tiettyyn prosessiin, puoltaa tutkielma kaikkiaan kokonaisvaltaista lähestymistapaa. Tämä malli huomioi tiedon laadun ja saatavuuden, läpinäkyvyyden ja tulkittavuuden, sekä yhteistyön, integraation sekä jatkuvan kehityksen merkityksen. Edesauttaen varmentamaan tekoälyn täyden potentiaalin ymmärtämisen ja minimomaan samalla black-box AI:n riskit ja haasteet.

Lähteet

Arif, Waqas, Liang Li, Yangyang Jia ja Jianxin Li. 2021. “Black-Box Machine Learning and Its Challenges in Industry 4.0: A Survey”. Teoksessa *2021 IEEE 16th International Conference on Intelligent Computation Technology and Automation (ICICTA),* 69–74. IEEE.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Alexandre Bennetot, Siham Tabik, Alberto Barbado ja Ana Garcia-Serrano. 2020. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. *Information Fusion* 58:82–115.

Barocas, Solon, Moritz Hardt ja Arvind Narayanan. 2018. “Fairness in Machine Learning”. *NIPS 2018 Tutorial.*

Bishop, C. M. 2006. *Pattern recognition and machine learning (Vol. 1).* Springer.

Breiman, Leo. 2001. “Random forests”. *Machine learning* 45 (1): 5–32.

Brynjolfsson, Erik, ja Andrew McAfee. 2014. “The second machine age: Work, progress, and prosperity in a time of brilliant technologies”. *Journal of Economic Literature* 52 (2): 385–94.

Carvalho, Rafael Nunes, Jos’e Carlos Bittencourt de Souza ja H’elio Pedrini Gomes. 2019. “Understanding machine learning interpretability through neural network representation”. *Expert Systems with Applications* 118:43–58.

Chauhan, Alka, Sandeep Khurana ja Vipul Jain. 2020. “Exploring the challenges and opportunities for adopting AI in manufacturing: a review”. *Journal of Manufacturing Technology Management* 31 (2): 385–415.

Chen, Zeyang, Yu Zhou ja Xiaobing Li. 2021. “Explainable artificial intelligence for predicting aircraft structural repair”. *Sensors* 21 (3): 862.

Chui, Michael, James Manyika ja Mehdi Miremadi. 2018. “AI, automation, and the future of work”. *McKinsey Global Institute* 1:1–7.

Coghlan, David, Regina Connolly, Rob Corbett ja Peter McNamara. 2021. “Organisational readiness for the adoption of artificial intelligence technologies: a systematic literature review”. *Journal of Decision Systems* 30 (2): 133–156.

Cortes, Corinna, ja Vladimir Vapnik. 1995. “Support-vector networks”. *Machine learning* 20 (3): 273–297.

Cournapeau, D., ja contributors. 2021. *scikit-learn: Machine Learning in Python.* Version 1.0. https://scikit-learn.org/stable/index.html.

Doshi-Velez, Finale, ja Been Kim. 2017a. *Towards A Rigorous Science of Interpretable Machine Learning.* arXiv preprint arXiv:1702.08608. Accessed on 20 Feb 2023.

———. 2017b. “Towards a rigorous science of interpretable machine learning”. *arXiv preprint arXiv:1702.08608.*

Duda, R., ja P. Hart. 2000. “Pattern classification and scene analysis”. *John Wiley and Sons* 1:2.

Forsyth, David A., ja Jean Ponce. 2003. “Computer Vision: A Modern Approach”.

Gandomi, A, ja M Haider. 2015. “Beyond the hype: Big data concepts, methods, and analytics”. *International Journal of Information Management* 35:137–144.

García-Sánchez, Francisco, Elena García-Sánchez, Ángel García-Crespo ja Mario Piattini. 2021. “Evaluating explainability techniques for machine learning models in predictive maintenance”. *Applied Soft Computing* 98:106887.

Gondzio, J, ja B Małachowski. 2018. “Predictive maintenance: a case study on rotating equipment”. *Journal of Intelligent Manufacturing* 29:1429–1442.

Goodfellow, Ian, Yoshua Bengio ja Aaron Courville. 2016. *Deep learning.* MIT Press.

Gunning, David. 2017. “Explainable artificial intelligence (xai)”. *Defense Advanced Research Projects Agency (DARPA)* 2:77–82.

Guo, Q., Y. Wang ja S. Dong. 2021. “Integrated Predictive Maintenance and Process Optimization in the Manufacturing Industry Based on Machine Learning”. *IEEE Transactions on Industrial Informatics* 17 (2): 1355–1364.

Han, Zhao, ja James EW Taylor. 2020. “Explainable artificial intelligence for sustainable and resilient food systems”. *Sustainable Production and Consumption* 24:98–111.

Hentze, Julian, Holger Wieker ja Jens Rettkowski. 2020. “Predictive Logistics: A Case Study in Industrial AI”. Teoksessa *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE),* 57–58. IEEE.

Hiatt, Jeff. 2018. *Change management: The people side of change.* Prosci.

Hofmann, Patricia, ja Matthias Tichy. 2020. “Data and AI ethics in Industry 4.0 and beyond: challenges, solutions, and future research directions”. *Journal of Business Research* 117:715–728.

Horvitz, Eric, Michael Muller, Jeff Ho, Richard Wang, Diane Tang ja Avi Char. 2019. “Principles of human-centered machine learning”. *Communications of the ACM* 62 (1): 68–77.

IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* IEEE Standards Association.

Juran, J. M., ja F. M. Gryna. 1988. *Juran’s Quality Control Handbook.* McGraw-Hill.

Kirkpatrick, Kathryn, Frank Pasquale, Victor Raskin ja Farhi Safaei. 2018. “The Need for Better Transparency and Accountability in Autonomous Vehicles”. *IEEE Intelligent Transportation Systems Magazine* 10 (2): 6–17.

Kusiak, Andrew, ja Anit Verma. 2018. “Smart manufacturing: Past research, present findings, and future directions”. *Journal of Manufacturing Systems* 48:1–14.

LeCun, Yann, Yoshua Bengio ja Geoffrey Hinton. 2015. “Deep learning”. *Nature* 521 (7553): 436–444.

Lee, Jinho, ja Ohbyung Kwon. 2020. “Explainable artificial intelligence (XAI) for robust decision making”. *Expert Systems with Applications* 155:113403.

Li, Y., C. Zhang, Z. Liu, J. Wang ja W. Xiong. 2020. “A Framework for Process Optimization in the Manufacturing Industry Based on Machine Learning Techniques”. *IEEE Transactions on Industrial Informatics* 16 (6): 4082–4091.

Li, Ziyi, Zhenhua Zhou, Henglin Chen, Yu-Kuan Lai ja Ling Shao. 2020. “Deep learning-based quality inspection of manufacturing”. *IEEE Transactions on Industrial Informatics* 16 (6): 4043–4053.

Lichman, M. 2013. *UCI Machine Learning Repository.* University of California, Irvine, School of Information ja Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Steel+Plates+Faults.

Liu, Bo, Xiao Ma, Hao Wang ja Jia Liu. 2021. “Towards a Sustainable and Ethical Development of Artificial Intelligence”. *Sustainability* 13 (4): 2179.

Liu, L, X Xu, J Li ja X Chen. 2019. “A survey of industrial applications of artificial intelligence towards data-driven smart manufacturing”. *Journal of Manufacturing Systems* 51:1–10.

Liu, X., X. Huang, Y. Zhang ja Y. Hu. 2021. “A Machine Learning-Based Approach for Predictive Maintenance in the Manufacturing Industry”. *IEEE Transactions on Industrial Informatics* 17 (6): 4146–4155.

Liu, Xin, Yan-Fu Li, Bo Li ja Dong Li. 2019. “A survey of deep learning-based approaches for predictive maintenance”. *IEEE Access* 7:142363–142375.

Liu, Yue, Peng Chen, Yong Shi, Jianqiang Yu ja Weiming Zhang. 2021. “Black-Box Machine Learning and Decision Making for Industry 4.0”. *IEEE Access* 9:54555–54569.

Lundberg, Scott M, ja Su-In Lee. 2017. “A unified approach to interpreting model predictions”. Teoksessa *Advances in neural information processing systems,* 4765–4774.

Lundberg, Scott M., ja Gabriel G. Erion. 2021. *SHAP (SHapley Additive exPlanations).* Version 0.39.0.

Løken, Susanne L., ja M Bültmann. 2021. *The Social and Economic Implications of Artificial Intelligence Technologies in the Health Care Sector.* Springer.

Miller, Tim. 2018. “Explanation in artificial intelligence: Insights from the social sciences”. *arXiv preprint arXiv:1806.07757.*

Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans.* W. W. Norton / Company.

Molnar, C. 2019. *Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning.* Springer.

———. 2019. *Interpretable Machine Learning.* Springer.

Patel, R, H Teng, R Zhang, Y Liu ja H Li. 2019. “A review of explainable AI (XAI) in the context of industrial applications”. *IEEE Access* 7:135287–135305.

Patel, Vivek, ja Dipti Mohapatra. 2019. “A review of natural language processing and human-computer interaction techniques for intelligent manufacturing systems”. *Journal of Intelligent Manufacturing* 30 (5): 1821–1842.

Research, Grand View. 2021. “Artificial Intelligence in Manufacturing Market Size, Share and Trends Analysis Report By Offering (Hardware, Software, Services), By Technology (Machine Learning, Computer Vision), By Application, By Region, And Segment Forecasts, 2021-2028”. *Grand View Research.*

Ribeiro, M. T., S. Singh ja C. Guestrin. 2016. “Why Should I Trust You? Explaining the Predictions of Any Classifier”. Teoksessa *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144.

Rodr’ıguez, Sara, David Mart’ın, Ricardo Corral, Aitor Alonso ja Eduardo Cuevas. 2018. “Predictive maintenance for gas turbines based on deep learning techniques”. Teoksessa *2018 IEEE International Conference on Big Data (Big Data),* 3808–3811. IEEE.

Sallam, Rita L., Faith Chong, Joao Tapadinhas ja Thomas Oestreich. 2019. *Market Guide for Augmented Analytics and BI.* Gartner.

Samek, Wojciech, Tobias Wiegand ja Klaus-Robert Muller. 2017. “Towards explainable artificial intelligence: methods and tools”. *arXiv preprint arXiv:1708.08296.*

Wachter, Sandra, Brent Mittelstadt ja Chris Russell. 2019. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”. *Harvard Journal of Law and Technology* 33 (2): 173–209.

Van Wynsberghe, Aimee, Sarah Robbins ja John Sullins. 2020. “Addressing ethical concerns in artificial intelligence and robotics”. *Science* 369 (6505): 1055–1056.

Wang, Dong, Wei Liao, Rui Zhou ja Yan Liu. 2020. “A review of predictive maintenance and its role in enhancing manufacturing performance”. *Journal of Manufacturing Systems* 56:258–271.

Wei, Jun, Fei Tao, Shaokun Fan, Xichun Luo, Bo Li ja Ling Li. 2020. “A survey of data-driven approaches in smart manufacturing”. *Journal of Intelligent Manufacturing* 31 (7): 1713–1732.

Veleso, M., M. Nanni ja V. Lippiello. 2017. “Human-in-the-Loop Machine Learning”. *ACM Transactions on Internet Technology (TOIT)* 17 (2): 15.

Wunderlich, Nico, Till Pfeiffer, Johannes Sch"onberger, Christopher Klinkm"uller, Christoph Hollauer ja Michael ten Hompel. 2019. “Explainable artificial intelligence for production in the age of Industry 4.0”. *Procedia CIRP* 84:15–20.

Yadav, Pradeep, ja R Srinivasan. 2021. “Industry 4.0 technologies for sustainable manufacturing: A systematic literature review and research agenda”. *Journal of Cleaner Production* 293:126186.

Yang, Guanyu, Lirong Peng, Jian Xu ja Huachun Wang. 2020. “Exploring the applications of explainable artificial intelligence in additive manufacturing”. *Robotics and Computer-Integrated Manufacturing* 63:101925.

Yang, Jing, ja Sung-Hee Kim. 2020. “Black-box explanation of decision-making based on deep learning: A review”. *Computers in Industry* 121:103295.

Zhao, Yuqian, Wei Chen, Rong Zhu, Yanqing Jiang ja Jiafu Li. 2021. “A survey on machine learning for predictive maintenance: Algorithms, tools, and applications”. *Engineering Applications of Artificial Intelligence* 98:104282.

Zheng, Xudong, Di Wu, Shanqi Li, Yulin Li ja Liang Li. 2021. “A novel explainable artificial intelligence model for production scheduling optimization in smart manufacturing”. *Journal of Intelligent Manufacturing* 32:1203–1216.

Zhou, Yitong, Bin Liu ja Zonghai Yu. 2020. “Predictive control with deep learning for batch processes with high variability”. *Computers and Chemical Engineering* 139:106909.