International Journal of Production Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tprs20

Human-centric artificial intelligence architecture for industry 5.0 applications

Jože M. Rožanec, Inna Novalija, Patrik Zajec, Klemen Kenda, Hooman Tavakoli Ghinani, Sungho Suh, Entso Veliou, Dimitrios Papamartzivanos, Thanassis Giannetsos, Sofia Anna Menesidou, Ruben Alonso, Nino Cauli, Antonello Meloni, Diego Reforgiato Recupero, Dimosthenis Kyriazis, Georgios Sofianidis, Spyros Theodoropoulos, Blaž Fortuna, Dunja Mladenić & John Soldatos

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2022.2138611

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Human-centric artificial intelligence architecture for industry 5.0 applications

Jože M. Rožanec[a,b,c], Inna Novalija[b], Patrik Zajec[a,b], Klemen Kenda[a,b,c], Hooman Tavakoli Ghinani[d], Sungho Suh[d], Entso Veliou[e], Dimitrios Papamartzivanos[f], Thanassis Giannetsos[f], Sofia Anna Menesidou[f], Ruben Alonso[g], Nino Cauli[g], Antonello Meloni[h], Diego Reforgiato Recupero[f,h], Dimosthenis Kyriazis[i], Georgios Sofianidis[i], Spyros Theodoropoulos[i,j], Blaž Fortuna, Dunja Mladenić[b] and John Soldatos[k]

[a]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia; [b]Jožef Stefan Institute, Ljubljana, Slovenia; [c]Qlector d.o.o., Ljubljana, Slovenia; [d]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany; [e]Department of Informatics and Computer Engineering, University of West Attica, Egaleo, Athens, Greece; [f]Digital Security & Trusted Computing Group, Ubitech Ltd., Athens, Greece; [g]R2M Solution Srl, Pavia, Italy; [h]Department of Computer Science, University of Cagliari, Cagliari, Italy; [i]Department of Digital Systems, University of Piraeus, Piraeus, Greece; [j]Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece; [k]Intrasoft International, Peania, Greece

ABSTRACT

Human-centricity is the core value behind the evolution of manufacturing towards Industry 5.0. Nevertheless, there is a lack of architecture that considers safety, trustworthiness, and human-centricity at its core. Therefore, we propose an architecture that integrates Artificial Intelligence (Active Learning, Forecasting, Explainable Artificial Intelligence), simulated reality, decision-making, and users' feedback, focussing on synergies between humans and machines. Furthermore, we align the proposed architecture with the Big Data Value Association Reference Architecture Model. Finally, we validate it on three use cases from real-world case studies.

ARTICLE HISTORY
Received 26 March 2022
Accepted 29 September 2022

KEYWORDS
Smart manufacturing; explainable artificial intelligence (XAI); active learning; demand forecasting; quality inspection

1. Introduction

The development of new technologies and their increasing democratisation have enabled the digitalisation of the manufacturing domain. Furthermore, the digitalisation and application of cutting-edge technologies have been fostered by many governments through national programs (e.g.*Industrie 4.0*, *Advanced Manufacturing Partnership*, or *Made in China 2025* and others) as means to revitalise the industry and face societal changes such as the increasingly aging population (Kuo, Shyu, and Ding 2019). The Industry 4.0 paradigm, first presented in the Hannover Industrial Fair in 2011, aims to leverage the latest technologies (e.g. Internet of Things, cloud computing, and artificial intelligence, among others), and paradigms (e.g.Cyber-Physical Systems (Lee 2006; Rajkumar et al. 2010; Xu and Duan 2019), and Digital Twins (Grieves 2015; Grieves and Vickers 2017; Tao and Zhang 2017)), to enable innovative manufacturing functionalities. Research regarding Industry 4.0 aims to devise means to realise mass customisation, predictive maintenance, zero-defect manufacturing, and smart product lifecycles management (Soldatos et al. 2019; Lim, Zheng, and Chen 2020)). Furthermore, there is an increasing awareness regarding the complementarity of skills between humans and machines and the opportunity to foster human-centric solutions, which is one of the core principles in the emerging Industry 5.0.

When speaking about artificial intelligence, we must distinguish between General Artificial Intelligence (which aims to create machines capable of thinking and reasoning like humans), and Narrow Artificial Intelligence (which aims to solve specific problems automating specific and repetitive tasks). Industrial Artificial Intelligence can be considered a specific case of Narrow Artificial Intelligence, applied to industry. Artificial intelligence is being increasingly adopted in manufacturing, leading to work design, responsibilities, and dynamics changes. Artificial intelligence techniques can provide insights and partially or even fully automate specific tasks, while human input or decision-making remains critical in some instances. When insights are required for decision-making, it is of utmost importance to understand the models' rationale and inner workings to ensure the models can be trusted and responsible

CONTACT Jože M. Rožanec joze.rozanec@ijs.si Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana 1000, Slovenia; Jožef Stefan Institute, Jamova 39, Ljubljana 1000, Slovenia; Qlector d.o.o., Rovšnikova 7, Ljubljana 1000, Slovenia

decisions made based on their outcomes (Ahmed, Jeon, and Piccialli 2022). Among the human-machine collaboration approaches, we find mutual learning, which considers learning to be a bidirectional process and a reciprocal collaboration between humans and machines when performing shared tasks (Ansari et al. 2018; Ansari, Erol, and Sihn 2018). Another possible approach is active learning, which assumes the machine learning model can learn from carefully selected data and leverage the knowledge and expertise of a human expert in a human-in-the-loop system. Furthermore, interactions between humans and machines can be enhanced by developing proper interfaces. For example, spoken dialog systems and voice-user interfaces attempt to do so by mimicking human conversations (McTear, Callejas, and Griol 2016; Jentzsch, Höhn, and Hochgeschwender 2019). To ensure artificial intelligence systems are human-centred, much research is being invested in ensuring such systems remain secure and comply with ethical principles (Shneiderman 2020). Security challenges involve multiple aspects, such as ensuring that the integration between business and industrial networks remains secure (Ani et al. 2017), and data-related aspects critical to artificial intelligence, such as protecting the Confidentiality, Integrity, and Availability (CIA) of data (Wu et al. 2018; Mahesh et al. 2020). Compliance with ethical principles can be realised through three building blocks: (i) provide a framework of ethical values to support, underwrite and motivate (SUM) responsible data design and the use of the ecosystem, (ii) a set of actionable principles to ensure fairness, accountability, sustainability and transparency (FAST principles), and (iii) a process-based governance framework (PGB framework) to operationalise (i) and (ii) (Leslie 2019).

The richness of manufacturing use cases and the high level of shared challenges require standards to build a common ground to ensure the components' interoperability and that best practices are applied to develop and integrate them. Furthermore, there is a need to develop a unified architecture based on standards and reference architecture components to tackle the challenges described above. Among the reference architectures relevant to the field of manufacturing, we find the Reference Architecture Model for Industry 4.0 (RAMI 4.0-presents the main building blocks of Industry 4.0 systems) (Schweichhart 2016), the Industrial Internet Reference Architecture (IIRA – specifies a common architecture framework for interoperable IoT systems) (IIR n.d.), the Industrial Internet Security Framework (IISF) (IIS 2016), and the Big Data Value Association (BDVA) Reference Architecture (Reñones, Carbonare, and Gusmeroli 2018). While the RAMI 4.0 and IIRA do not address the security and safety aspects, these are addressed by the ISSF and BDVA. Furthermore, the BDVA reference architecture provides structure and guidelines to structure big data software, foster data sharing, and enable the use of artificial intelligence in its components while ensuring compliance with standards. The requirements that emerge from the application of artificial intelligence have also crystallized in specific architectures. Spelt, Knee, and Glover (1991) describes how machine learning models and expert systems can be combined, to leverage their complimentary strengths. Xu et al. (2020) surveys multiple authors on how they realise edge intelligence, considering three key components: data, model, and computation. In the same line, Yang et al. (2019) surveys authors on federated learning, considering horizontal and vertical federated learning, and federated transfer learning, while describing architectures that implement them. A slightly different approach is considered by Wan et al. (2020), who describe an architecture combining cloud computing, edge computing, and local computing paradigms. The authors consider four major architecture components: smart devices, smart interaction, an artificial intelligence layer, and smart services.

In this work, we evolve and detail the architecture proposed in Rožanec et al.,"STARdom: An Architecture for Trusted and Secure Human-Centered Manufacturing Systems," (2021), which addresses the void of an architecture specification that tackles the needs of trusted and secure artificial intelligence systems in manufacturing, seeking human-machine synergies by considering humans-in-the-loop. The human at the centre of the manufacturing evolution represents the core of the evolution towards Industry 5.0 (Nahavandi 2019; EC2 2020). Furthermore, we map the proposed architecture modules to the BDVA reference architecture and ISSF framework to ensure compatibility and show how their complimentary views can coincide in a single solution.

The rest of this paper is structured as follows: Section 2 presents related work, and Section 3 introduces three values-based principles and describes the proposed architecture. Next, Section 4 describes the validating use cases, while Section 5 describes the experiments we conducted and the results we obtained. Finally, in Section 6, we provide our conclusions and outline future work.

2. Related work

2.1. *Industry 5.0*

Industry 4.0 was introduced at the Hannover Trade Fair in 2011, aiming to introduce new technologies into manufacturing with the purpose of achieving high levels of operational efficiency and productivity (Sanchez, Exposito, and Aguilar 2020). While technology is emphasised as a means toward greater efficiency and productivity to enhance competitiveness in the global market, some emphasis have been placed on using such technology to reach certain level of human-centricity through the concept of Operator 4.0. Operators 4.0 are operators who will be assisted by systems providing relief from physical and mental stress, without compromising the production objectives (Romero et al. 2016; Romero, Stahre, and Taisch 2020; Kaasinen et al. 2020). Industry 5.0 is envisioned as a co-existing industrial revolution (Xu et al. 2021), for which two visions have emerged: (i) one that refers to human-robot co-working, and (ii) a second one as a bioeconomy where renewable biological resources are used to transform existing industries (Demir, Döven, and Sezen 2019). In this work, we focus on Industry 5.0 as a value-driven manufacturing paradigm and revolution that highlights the importance of research and innovation to support the industry while placing the well-being of the worker at the centre of the production process (Xu et al. 2021). Such a revolution must attempt to satisfy the needs placed in the Industrial Human Needs Pyramid, which range from workplace safety to the development of a trustworthy relationship between humans and machines that enables the highest level of self-esteem and self-actualisation, realising and fulfilling their potential (Lu et al. 2022). It aims to intertwine machines and humans in a synergistic collaboration to increase productivity in the manufacturing industry while retaining human workers. Furthermore, it seeks to develop means that enable humans to unleash their critical thinking, creativity, and domain knowledge. At the same time, the machines can be trusted to autonomously assist on repetitive tasks with high efficiency, anticipating the goals and expectations of the human operator, and leading to reduced waste and costs (Nahavandi 2019; Demir, Döven, and Sezen 2019; Maddikunta et al. 2022). Such communication and collaborative intelligence enable the development of trustworthy coevolutionary relationships between humans and machines. To foster the development of trustworthy coevolutionary relationships, interfaces must consider the employee's characteristics (e.g. age, gender, and level of education, among others) and the organisational goals. One example of collaboration between humans and machines is realised with cobots, where the cobots share the same physical space, sense and understand the human presence, and can perform tasks either independently, simultaneously, sequentially, or in a supportive way (El Zaatari et al. 2019).

In order to realise the Industry 5.0 vision, the focus must be shifted from individual technologies to a systematic approach rethinking how to (a) combine the strengths of humans and machines, (b) create digital twins of entire systems, and (c) widespread use artificial intelligence, with a particular emphasis generation of actionable items for humans. While research regarding Industry 5.0 is incipient, it has been formally encouraged by the European Commission through a formal document released back in 2021 (Ind n.d.).

2.2. Considering standards and regulations

In order to realise the vision laid out for Industry 5.0, constraints and directions imposed by existing regulations must be considered. Furthermore, standards should be taken into account to ensure that the fundamental blocks can be universally understood and adopted to achieve compatibility and interoperability.

Cybersecurity is considered a transversal concern in the architecture presented in this work. Among the standards and regulations that relate to it we must mention the ISO 27000 family of standards (ISO n.d.), the USA Cybersecurity Information Sharing Act (CISA) (CIS n.d.), the EU Cybersecurity Act (cyb n.d.), and the EU Network and Information Security Directive II (NIS II) (NIS n.d.). The ISO 27000 standards defined a common vocabulary and provided an overview of information security management systems. The NIS II directive aimed to force certain entities and sectors of the European Union to take measures to increase the overall cybersecurity level in Europe. The European Union Cybersecurity Act provided complementary legislation by establishing a cybersecurity certification framework for products and services and granted a permanent mandate to the EU agency for cybersecurity (ENISA) to inform the public regarding certification schemas and issue the corresponding certificates. Finally, the CISA established a legal ground for information sharing between the USA government agencies and non-government entities for cyberattack investigations.

When considering data management, much emphasis is being put on privacy. The General Data Protection Regulation (GDPR) (GDP n.d.), ePrivacy directive (ePr n.d.), or Data Governance Act (dat n.d.), issued by the European Union, are relevant when managing and sharing data, especially personal or sensitive data. The GDPR establishes a legal framework setting guidelines to process and collect personal information of persons living in the European Union. The ePrivacy directive regulates data protection and privacy, emphasising issues related to confidentiality of information, treatment of spam, cookies, and traffic data. Finally, the Data Governance Act promotes wider re-use of data, using secure processing environments, data anonymization techniques (e.g. differential privacy), and synthetic data creation; and establishes a licensing regime for data intermediaries between data holders and data users. While these regulations and directives must be considered, we provide no systemic solution from an architectural point of view.

Finally, given the increasing adoption of artificial intelligence, a legislative effort is being made to regulate its use. For example, the Artificial Intelligence Act (AIA n.d.a), issued in the European Union, was the first law of this kind issued by a significant regulator worldwide. The law categorises artificial intelligence applications into three risk categories: (a) unacceptable risk (e.g. social scoring systems), which are banned, (b) high-risk (e.g. resume scanning applications), which are subject to specific legal requirements, and (c) applications that do not fall into categories (a) and (b), which remain unregulated. Another example is a law issued by the Federative Republic of Brazil (AIA n.d.b), which establishes the principles, obligations, rights, and governance instruments regarding the use of artificial intelligence.

While the abovementioned list is not exhaustive, it provides a high-level view of the main concerns and topics that must be considered.

2.3. Enabling technologies

In order to realise a human-centric artificial intelligence architecture for Industry 5.0 applications, a set of technologies that enable a human-centric approach we consider a set of technologies must be taken into account. We consider five of them related to the field of artificial intelligence: (i) active learning, (ii) explainable artificial intelligence, (iii) simulated reality, (iv) conversational interfaces, and (v) security. Below we introduce some related work regarding each of them, and in Section 3 describe the corresponding architecture building blocks.

2.3.1. Active learning

The adoption of artificial intelligence in manufacturing and the complementarity of the machine and human capabilities is reshaping jobs, and human-machine cooperation opportunities are emerging. One way to realise such human-machine cooperation is through the Active Learning paradigm, which considers an artificial intelligence model can be improved by carefully selecting a small number of data instances to satisfy a learning objective (Settles 2009). Active Learning is built upon three assumptions: (i) the learner (artificial intelligence model) can learn by asking questions (e.g. request a target variable's data), (ii) there is an abundance of questions that can be asked (e.g. data, either gathered or synthetically created, without a target value), and (iii) there is a constrained capacity to answer such questions (and therefore, the questions must be carefully selected) (Elahi, Ricci, and Rubens 2016). Therefore, applied research is focussed on how to structure use case solutions so that through a human-in-the-loop, artificial intelligence models can benefit from human expertise to make decisions and provide valuable input, which is later used to enhance the models (Kumar and Gupta 2020; Schröder and Niekler 2020; Budd, Robinson, and Kainz 2021).

We discriminate between data obtained from real sources and synthetic data (created through some procedure) regarding the source of the data. Synthetic data is frequently used to enlarge the existing data or to generate instances that satisfy specific requirements when similar data is expensive to obtain. While many techniques and heuristics have been applied in the past to generate synthetic data, the use of Generative Adversarial Networks (GANs) has shown promising results and been intensely researched Zhu and Bento (2017), Mahapatra et al. (2018), Sinha, Ebrahimi, and Darrell (2019), Mayer and Timofte (2020). Strategies related to data selection are conditioned by how data is generated and served. If the data is stored, data instances can be scanned and compared, and some latency can be tolerated to make a decision. On the other hand, decisions must be made at low latency in a streaming setting, and the knowledge is constrained to previously seen instances. Data selection approaches must consider informativeness (quantifying the uncertainty associated to a given instance, or the expected model change), representativeness (number of samples similar to the target sample), or diversity criteria (selected samples scatter across the whole input space) (Wu 2018). Popular approaches for classification problems are the random sampling, query-by-committee (Seung, Opper, and Sompolinsky 1992), minimisation of the Fisher information ratio (Padmanabhan et al. 2014), or hinted sampling with Support Vector Machines (Li, Ferng, and Lin 2015).

Active learning has been applied to several manufacturing use cases. Nevertheless, applied research in the manufacturing sector remains scarce (Samsonov et al. 2019; Meng et al. 2020), but its relevance increases along with the proliferation of digital data and democratisation of artificial intelligence. In the scientific literature, authors report using Active Learning to tackle quality control, predictive modelling, and demand forecasting. For example, active learning for quality control was applied to predict the local displacement between two layers on a chip (Dai et al. 2018) or gather users' input in visual quality inspection of printed company logos on the manufactured products (Trajkova et al. 2021). In predictive modelling, it was applied in the aerospace industry to assist a model in predicting the shape control of a composite fuselage (Yue et al. 2020). Finally, in the demand forecasting use case, the authors explored using active learning to recommend media news and broaden the logisticians' understanding of the domain while informing relevant events that could affect the demand, to reach better decisions (Zajec et al. 2021). Regardless of the successful application in several manufacturing use cases, active learning is not widely adopted in manufacturing and could be applied to enhance cybersecurity capabilities and fatigue monitoring systems (Li et al. 2019), among others.

2.3.2. Explainable artificial intelligence

While artificial intelligence was applied to manufacturing problems in the past (Bullers, Nof, and Whinston 1980), it has become increasingly common to rely on artificial intelligence models to automate certain tasks and provide data-based insights (Chien et al. 2020). When human decision-making relies on artificial intelligence models' outcomes, enough information regarding the models' rationale for such a forecast must be provided (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). Such information enables the user to assess the trustworthiness and soundness of the provided forecast and therefore ensure decisions are made responsibly (Das and Rad 2020). Research on techniques and approaches that convey information regarding the rationale behind the artificial intelligence models, or an approximation to it, and how such information is best presented to the users, is done in a sub-field of artificial intelligence, known as eXplainable Artificial Intelligence (XAI) (Henin and Métayer 2021). Such approaches can be classified according to different taxonomies. Among them, there is consensus that artificial intelligence models can be considered either *white-box models* (inherently interpretable models), or *black-box models* (models that remain opaque to the users) (Loyola-Gonzalez 2019). Regarding the characteristics of the explanation, Angelov et al. (2021) divide XAI methods into four groups, considering whether (i) the explanations are provided at a local (for a specific forecast) or global (for the whole model) level, (ii) the models are transparent or opaque to the users, (iii) the explainability techniques are model-specific or model-agnostic, and (iv) the explanations are conveyed through visualisations, surrogate models or taking into account features relevance.

The scientific literature reports an increasing amount of use cases where explainable artificial intelligence is applied. Meister et al. (2021) applied deep learning models to automate defect detection on composite components built with a fibre layup procedure. Furthermore, the authors explored using three Explainable Artificial Intelligence techniques (Smoothed Integrated Gradients (Sundararajan, Taly, and Yan 2017), Guided Gradient Class Activation Mapping (Shrikumar, Greenside, and Kundaje 2017) and DeepSHAP (Selvaraju et al. 2017)), to understand whether the model has learned and thus can be trusted that it will behave robustly. Senoner, Netland, and Feuerriegel (2021) developed an approach to creating insights on how production parameters contribute to the process quality based on the estimated features' relevance to the forecast estimated with the Shapley additive explanations technique (Lundberg and Lee 2017). Finally, Serradilla et al. (2020) implemented multiple machine learning regression models to estimate the remaining life of industrial machinery and resorted to the Local Interpretable Model-Agnostic Explanations technique (Ribeiro, Singh, and Guestrin 2016) to identify relevant predictor variables for individual and overall estimations. Given research had little consideration for the complexity of interactions between humans and their environment in manufacturing, much can be done to develop XAI approaches that consider anthropometrics, physiological and psychological states, and motivations to not only provide better explanations, but also enhance the workers' self-esteem and help them towards their self-actualisation (Lu et al. 2022).

2.3.3. Simulated reality

Under simulated reality, we understand any program or process that can generate data resembling a particular aspect of reality. Such a process can take inputs and produce outputs, such as synthetic data or outcomes that reflect different scenarios or process changes.

Machine learning models can solve complex tasks only if provided with data. Acquiring high-quality data can be a complex and expensive endeavour: lack of examples concerning faulty items for defect detection systems, wearing down and damaging a robotic system during data collection, or human errors when labelling the data are just some examples. Synthetic data is envisioned as a solution to such challenges. Much research is invested in making it easy to generate while avoiding annotation pitfalls, ethical and practical concerns and promising an unlimited supply of data (de Melo et al. 2022). In addition, much research was invested in the past regarding synthetic data generation to cope with imbalanced datasets.

Nevertheless, the development of GANs opened a new research frontier, leading to promising results (Creswell et al. 2018). GANs consist of two networks: a generator (trained to map some noise input into a synthetic data sample) and a discriminator (that, given two examples, tries to distinguish the real from the synthetic one). This way, the generator learns to generate higher-quality samples based on the discriminator's feedback. While they were first applied to images (Goodfellow et al. 2014), models have been developed to enhance the quality of synthetic images and to apply them to other types of data, too (Patki, Wedge, and Veeramachaneni 2016; Xu et al. 2019).

Simulated reality can be considered a key component of Reinforcement Learning. The reinforcement learning agent can explore an approximation of the real world through the simulator and learn efficient policies safely and without costly interactions with the world. Furthermore, by envisioning the consequences of an action, simulations can help to validate desired outcomes in a real-world setting (Amodei et al. 2016).

Simulated reality has been applied in a wide range of manufacturing use cases. Neural Style Transfer (Wei et al. 2020) has been successfully used to generate synthetic samples by fusing defect snippets with images of non-defective manufactured pieces. Such images can be later used to enhance the algorithm's predictive capacity. Simulators have been widely applied to train Reinforcement Learning models in manufacturing. Mahadevan and Theocharous (1998) used them to simulate a production process and let the RL algorithm learn to maximise the throughput in assembly lines, regardless of the failures that can take place during the manufacturing process. Oliff et al. (2020) simulated human operators' performance under different circumstances (fatigue, shift, day of week) so that behavioural policies could be learned for robotic operators and ensure they provided an adequate response to the operators' performance variations. Finally, Johannink et al. (2019) used Reinforcement Learning to learn robot control and evaluated their approach in both real-world and simulated environments. Bridging the gap from simulated to real-world knowledge remains a challenge.

2.3.4. Intention recognition in manufacturing lines

The development of Industrial Internet of Things (IIoT) technologies and the availability of low-cost wearable sensors have enabled access to big data and the utilisation of the sensors for the manufacturing industry (Jeschke et al. 2017). Recently, various deep learning-based methods have been proposed to learn valuable information from big data and improve the effectiveness and safety of the manufacturing lines. In particular, worker's activity and intention recognition can be used for quantification and evaluation of the worker's performance and safety in the manufacturing lines. In addition, with the introduction of autonomous robots for effective manufacturing, efficient collaboration between robots and workers and the safety of workers are also becoming increasingly important.

Thanks to the miniaturisation and reduction of costs, the adoption of wearable sensors has also been growing in the industrial context to investigate workers' conditions and well-being. Worker's health is a key factor in determining the organisation's long-term competitiveness, and it is also directly related to production efficiency. The cumulative effect of positive impacts on the human factor brings economic benefit through productivity increase, scrap reduction, and decreased absenteeism. Few research works have been recently developed, where workers' physiological data are used to infer the insurgence of phenomena such as fatigue (Maman et al. 2017, 2020) and mental stress (Villani et al. 2020), which have a relevant impact on process performance. Another research line adopted eye-trackers, together with wearables and cameras, to estimate workers' attention and stress levels, understand assembly sequence, and identify the criticalities in the product design affecting the assembly process (Peruzzini, Grandi, and Pellicciari 2017).

Human-robot collaboration in open workspaces is realised through cobots, for which additional mechanisms must be developed to ensure the workers' safety, given that humans can be easily hurt in case of contact due to the large workload or moving mass (Bi et al. 2021). To realise such collaboration, human movement prediction is of utmost importance to avoid collisions and minimise injuries caused by such collisions (Buerkle et al. 2021). Many researchers have developed various activity and intention recognition methods by using machine learning-based algorithms and wearable sensors. Malaisé et al. (2018) proposed activity recognition with Hidden Markov Model (HMM)-based models and multiple wearable sensors for a manufacturing scenario. Tao et al. (2018) proposed an activity recognition method using Inertial Measurement Unit (IMU) and surface electromyography (sEMG) signals obtained from a Myo armband. They combined the IMU and sEMG signals and fed them into convolutional neural networks (CNN) for worker activity classification. Kang and Kim (2018) developed a motion recognition system for worker safety in manufacturing work cells, leveraging a vision system. Forkan et al. (2019) introduced an IIoT solution for monitoring, evaluating, and improving worker and related plant productivity based on worker activity recognition using a distributed platform and wearable sensors. Günther, Kärcher, and Bauernhansl (2019) proposed a human activity recognition approach to detect assembly processes in a production environment by tracking activities performed with tools. Tao, Leu, and Yin (2020) proposed a multi-modal activity recognition method by leveraging information from different wearable sensors and visual cameras. Finally, Buerkle et al. (2021) proposed using a mobile electroencephalogram and machine learning models to forecast operators' movements based on two neurophysiological phenomena that can be measured before the actual movement takes place: (a) a weak signal that occurs about 1.5s before a movement, and (b) a strong signal that occurs about 0.5s before any movement.

2.3.5. Conversational interfaces

Spoken dialog systems and conversational multimodal interfaces leverage artificial intelligence and can reduce friction and enhance human-machine interactions (Klopfenstein et al. 2017; Vajpai and Bora 2016; Maurtua et al. 2017) by approximating a human conversation. However, in practice, conversational interfaces mostly act as the first level of support and cannot offer much help as a knowledgeable human. They can be classified into three broad categories: (i) basic-bots, (ii) text-based assistants, and (iii) voice-based assistants. While basic bots have a simple design and allow basic commands, the text-based assistants (also known as chatbots) can interpret users' text and enable more complex interactions. Both cases require speech-to-text and text-to-speech technologies, especially if verbal interaction with the conversational interface is supported. Many tools have been developed to support the aforementioned functionalities. Among them, we find the Web Speech API,[1] which can be configured to recognise expressions based on a finite set of options defined through a grammar.[2] Most advanced version of conversational interfaces are represented by voice assistants, such as the Google Assistant,[3] Apple's Siri,[4] Microsoft's Cortana,[5] or Amazon's Alexa.[6] They can be integrated into multiple devices and environments through publicly available application development interfaces (APIs), enabling new business opportunities (Erol et al. 2018). Given voice interfaces can place unnecessary constraints in some use cases, they can be complemented following a multimodal approach (Kouroupetroglou et al. 2017).

A few implementations were described in an industrial setting. Silaghi et al. (2014) researched the use of voice commands in noisy industrial environments, showing that noises can be attenuated with adequate noise filtering techniques. Wellsandta et al. (2020) developed an intelligent digital assistant that connects multiple information systems to support maintenance staff on their tasks regarding operative maintenance. They exploit the fact that access to the voice assistant's functions is hand free and that voice operation is usually faster than writing. Afanasev et al. (2019) developed a method to integrate a voice assistant and modular cyber-physical production system, where the operator could request help to find out-of-sight equipment or get specific sensor readings. Finally, Li et al. (2022) developed a virtual assistant to assist workers on dangerous and challenging manufacturing tasks, controlling industrial mobile manipulators that combine robotic arms with mobile platforms used on shop floors. The assistant uses a language service to extract keywords, recognise intent, and ground knowledge based on a knowledge graph. Furthermore, conversation strategies and response templates are used to ensure the assistant can respond in different ways, event when the same question is asked repeatedly.

2.3.6. Security

While the next generation of manufacturing aims to incorporate a wide variety of technologies to enable more efficient manufacturing and product lifecycles, at the same time, the attack surface increases, and new threats against confidentiality, integrity, and availability are introduced (Chhetri et al. 2017, 2018). These are exacerbated by the existence of a large number of legacy equipment, the lack of patching and continuous updates on the industrial equipment and infrastructure, and the fact that cyberattacks on cyber-physical systems achieve a physical dimension, which can affect human safety (Elhabashy, Wells, and Camelio 2019). Artificial intelligence has proved its efficiency for threat intelligence sensing, intrusion detection, and malware classification, while how to ensure a model itself was not compromised remains a topic of major research (Conti, Dargahi, and Dehghantanha 2018; Li 2018).

Multiple cyberattack case studies in manufacturing have been analysed in the scientific literature. Zeltmann et al. (2016) studied how embedded defects during additive manufacturing can compromise the quality of products without being detected during the quality inspection procedure. The attacks can be fulfilled by either compromising the CAD files or the G-codes. Ranabhat et al. (2019) demonstrated sabotage attacks on carbon fibre reinforced polymer by identifying critical force bearing plies and rotating them. Therefore, the resulting compromised design file provides a product specification that renders the manufactured product useless. Finally, Liu et al. (2020) describe a data poisoning attack through which the resulting machine learning model is not able to detect hotspots in integrated circuit boards.

In order to mitigate the threats mentioned above, steps must be taken to prevent the attacks, detect their effects, and respond, neutralising them and mitigating their consequences (Elhabashy, Wells, and Camelio 2019). On the prevention side, Wegner, Graham, and Ribble (2017) advocated for the extensive use of authentication and authorisation in the manufacturing setting.

To that end, the authors proposed using asymmetric encryption keys to enable encrypted communications, a *comptroller* (software authorising actions in the manufacturing network, and encryption key provider) to ensure the input data is encoded and handled to a Manufacturing Security Enforcement Device, which then ensures the integrity of the transmitted data. A security framework for cyber-physical systems was proposed by Wu and Moon (2018), defining five steps: Define, Audit, Correlate, Disclose, Improve (DACDI). *Define* refers to the scope of work, considering the architecture, the attack surface, vector, impact, target, and consequence, and the audit material. *Audit* relates to the process of collecting cyber and physical data required for intrusion detection. Artificial intelligence is being increasingly used in this regard, leveraging paradigms such as active learning to combine machine and human strengths (Klein et al. 2022). *Correlate* attempts to establish relationships between cyber and physical data considering time and production sequences, the scale and duration of the attack, and therefore reduce the number of false positives and assist in identifying the root causes of alerts. *Disclose* establishes a set of methods used to stop the intrusion as quickly as possible. Finally, *Improve* aims to incrementally enhance the security policies to avoid similar issues in the future. Another approach was proposed by Bayanifar and Kühnle (2017), who described an agent-based system capable of real-time supervision, control, and autonomous decision-making to defend against or mitigate measured risks.

3. Safe, trusted, and human-centred architecture

3.1. Architecture values-based principles

The proposed architecture is designed to comply with three key desired characteristics for the manufacturing environments in Industry 5.0: safety, trustworthiness, and human centricity. *Safety* is defined as the condition of being protected from danger, risk, or injury. In a manufacturing setting, *safety* can refer either to *product safety* (quality of a product and its utilisation without risk), or *human safety* (accident prevention in work situations), and the injuries usually relate to occupational accidents, or bad ergonomics (Wilson-Donnelly et al. 2005; Sadeghi et al. 2016). Trustworthiness is understood as the quality of deserving trust. In the context of manufacturing systems, it can be defined as a composite of transparency, reliability, availability, safety, and integrity (Yu et al. 2017). In manufacturing, trustworthiness refers to the ability of a manufacturing system to perform as expected, even in the face of anomalous events (e.g.cyberattacks), and whose inner workings are intelligible to the human persons who interact with them. Human-centricity in production systems refers to designs that put the human person at the centre of the production process, taking into account their competencies, needs, and desires, and expecting them to be in control of the work process while ensuring a healthy and interactive working environment (May et al. 2015). We consider *Safety* and *Trustworthiness* are critical to a human-centric approach, and therefore render them as supporting pillars of the *Human Centricity* values-based principle in Figure 1.

We depict the above-listed architecture value-based principles in Figure 1, and how do the building blocks, detailed in Section 2, relate to them. *Cybersecurity* is considered at the intersection of safety and trustworthiness since it ensures manufacturing systems and data are not disrupted through cybersecurity attacks (e.g. data poisoning or malware attacks). The *Worker Intention Recognition* is found at the intersection of safety and human centricity since it aims to track better and understand the human person to predict its intentions (e.g. movements) and adapt to the environment according to this information. *Explainable Artificial Intelligence* provides insights regarding the inner workings of artificial intelligence models and therefore contributes to the trustworthiness while being eminently human-centric. *Conversational Interfaces* and *Active Learning* place the human person at their centre, either by easing interactions between humans and machines or seeking synergies between their strengths to enhance Artificial Intelligence models' learning. Finally, *Standards and Regulations* are considered at the intersection of the three aforementioned value-based principles, given they organise and regulate aspects related to each of them.

3.2. Architecture for safe, trusted, and human-centric manufacturing systems

We propose a modular architecture for manufacturing systems, considering three core value-based principles: safety, trustworthiness, and human centricity. The proposed architecture complies with the BDVA reference architecture (see Figure 2) and considers cybersecurity a transversal concern, which can be implemented following guidelines from the IISF or ISO 27000, along with other security frameworks and standards. The cybersecurity layer transversal implements a *Security Policies Repository* and a *Policy Manager*. The *Security Policies Repository* associates risk-mitigation and cyber defense strategies to potential vulnerabilities and specific cyberattacks. The *Policy Manager*, on the other hand, configures

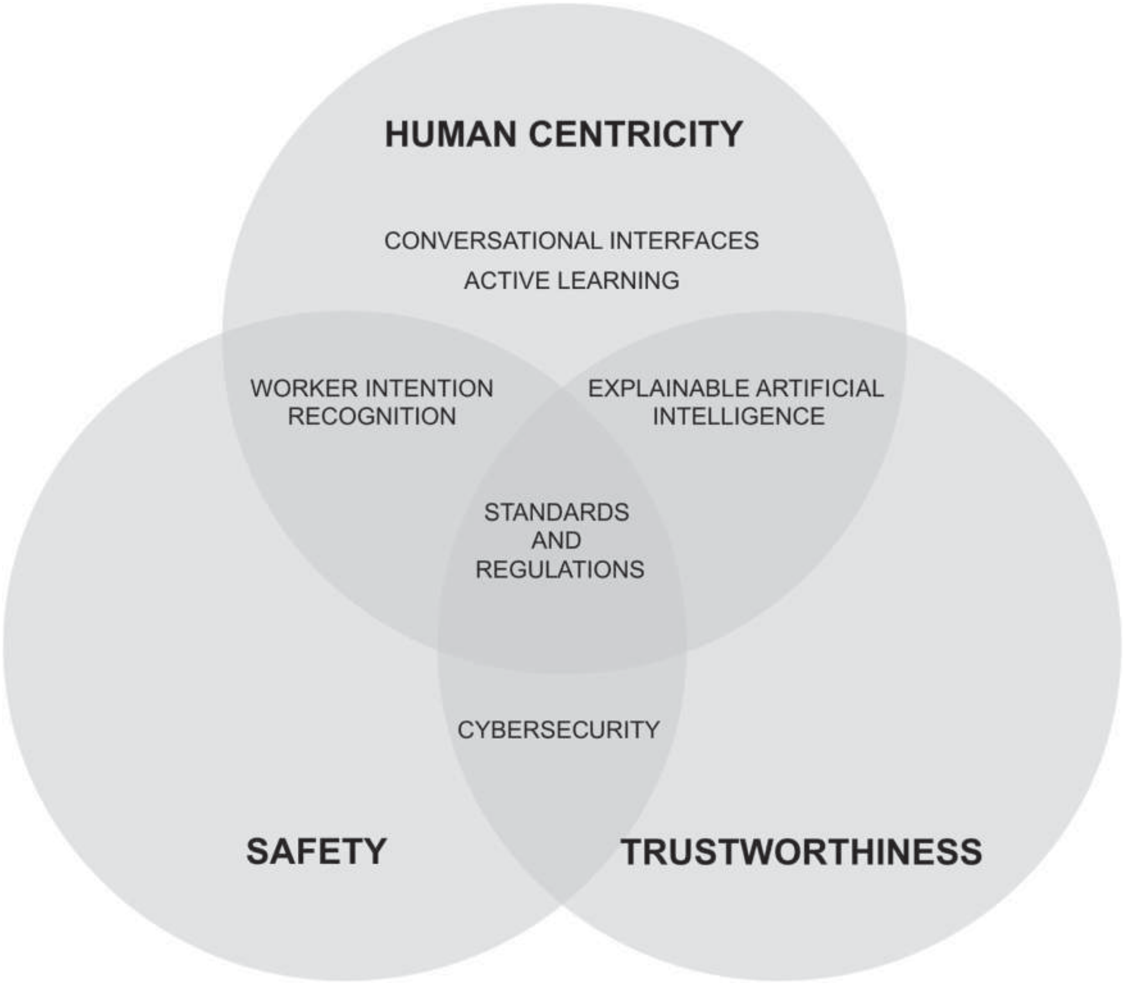
HUMAN CENTRICITY
CONVERSATIONAL INTERFACES
ACTIVE LEARNING
WORKER INTENTION RECOGNITION
EXPLAINABLE ARTIFICIAL INTELLIGENCE
STANDARDS AND REGULATIONS
CYBERSECURITY
SAFETY
TRUSTWORTHINESS

Figure 1. Intersection of architecture value-based principles, and architecture building blocks addressing them.

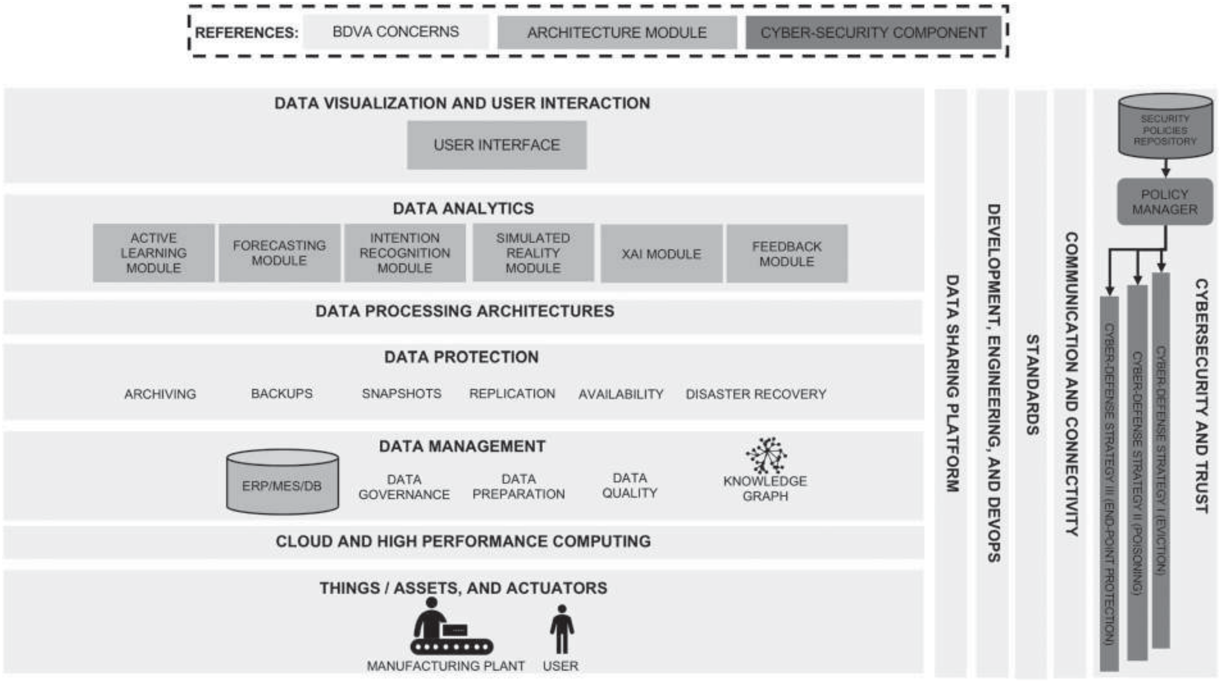

Figure 2. Proposed architecture contextualised within the BDVA reference architecture components.

security policies and ensures they are deployed, changing the security operations.

The architecture evolved organically from a set of use cases developed for two EU H2020 projects. It comprises the following modules, whose interaction is depicted in Figure 3:

- *Simulated Reality Module*: uses heuristics, statistical, and machine learning models to either create alternative scenarios or generate synthetic data. Synthetic data is frequently used to mitigate the lack of data, either by replacing expensive data gathering procedures or enriching the existing datasets. On the other hand, the simulated scenarios are frequently used in Reinforcement Learning problems to foster models' learning while avoiding the complexities of a real-world environment. Furthermore, simulations can also be used to project possible outcomes based on potential users' decisions. Such capability enables what-if scenarios, which can be used to inform better decision-making processes. The *Simulated Reality Module* provides synthetic data instances to the *Active Learning Module*, simulated scenarios to the *Decision-Making Module*, and simulation outcomes to the user (through a *User Interface*).
- *Forecasting Module*: provides forecasts for a wide range of manufacturing scenarios, leveraging artificial intelligence and statistical simulation models. The outcomes of such models depend on the goal to be solved (e.g. classification, regression, clustering, or ranking). While machine learning models require data to learn patterns and create inductive predictions, simulation models can predict future outcomes based on particular heuristic and configuration parameters that define the problem at hand. The *Forecasting Module* can receive inputs either from the storage or the *Active Learning Module*. At the same time, it provides forecasts to the user, the *Simulated Reality Module*, and the *XAI Module*. With the former, it can also share relevant information regarding the forecasting model to facilitate the creation of accurate explanations.
- *XAI Module*: is concerned with providing adequate explanations regarding artificial intelligence models and their forecasts. Such explanations aim to inform the user regarding the models' rationale behind a particular forecast and must be tailored to the users' profile to ensure the appropriate vocabulary, level of detail, and explanation type (e.g. feature ranking, counterfactual explanation, or contrastive explanation) is provided. Furthermore, the module must ensure that no sensitive information is exposed to users who must not have access to it. Finally, the explanations can be enriched with domain knowledge and information from complementary sources. Such enrichment can provide context to enhance users' understanding and, therefore, enable the user to evaluate the forecast and decision-making. The *XAI Module* provides input to the *Decision-Making Module* and explanations to the user.
- *Decision-Making Module*: is concerned with recommending decision-making options to the users. Envisioned as a recommender system, it can leverage expert knowledge and predictions obtained from inductive models and simulations and exploits it using heuristics and machine learning approaches. Given a particular context, it provides the user with the best possible decision-making options available to achieve the desired outcome. It receives input from the *XAI Module* and *Forecasting Module* and can retrieve expert knowledge encoded in the storage (e.g. a knowledge graph).
- *Active Learning Module*: implements a set of strategies to take actions on how data must be gathered to realise a learning objective. In supervised machine learning models, this is realised by selecting unlabelled data, which can lead to the best models' learning outcomes, and request labels to a human annotator. Another use case can be the data gathering that concerns a knowledge base enrichment. To that end, heuristics can be applied to detect missing facts and relationships ask for and store locally observed collective knowledge not captured by other means (Preece et al. 2015). The *Active Learning Module* interacts with the storage and the *Simulated Reality Module* to retrieve data, and the *Feedback Module* to collect answers to queries presented to the user.
- *Feedback Module*: collects feedback from users, which can be either explicit (a rating or an opinion) or implicit (the lack of feedback can be itself considered a signal) (Oard and Kim 1998). The feedback can refer to feedback regarding given predictions from the *Forecasting Module*, explanations provided by the *XAI Module*, or decision-making options recommended by the *Decision-Making Module*. It directly interacts with the *Active Learning Module* and the user (through the *User Interface*), and indirectly (registering and storing the feedback) with other modules' feedback functionality exposed to the user.
- *Intention Recognition Module*: predicts the user's movement trajectory based on artificial intelligence models and helps to decide whether the mobile robot should move faster, slower, or completely stop. The module receives sensor and video data. The data is captured by cameras in the manufacturing line or by sensors attached to the user's body. The recognised worker's activity and intention can be used by the

Forecasting Module and the *XAI Module* to decide the following action of the mobile robot in the manufacturing line.

- *User Interface*: enables users' multimodal interactions with the system, e.g. complementing the voice interactions with on-screen forms. Furthermore, it enables the machine to provide information to the user through audio, natural language, or other means such as visual information.

Interactions with the human persons are realised through a *User Interface*, while the data is stored in a *Storage*, which can be realised by different means (e.g. databases, filesystem, knowledge graph) based on the requirements of each module. The *User Interface* can be implemented taking into account multiple modalities. While the use of graphical user interfaces is most extended, there is increasing adoption of voice agents.

Based on the modules described above, multiple functionalities can be realised. While the *Storage* can store near data collected from the physical world, the *Simulated Reality Module* and the *Forecasting Module* can provide behaviour to Digital Twins mirroring humans (e.g. to monitor fatigue or emotional status), machines (e.g. for predictive maintenance), and manufacturing processes (e.g.supply chain and production). Furthermore, the *Forecasting Module* can be used to recognise and predict the workers' intention and expected movement trajectories. This information can then be used to adapt to the environment, e.g. by deciding whether autonomous mobile robots should move faster, slower, or completely stop. Finally, the *Active Learning Module* and *Explainable Artificial Intelligence Module* can be combined to create synergic relationships between humans and machines. While the *Active Learning Module* requires the human to provide expert knowledge to the machines and teach them, the *Explainable Artificial Intelligence Module* enables the humans to learn from machines.

4. Validating use cases

We propose three validating use cases (demand forecasting, quality inspection, and intention recognition), which align with the human-centricity, trustworthiness, and safety Industry 5.0 core values. Through the first two use cases, we aimed to realise the research gap highlighted by Lu et al. (2022), considering the interactions between human beings and their environment, developing a collaborative dynamic between humans and machines leveraging active learning and explainable artificial intelligence along with other relevant technologies. With the third use case, we aim to describe how intention recognition is considered in the proposed architecture to enable safe collaboration with cobots (Hentout et al. 2019). Finally, we evaluated the proposed architecture through the proposed use cases, assessing the internal validity (the results obtained can only be attributed to the manipulated variable) and external validity (which describes the generalizability of our architecture, demonstrated through three use cases) (Salkind and Rainwater 2006). It must be noted, that the components were implemented separately and not deployed into a productive environment.

4.1. Demand forecasting

Research regarding demand forecasting was performed with data provided by a European original equipment manufacturer targeting the global automotive industry market. Demand forecasting aims to estimate future customer demand over a defined period of time, using historical data and other sources of information. The ability to accurately forecast future demand allows to reduce operational inefficiencies (e.g. high stocks or stock shortages), which have a direct impact on goods produced in the supply chain and therefore can affect the brands' reputation (Brühl et al. 2009). Furthermore, insights into future demand enable better decision-making on various levels (e.g. regarding resources, workers, manufactured products, and logistics) (Thomopoulos 2015). While human forecasts are prone to multiple biases, the statistical and machine learning models can be used to learn patterns present in data obtained from multiple information sources to create accurate forecasts (Corredor, Ferrer, and Santamaria 2014; Hogarth and Makridakis 1981). Such models do not replace humans but provide a means to establish a synergic relationship. The model provides a forecast, and the user can make judgmental adjustments and make decisions based on them. Judgemental adjustments need to be done when some information is not available to the model, e.g.knowledge regarding future and extraordinary events that the model cannot capture from the existing information sources (Fildes and Goodwin 2021). Furthermore, it is recommended to record such forecast adjustments and the reasons behind them to be evaluated in retrospect. Such records can provide valuable input to improve the demand forecasting models. It must be noted that the demand forecasting models, regardless of their accuracy, must be considered tools to ease the planning duties. The planners are responsible for the decisions taken, regardless of the forecast outcomes.

Given that the planners hold responsibility for their decisions, the forecast must be complemented with insights regarding the models' rationale to enable responsible decision-making (Almada 2019; Wang, Xiong, and Olya 2020). Furthermore, an explanation regarding a

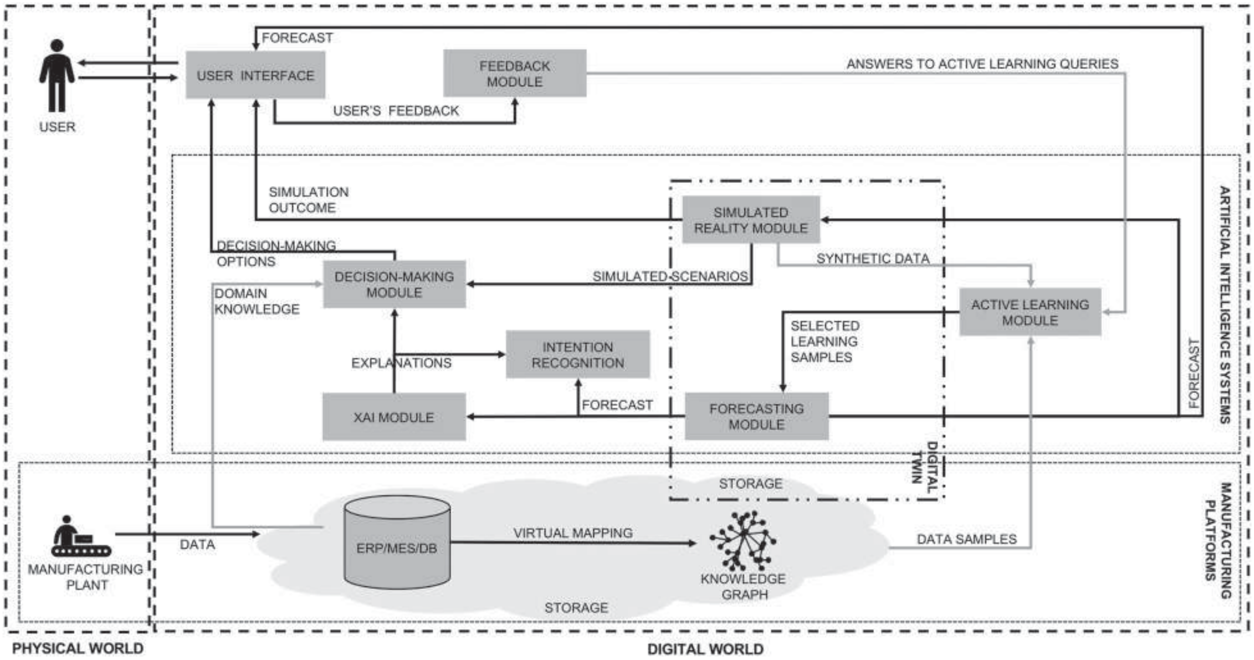

Figure 3. The proposed architecture modules, a storage layer, and their interactions. In addition, we distinguish (a) the physical and digital worlds, (b) manufacturing platforms, (c) artificial intelligence systems, and (d) digital twin capabilities.

particular forecast is sometimes legally required (Goodman and Flaxman 2017). Such insights can be either derived from the model or attained through specific techniques. The insights can be served as explanations to the users, tailoring them according to their purpose and the target stakeholders (Samek and Müller 2019). Such explanations must convey enough and relevant information, resemble a logical explanation (Doran, Schulz, and Besold 2017; Pedreschi et al. 2018), focus on aspects the stakeholder can act on to change an outcome (Verma, Dickerson, and Hines 2020; Keane et al. 2021), and ensure confidentiality is preserved (Rožanec, Fortuna, and Mladenić 2022). Good decision-making can require not only understanding the models' rationale but having relevant domain knowledge and contextual information at hand too (Arrieta et al. 2020; Rozanec 2021; Zajec et al. 2021). Furthermore, the users' perception of such explanations must be assessed to ensure their purpose is achieved (Mohseni, Zarei, and Ragan 2021; Sovrano and Vitali 2021).

4.2. Quality inspection

Research regarding quality inspection was performed with data provided by *Philips Consumer Lifestyle BV*. The dataset consisted of images focussed on the company's logo printed on manufactured shavers. The visual quality inspection aims to detect defective logo printing on the shavers, focussing on printing pads used for a wide range of products and logos. Currently, two types of defects are classified related to the printing quality of the logo on the shaver (see Figure 4): double printing and interrupted printing. Handling, inspecting and labelling the products can be addressed with robotics and artificial intelligence. It is estimated that automating the process mentioned above could speed it up by more than 40%, while the labelling effort can be alleviated by incorporating active learning (Trajkova et al. 2021; Rožanec et al.,"Streaming Machine Learning and Online Active Learning for Automated Visual Inspection," 2021).

When addressing the quality inspection use case, many challenges must be solved. We focussed on four of them: (i) automate the visual inspection, (ii) address data imbalance, (iii) understand the models' rationale behind the forecast, and (iv) enhance the manual revision process. While class imbalance is natural to quality inspection problems, its acuteness only increases over time: the greater the quality of the manufacturing process, the higher the scarcity of defective samples will be. The class imbalance has at least two implications. First, the increasing scarcity of defective parts affects the amount of data available to train defect detection models, affecting the capacity to improve them. Second, the higher the imbalance between good and defective products, the higher the risk that the inspection operators will not detect defective parts due to fatigue. It is thus necessary to devise mechanisms to mitigate such scenarios, ensuring high-quality standards are met while also enhancing the operators' work experience.

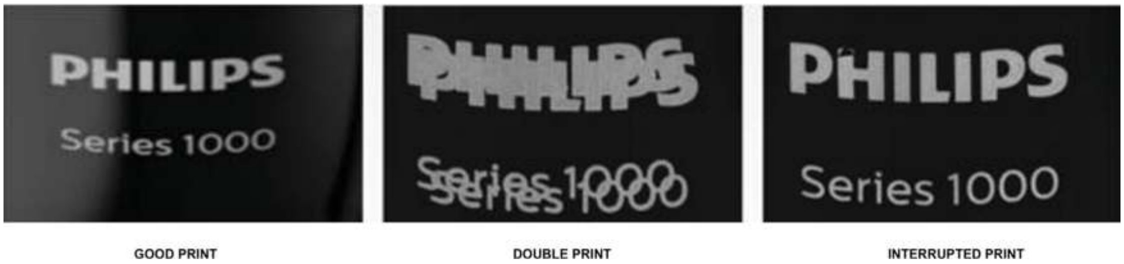

Figure 4. Samples of three types of images: (a) good (no defect), (b) double-print (defect), and (c) interrupted print (defect).

4.3. Human behaviour prediction and safe zone detection

SmartFactoryKL is an industry test-bed owned by the Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), which demonstrates the latest technologies from the industry domain building industry-standard demonstrators. One of the demanding aspects to our demonstrators would be adding safety considerations while working with an autonomous robot while combining artificial intelligence technologies. In this context, SmartFactoryKL aims to improve its demonstrator to have a high production rate while keeping workers and the hardware equipment safe using AI technologies.

In order to achieve this, three use cases were initialised for this test-bed:

(1) Human intention recognition
(2) Robot reconfiguration based on the dynamic layout
(3) Dynamic path planning using both aforementioned use cases.

The first use case aimed to detect human activities and predict their next actions, which then will be combined with robot navigation to create a safer environment. For this matter, DFKI initialised some workflows as the typical worker scenarios happening during regular daily work. For the training phase, the behaviours of more than ten participants were recorded, who were supposed to follow the same or similar flows. The recordings were made using wrist sensors which are then analysed in detail to detect different activities they are currently performing. The use case converts algorithm-based human activities into an artificial intelligence approach and predicts their following actions based on their daily activities.

The second use case aims to dynamically update the navigation route of the mobile robot by considering static and dynamic objects in the environment, which can be human and other (non-)moving objects. A desired outcome of the use case is the ability to ease the robot reconfiguration given the environment layout (including the production stations) changes.

Finally, both use cases mentioned above were combined into a third one to ensure a safe environment for workers and hardware equipment. The newly received coordinates of the stations were used to set the robot's destinations. The speed of the robot and the objects in the layout were also considered to create a collision-free navigation path for the robot. Moreover, predicting human behaviour is essential to configure the mobile robot, avoid possible collisions, and create safety zones. Human movement intention forecasting was used to plan an optimal robot route without collision risks. Such a route relies on the defined workflow progress done by workers to respond to production requirements. The use case's goal is to keep the production level high and maintain human safety during robot activity.

5. Experiments and results

5.1. Demand forecasting

We addressed the demand forecasting use case in four parts: (i) development of forecasting models, (ii) models' explainability, (iii) decision-making options recommendation, and (iv) the development of a voice interface.

To enable demand forecasts, we developed multiple statistical and machine learning models for products with smooth and erratic demands (Rožanec et al.,"Automotive OEM Demand Forecasting: A Comparative Study of Forecasting Algorithms and Strategies," 2021) and products with lumpy and intermittent demand (Rožanec and Mladenić 2021) (*Forecasting Module*). The models were developed based on real-world data from a European original equipment manufacturer targeting the global automotive industry market. For products with smooth and erratic demand, we found that the best results were obtained with global models trained across multiple time series, assuming that there is enough similarity between them to enhance learning. Furthermore, our research shows that the forecast errors of such models can be constrained by pooling product demand time series based on the past demand magnitude. On the other hand, for products with lumpy and intermittent demand we found best results were obtained applying a two-fold approach, which was more than 30% more precise than best existing approaches when predicting demand occurrence, resulting in important gains when considering Stock-keeping-oriented Prediction Error Costs (Martin, Spitzer, and Kühl 2020).

Demand forecasts influence the supply chain managers' decision-making process, and therefore additional insights, obtained through Explainable Artificial Intelligence (*XAI Module*), must be provided to understand the model's rationale behind a forecast. To that end, we explored the use of surrogate models to understand which features were most relevant to a particular forecast and used a custom ontology model to map relevant concepts to the aforementioned features (Rozanec 2021; Rožanec, Fortuna, and Mladenić 2022; Rožanec et al.,"Enriching Artificial Intelligence Explanations With Knowledge Fragments," 2022). Such mapping hides sensitive details regarding the underlying model from the end-user. It ensures meaning is conveyed with high-level concepts intelligible to the users while remaining faithful to the ranking of the features. Furthermore, we enriched the explanations by providing media news information regarding events that could have influenced the demand in the past and searched for open datasets that could be used to enrich the models' data to lead to better results in the future. Our demand forecasting models achieved state-of-the-art performance, while the enriched explanations displayed a high-degree of precision: for the worst cases, we achieved a precision of 0,95 for the media events displayed, a precision of 0,71 for the media keywords, and a precision of 0,56 for datasets displayed to the users.

Finally, we developed a heuristic recommender system to advise logisticians on decision-making options based on the demand forecasting outcomes (Rožanec et al.,"XAI-KG: Knowledge Graph to Support XAI and Decision-Making in Manufacturing," 2021) (*Decision-making Module*). The prototype application supported gathering (i) feedback regarding existing decision-making options (*Feedback Module*) and (ii) new knowledge to mitigate scenarios where the provided decision-making options did not satisfy the user. Feedback and new knowledge were persisted into a knowledge graph modelled after an ontology developed for this purpose. Furthermore, the user interface was developed to support interactions either through a graphical user interface or voice commands[7]. Future work will develop voice interfaces that are robust to noisy industrial environments.

Among the main challenges faced to create the demand forecasting models and provide models' explainability were the data acquisition and ensuring data quality. Data acquisition required working with different environments, application programming interfaces (e.g. to retrieve media news, information regarding demand, or other complementary information), and query constraints related to those environments and interfaces. Multiple iterations were performed to validate successive model versions and information displayed to the experts. Furthermore, through the iterations we worked on enhancing the models' performance, and expand the models' scope towards a greater number of products.

5.2. Quality inspection

Our work regarding quality inspection was addressed in four parts: (i) development of machine learning models for automated visual inspection of manufactured products (*Forecasting Module*), (ii) use of active learning to reduce the manual labelling efforts (*Active Learning Module*), (iii) use of simulated reality to generate synthetic images (*Simulated Reality Module*), and (iv) explore techniques to hint the user where defect could be (some hints were retrieved from the *XAI Module*).

To automate visual inspection, we explored batch and streaming models (Trajkova et al. 2021; Rožanec et al.,"Streaming Machine Learning and Online Active Learning for Automated Visual Inspection," 2021). While the batch models usually achieve better performance, they cannot leverage new data as it becomes available, but rather a new retrained model has to be deployed. Furthermore, while using all available data can help the model achieve better discriminative power, it is desirable to minimise the labelling and manual revision efforts, which can be achieved through active learning. We found that while models trained through active learning had a slightly inferior discriminative power, their performance consistently improved over time. In an active learning setting, we found that the best batch model (multilayer perceptron) achieved an average performance of 0,9792 AUC ROC, while the best streaming model (streaming kNN) lagged by at least 0,16 points. Both models were built using a ResNet-18 model (He et al. 2016) to extract embeddings from the Average Pooling layer. We selected a subset of features based on their mutual information ranking and evaluated the models with a stratified 10-fold cross-validation (Zeng and Martinez. 2000). Given the performance gap between both types of models and the high cost of miss-classification, batch models were considered the best choice in this use case. To decouple specific model implementations and their predictions from any service using those predictions, machine learning models can be calibrated to produce calibrated probabilities. We noted that in some cases, such model calibration further enhanced the models' discriminative power (Rožanec et al.,"Active Learning and Approximate Model

Calibration for Automated Visual Inspection in Manufacturing," 2022).

Given that defective parts always concern a small proportion of the overall production, it would be natural that the datasets are skewed, having a strong class imbalance. Furthermore, such imbalance is expected to increase over time as the manufacturing quality improves. Therefore, the *Simulated Reality Module* was used to generate synthetic images with two purposes. First, they were used to achieve greater class balance, leading to nearly perfect classification results (Rožanec et al.,"Synthetic Data Augmentation Using GAN For Improved Automated Visual Inspection," 2022). Second, synthetic images were used to balance data streams in manual revision to ensure attention is maximised and that defective pieces are not dismissed as good ones due to inertia (Rožanec et al.,"Towards a Comprehensive Visual Quality Inspection for Industry 4.0," 2022; Rožanec et al.,"Enhancing Manual Revision in Manufacturing With AI-Based Defect Hints," 2022). To that end, we developed a prototype application, that simulated a manual revision process and collected users' feedback (see Figure 5). Furthermore, cues were be provided to the users, to help them identify possible defects. To that end we explored three techniques: (a) Grad-CAM (Selvaraju et al. 2017), (b) DRAEM, and (c) the most similar labelled image. GradCAM is an Explainable Artificial Intelligence method suitable for deep learning models. It uses the gradient information to understand how strongly does each neuron activate in the last convolutional layer of the neural network. They are then combined with existing high-resolution visualisations to obtain class-discriminative guided visualisations as saliency masks. DRAEM (Zavrtanik, Kristan, and Skočaj 2021) is a state-of-the-art method for unsupervised anomaly detection. It works by training an autoencoder on anomaly-free images and using it to threshold the difference between the input images and the autoencoder reconstruction. Finally, the most similar labelled images were retrieved considering the structural similarity index measure (Wang et al. 2004). From the experiments performed (Rožanec et al.,"Towards a Comprehensive Visual Quality Inspection for Industry 4.0," 2022; Rožanec et al.,"Enhancing Manual Revision in Manufacturing With AI-Based Defect Hints," 2022), we found that the best results were obtained when hinting the users with the images and labels with the closest structural similarity index. This resulted in an increased mean labelling time (by 30%), but a higher quality of labelling (three times the original labelling precision, and two times the original F1 score). In addition, the number of unidentified defects was reduced by more than 80%. Future work will explore how users' feedback can lead to discovering new defects and whether users' fatigue can be detected to alternate types of work or suggest breaks to the operators, enhancing their work experience and the quality of the outcomes.

While for this particular use case we succeeded on gathering a good quality dataset of labelled images, the process has not been straightforward on similar use cases, where multiple iterations were required, to ensure enough and high-quality data was provided. Furthermore, much work was invested towards getting few people to perform a set of lengthy experiments that provided new insights on which cues helped best towards enhancing the quality of labelling. Nevertheless, their data provided valuable insights and findings, and ground towards new directions of research.

5.3. Human behaviour prediction and safe zone detection

We have designed a sensing prototype and simple neural networks to evaluate our human activity recognition module to predict the user's movement trajectory based on artificial intelligence models. The sensing prototype (see Figure 6) combines three boards: an nRF52840 backend motherboard from Adafruit Feather, a customised human body capacitance sensing board, and a data logger board. The nRF52840 board supplies three axes of Inertial Measurement Unit (IMU) data, including acceleration, gyroscope, and magnet. The customised body capacitance board is verified efficiently to sense both the body movement (Bian et al., "Passive Capacitive Based Approach for Full Body Gym Workout Recognition and Counting," 2019) and the environmental context (Bian et al., "Wrist-Worn Capacitive Sensor for Activity and Physical Collaboration Recognition," 2019) by measuring the skin potential signal. The sensor data is stored in a Secure Digital card on the data logger board at a rate of 20 Hz. Since the environment in the factory is full of 2.4 G Hz wireless signals, such as WiFi and Bluetooth, to avoid data package loss, an SD card has been used to record the data locally and finally synchronised by checking some predefined actions. The sensing component, the IMU and body capacitance sensor, consumes the power with a level of sub-mW. A 3.7 V chargeable lithium battery is used for the power supply. The feasibility of developing machine learning models for human activity recognition was validated through the development of an adversarial encoder-decoder structure with maximum mean discrepancy to realign the data distribution over multiple subjects, and tested on four open datasets. Suh, Rey, and Lukowicz (2022a) report that the results obtained outperformed state-of-the-art methods and improved generalisation capabilities. The

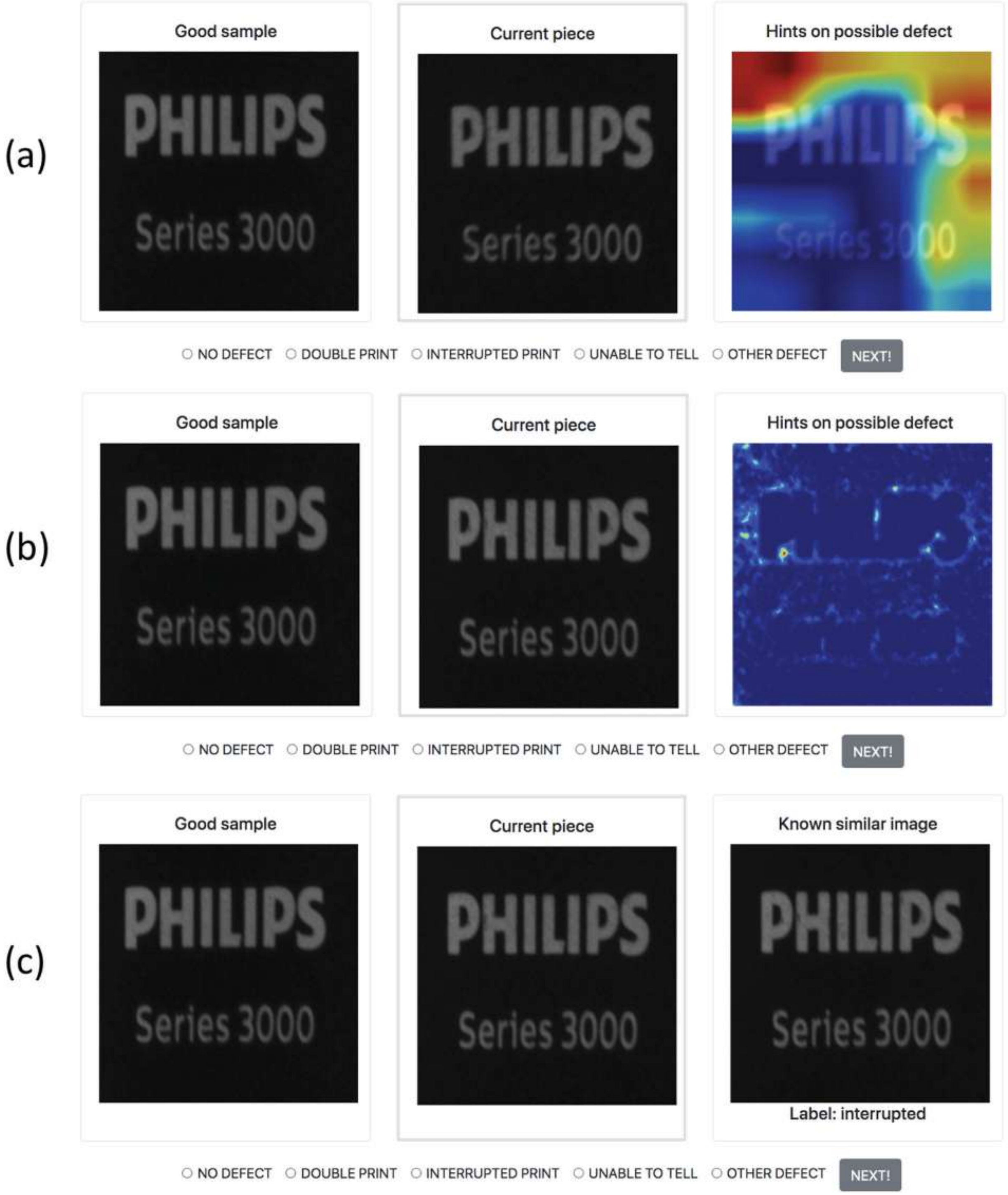

Figure 5. Sample screen for the manual revision process. We provide the operator an image of a non-defective part, the image of the component being inspected, and the hints regarding where we do expect the error can be. The images correspond to cases where the hints were created with (a) GradCAM, (b) DRAEM, and (c) the most similar labelled image.

same authors also proposed TASKED (Transformer-based Adversarial learning framework for human activity recognition using wearable sensors via Self-KnowledgE Distillation), a deep learing architecture capable of learning cross-domain feature representations using adversarial learning and maximum mean discrepancy to align data distributions from multiple domains (Suh, Rey, and Lukowicz 2022b). Future work will address the development of models for human intention recognition and how such predictions can be leveraged for safe zone detection when routeing autonomous mobile robots in the manufacturing context.

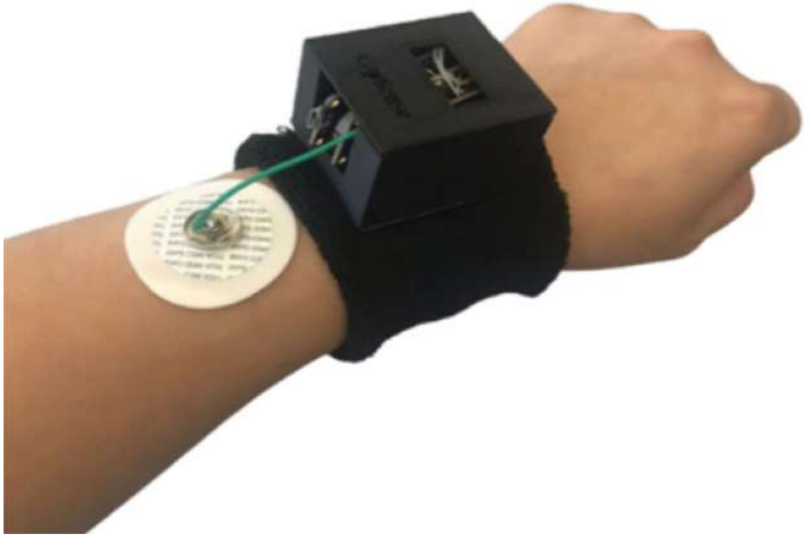

Figure 6. The sensing prototype for worker's activity recognition worn on the wrist.

6. Conclusions

The increasing digitalisation of the manufacturing processes has democratised the use of artificial intelligence across manufacturing. Consequently, jobs are being reshaped, fostering the development of human-machine collaboration approaches. Humans and machines have unique capabilities, which can be potentiated through a synergistic relationship. A systemic approach is required to realise such a collaboration at its fullest. Furthermore, an architecture must be devised to support it. In particular, such an architecture must consider modules related to forecasting, explainable artificial intelligence, active learning, simulated reality, decision-making, and human feedback.

We validated the feasibility of the proposed architecture through three real-world use cases (demand forecasting, quality inspection, human behaviour prediction, and safe zone detection). The experiments and results obtained in each case show how artificial intelligence can be used to achieve particular goals in manufacturing. Furthermore, it confirms the interplay between the architecture modules to deliver a human-centric experience aligned with the Industry 5.0 paradigm. Nevertheless, further research work is required to hone and highlight aspects related to workers' safety and how the current architecture supports this.

Ongoing and future work is and will be focussed on three research directions. First, we are researching human intention recognition to enhance workers' safety in industrial settings using wearable technologies. Second, we will explore active learning approaches for cybersecurity to enable real-time assessment of cyberattacks and proactively act on them. Finally, we will develop machine learning and active learning approaches for human fatigue monitoring to enhance workers' well-being in manufacturing settings.

Notes

1. https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API
2. https://www.w3.org/TR/jsgf/
3. https://assistant.google.com/
4. https://www.apple.com/siri/
5. https://www.microsoft.com/en-in/windows/cortana
6. https://developer.amazon.com/alexa
7. A video of the application was published in https://www.youtube.com/watch?v = EpFBNwz6Klk

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 program projects FACTLOG and STAR under grant agreements numbers H2020-869951 and H2020-956573.

Notes on contributors

Jože Martin Rožanec is a PhD candidate in Information and Communication Technologies at Jožef Stefan International Postgraduate School. He is a Researcher at the Artificial Intelligence Laboratory (Jožef Stefan Institute), a machine learning engineer at Qlector d.o.o. (developing intelligent solutions for smart factories). He collaborates with the American Slovenian Education Foundation, where he leads multiple activities for Fellows and Alumni. Over more than ten years he worked for several companies (e.g.

Mercado Libre, Navent, Globant) in software engineering and machine learning-related roles. His research interests include machine learning methods for recommendations, fraud detection, demand forecasting, active learning, and explainable artificial intelligence (XAI).

Inna Novalija Dr. works as a researcher and advisor at the Artificial Intelligence Laboratory, Jožef Stefan Institute. She is an expert in research and development in the area of Artificial Intelligence and Data Science. In particular, she is working on utilising text analysis techniques for modelling and forecasting and data mining techniques for developing applicable tools and technologies. She participated as a team leader and team member in EU projects, national and international projects. In 2012–2013 she was a visiting researcher in Knowledge Media Institute (the Open University, UK), working in the area of Semantic Technologies and Semantic Web with Prof. John Domingue et al. She is a program committee member for international conferences and reviewer for several international journals.

Patrik Zajec is a master's student at the Jožef Stefan International Postgraduate School. His research is mainly in the field of natural language processing, more specifically he develops methods for unsupervised and semi-supervised information extraction and retrieval. He is also involved in the European Union Horizon 2020 program project STAR, where he is working on synthetic data augmentation techniques and a human-machine collaboration system using active learning, XAI, and semantic technologies.

Klemen Kenda (M) is a researcher at Jozef Stefan Institute. He obtained his diploma in physics at University of Ljubljana and is pursuing his PhD in Information and Communication Technologies at Jožef Stefan International Postgraduate School. He has been involved with machine learning and stream mining of heterogeneous data sources. Applications of his work have been made in the fields of environmental intelligence and energy management. He has contributed to several EU FP7 and H2020 projects (Planetdata, Envision, NRG4CAST, Sunseed, PerceptiveSentinel, EnviroLENS, Factlog) from 2011 and acted as a leader of several technical work packages. He is the author of a chapter on mashups for environmental intelligence in the book 'Semantic Mashups' (Springer, 2013). He is one of the contributing authors of QMiner, open-source data analytics platform for processing large-scale real-time streams.

Hooman Tavakoli Ghinani is a PhD student and project manager in the Smartfactory-DFKI in Kaiserslautern, Germany. He finished his bachelor's at the computer science department of the Azad University of Najafabad, Iran. He continued his study toward the master's degree in the field of intelligent systems at the department of computer science and completed that in 2020 at the technical university of Kaiserslautern, Germany. His master thesis's topic was 'Extensive study of probabilistic regression using GANs '. His fields of research are machine learning approaches, very deep learning models, GANs, augmented reality, synthetic data generations, and computer vision approaches.

Sungho Suh received the PhD degree in Computer Science from the University of Kaiserslautern, Germany in 2021, and the B.S. and M.S. degrees in Electrical Engineering and Computer Science from Seoul National University, Seoul, Republic of Korea, in 2009 and 2011, respectively. From 2018 to 2021, he was with KIST Europe, Saarbrücken, Germany. From 2011 to 2018, he was a research engineer with Samsung Electro-Mechanics, Suwon, Republic of Korea. He is currently a Senior Researcher with German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His research interests are machine learning algorithms, such as domain adaptation, GAN, and object segmentation, with a focus on industrial applications.

Entso Veliou From 2010 to 2014, he attended the University of Piraeus and earned a bachelor's degree in computer science as well as a master's degree in information security from 2014 to 2016. In 2014, he won an entrance into HUAWEI's seeds for the future program in China, which he finished with honours. He is now pursuing a PhD in machine learning and cybersecurity at UNIWA. His primary area of expertise is cybersecurity blue and red teaming, and he is OSCP, OSCE, and ISO27001 certified. Furthermore, his primary responsibilities at Ubitech include security integration and implementation procedures that are cutting-edge.

Dimitrios Papamartzivanos Dr. holds a PhD on advanced AI methods for network intrusion detection systems (IDS). His research deals both with misuse and anomaly detection systems, AI-enabled adaptive IDS, while his focus now is on the development of adversarial and defence strategies in AI. His research interests revolve around S&P for IoT, AI and Bio-inspired Algorithms, and Trusted Computing. Dimitris has worked as an intern researcher in the Security R&D group of NEC Europe, as an exchanged researcher in the University of Murcia and, currently, as Cybersecurity Research Associate in the Digital Security & Trusted Computing Group at UBITECH Ltd.

Thanassis Giannetsos received the PhD degree from the University of Aalborg, Aalborg, Denmark, in 2012. He was an Assistant Professor with the Department of Computer Science, University of Surrey, Guildford, U.K. He is currently the Head of Digital Security and Trusted Computing Group at Ubitech Ltd. and an Adjunct Associate Professor within the Cyber Security Section, Technical University of Denmark. He has expertise in the design and implementation of secure and privacy-preserving protocols and risk management. His current research interests include applied cryptography to security and privacy in information technology.

Sofia-Anna Menesidou Dr. holds a PhD in the area of Cryptographic Key Management in Delay Tolerant Networks from the department of Electrical and Computer Engineering of Democritus University of Thrace and an MSc degree in Information and Communication Systems Security of the University of the Aegean. She has published several papers in international journals and conferences, while she has participated in several research projects including SecureIoT, FutureTPM, SECONDO, InteropEHRate, CUREX, PUZZLE etc. Dr Menesidou is currently research associate of the Digital Security & Trusted Computing Research Group at UBITECH and her research interests are in the field of cyber security.

Rubén Alonso holds a Diploma of Advanced Studies on Computing Systems from the University of the Basque Country and a Computer Science and Engineering degree from the University of Deusto. He has been involved in industrial and research projects since 2004, mainly related to computational trust, security and embedded systems. He was part of the Trust and Security Unit of the European Software Institute and the innovation group of Visual Tools. He was also a visiting researcher in VTT Technical Research Centre of Finland and the University of Trento. Currently, he coordinates the ICT&Robotics Team of R2M Solution.

Nino Cauli is researcher at the University of Cagliari. He recently won a Marie Curie Fellowship for the project 'Dr. VCoach'. His research focuses on the study of Computer Vision deep models for Human Robot Interaction. He received his PhD in Biorobotics from the BioRobotics Institute of Scuola Superiore Sant'Anna, Pisa, in 2014. He was Postdoctoral Fellow at the BioRobotics Institute in Pisa (2014/2015), at the ISR in Lisbon (2016/2018), at the Ryerson University in Toronto (2019), and at the University of Catania (2019/2021). He collaborated in a number of international projects (RoboSom, Human Brain Project, Augmented Human Assistance (AHA)).

Antonello Meloni holds a Bachelor's in Computer Sciences and Technologies and a Master's Degree in Computer Science at the University of Cagliari. He is currently a PhD student in Mathematics and Computer Science at the University of Cagliari. He is involved in research projects, mainly related to Computer Science, Natural Language Processing, Human Robot Interaction and Semantic Web. Since 1998 he has been working for local public administrations as system administrator, expert in Geographic Information Systems and CAD, software developer and IT infrastructure designer.

Diego Reforgiato Recupero is a Full Professor at the Department of Mathematics and Computer Science of the University of Cagliari, Italy. He is director and creator of the Human-Robot-Interaction Laboratory (http://hri.unica.it), co-director and creator of the Artificial Intelligence and Big Data Laboratory (http://aibd.unica.it), co-director of the Semantic Web laboratory (http://swlab.unica.it/), and member of the Commission for start-up and spin-off. He is the coordinator and co-founder of the new bachelor's degree 'Applied Computer Science and Data Analytics' of the University of Cagliari. He teaches 'Computers Architecture', 'Big Data' and 'Deep Learning and Applications' for the master degree in Computer Science.

Dimosthenis Kyriazis is an Associate Professor at the Department of Digital Systems, Director of the Network-Based Systems and Services Laboratory, and Director of the Postgraduate Program on Information Systems and Services. He graduated from the School of Electrical and Computer Engineering of the National Technical University of Athens / NTUA. Since 2007 he holds a PhD from the Electrical and Computer Engineering School of NTUA in the area of service-oriented architectures with an emphasis on quality of service and workflow management. His research interests focus on service-oriented, distributed, and heterogeneous systems, as well as software and data analysis technologies. He has participated in a number of European and National projects (e.g. BigDataStack, INFINITECH, CrowdHEALTH, iHELP, 5GTANGO, CYBELE) with emphasis on research topics related to the provision of quality of service guarantees, fault tolerance, resources management, performance modelling in service-oriented environments and application domains such as multimedia, post-production, virtual reality, finance, e-health, etc. He is currently focussing on virtualisation technologies for high availability of cloud computing infrastructures, resource management techniques in cloud computing and edge computing environments, and data management and analytics in the aforementioned environments. The outcomes of his research are reflected in more than 150 publications in international scientific journals and conferences.

Georgios Sofianidis is an electrical and computer engineer who graduated from the National Technical University of Athens (N.T.U.A.) in 2019. He participated in European projects, building machine learning applications while working as a researcher for the University of Piraeus Research Center (U.P.R.C.).

Spyros Theodoropoulos is a PhD Candidate at the School of Electrical and Computer Engineering of the National Technical University of Athens. He is a graduate of the same school and also holds a MSc in Machine Learning from Imperial College London. He has worked in industry as a Software and Big Data Engineer and is now a member of the Data & Cloud Research Group at the University of Piraeus. His research is focussed on deep learning, reinforcement learning and the use of simulation for their deployment in real-life dynamic environments.

Blaž Fortuna is the managing director at Qlector, a startup developing artificial intelligence-based solutions for the manufacturing, logistics, finance, and senior researcher at Jozef Stefan Institute. He is the initiator and primary contributor to QMiner, the open-source data analytics platform for processing large-scale real-time streams containing structured and unstructured data and co-contributor to Event Registry. He did his Ph.D. at Jožef Stefan Institute. He was a research consultant for Bloomberg L.P., a Marie Curie Fellow at Stanford University, a postdoc at IBCN (Ghent University, Belgium), and the project manager for the XLike project.

Dunja Mladenić Prof. Dr. http://ailab.ijs.si/dunja_mladenic/ works as a researcher and a project leader at Jožef Stefan Institute, Slovenia, leading Artificial Intelligence Department and teaching at Jožef Stefan International Postgraduate School, University of Ljubljana and the University of Zagreb. She has extensive research experience in the study and development of Machine Learning, Big Data/Text Mining, the Internet of Things, Data Science, Semantic Technology techniques, and their application to real-world problems. She has published papers in refereed journals and conferences, co-edited several books, served on program committees of international conferences, and organised international events. She serves as a project evaluator of project proposals for the European Commission and USA National Science Foundation. From 2013–2017 she served on the Institute's Scientific Council as a vice president (2015–2017). She serves on the Executive Board of Slovenian Artificial Intelligence Society SLAIS (as a president of SLAIS (2010–2014)) and on the Advisory board of ACM Slovenija.

John Soldatos (http://gr.linkedin.com/in/johnsoldatos) holds a PhD in Electrical & Computer Engineering from the National Technical University of Athens (Greece) and is currently Senior R&D Consultant Innovation Delivery Specialist with Netcompany-Intrasoft (Luxembourg) and Honorary Research Fellow at the University of Glasgow (UK). He has played a leading role in the successful delivery of over sixty (commercial-industrial, research, consulting) projects. He is co-founder of the open source OpenIoT platform. His current research interests are in Internet of Things (IoT) and Artificial Intelligence (AI). Dr. Soldatos has published more than 200 articles in international journals, books, and conference proceedings, while he has coedited and co-authored seven books in IoT, Industry 4.0 and AI.

Data availability statement

Data not available due to restrictions.

ORCID

Jože M. Rožanec http://orcid.org/0000-0002-3665-639X
Inna Novalija http://orcid.org/0000-0003-2598-0116
Patrik Zajec http://orcid.org/0000-0002-6630-3106
Klemen Kenda http://orcid.org/0000-0002-4918-0650
Sungho Suh http://orcid.org/0000-0003-3723-1980
Entso Veliou http://orcid.org/0000-0001-9730-1720
Dimitrios Papamartzivanos http://orcid.org/0000-0002-9471-5415
Thanassis Giannetsos http://orcid.org/0000-0003-0663-2263
Sofia Anna Menesidou http://orcid.org/0000-0003-2446-5470
Nino Cauli http://orcid.org/0000-0002-9611-0655
Antonello Meloni http://orcid.org/0000-0001-6768-4599
Diego Reforgiato Recupero http://orcid.org/0000-0001-8646-6183
Dimosthenis Kyriazis http://orcid.org/0000-0001-7019-7214
Georgios Sofianidis http://orcid.org/0000-0002-9640-6317
Blaž Fortuna http://orcid.org/0000-0002-8585-9388
Dunja Mladenić http://orcid.org/0000-0002-0360-6505
John Soldatos http://orcid.org/0000-0002-6668-3911

References

Afanasev, Maxim Ya, Yuri V. Fedosov, Yuri S. Andreev, Anastasiya A. Krylova, Sergey A. Shorokhov, Kseniia V. Zimenko, and Mikhail V. Kolesnikov. 2019. "A Concept for Integration of Voice Assistant and Modular Cyber-Physical Production System." In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, Vol. 1, 27–32. IEEE.

Ahmed, Imran, Gwanggil Jeon, and Francesco Piccialli. 2022. "From Artificial Intelligence to EXplainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where." *IEEE Transactions on Industrial Informatics* 18 (8): 5031–5042. doi:10.1109/TII.2022.3146552.

AIA. n.d.a. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." Accessed 15 March 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri = CELEX:52021PC0206.

AIA. n.d.b. "Projeto de Lei N 21 DE 2020." Accessed 15 March 2022. https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor = 1853928.

Almada, Marco. 2019. "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems." In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2–11.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." arXiv preprint arXiv:1606.06565.

Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. 2021. "Explainable Artificial Intelligence: An Analytical Review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11 (5): Article ID e1424.

Ani, Uchenna P. Daniel, Hongmei He, and Ashutosh Tiwari. 2017. "Review of Cybersecurity Issues in Industrial Critical Infrastructure: Manufacturing in Perspective." *Journal of Cyber Security Technology* 1 (1): 32–74.

Ansari, Fazel, Selim Erol, and Wilfried Sihn. 2018. "Rethinking Human-Machine Learning in Industry 4.0: How Does the Paradigm Shift Treat the Role of Human Learning?" *Procedia Manufacturing* 23: 117–122.

Ansari, Fazel, Selim Erol, and Wilfried Sihn. 2018. "Rethinking Human-Machine Learning in Industry 4.0: How Does the Paradigm Shift Treat the Role of Human Learning?" *Procedia Manufacturing* 23: 117–122. doi:10.1016/j.promfg.2018.04.003.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, and Salvador García, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Bayanifar, Hessamedin, and Hermann Kühnle. 2017. "Enhancing Dependability and Security of Cyber-Physical Production Systems." In *Doctoral Conference on Computing, Electrical and Industrial Systems*, 135–143. Springer.

Bi, Zhu Ming, Chaomin Luo, Zhonghua Miao, Bing Zhang, W. J. Zhang, and Lihui Wang. 2021. "Safety Assurance Mechanisms of Collaborative Robotic Systems in Manufacturing." *Robotics and Computer-Integrated Manufacturing* 67: Article ID 102022.

Bian, Sizhen, Vitor F. Rey, Peter Hevesi, and Paul Lukowicz. 2019. "Passive Capacitive Based Approach for Full Body Gym Workout Recognition and Counting." In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 1–10. IEEE.

Bian, Sizhen, Vitor F. Rey, Junaid Younas, and Paul Lukowicz. 2019. "Wrist-Worn Capacitive Sensor for Activity and Physical Collaboration Recognition." In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 261–266. IEEE.

Brühl, Bernhard, Marco Hülsmann, Detlef Borscheid, Christoph M Friedrich, and Dirk Reith. 2009. "A Sales Forecast Model for the German Automobile Market Based on Time Series Analysis and Data Mining Methods." In *Industrial Conference on Data Mining*, 146–160. Springer.

Budd, Samuel, Emma C Robinson, and Bernhard Kainz. 2021. "A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis." *Medical Image Analysis* 71: Article ID 102062.

Buerkle, Achim, William Eaton, Niels Lohse, Thomas Bamber, and Pedro Ferreira. 2021. "EEG Based Arm Movement Intention Recognition Towards Enhanced Safety in Symbiotic Human-Robot Collaboration." *Robotics and Computer-Integrated Manufacturing* 70: 102137.

Bullers, William I., Shimon Y. Nof, and Andrew B. Whinston. 1980. "Artificial Intelligence in Manufacturing Planning and Control." *AIIE Transactions* 12 (4): 351–363.

Chhetri, Sujit Rokka, Sina Faezi, Nafiul Rashid, and Mohammad Abdullah Al Faruque. 2018. "Manufacturing Supply Chain and Product Lifecycle Security in the Era of Industry 4.0." *Journal of Hardware and Systems Security* 2 (1): 51–68.

Chhetri, Sujit Rokka, Nafiul Rashid, Sina Faezi, and Mohammad Abdullah Al Faruque. 2017. "Security Trends and Advances in Manufacturing Systems in the Era of Industry 4.0." In *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1039–1046. IEEE.

Chien, Chen-Fu, Stéphane Dauzère-Pérès, Woonghee Tim Huh, Young Jae Jang, and James R. Morrison. 2020. "Artificial Intelligence in Manufacturing and Logistics Systems: Algorithms, Applications, and Case Studies."

CIS. n.d. "Cybersecurity Information Sharing Act of 2015." Accessed 15 March 2022. https://www.cisa.gov/sites/default/files/publications/Cybersecurity%20Information%20Sharing%20Act%20of%202015.pdf.

Conti, Mauro, Tooska Dargahi, and Ali Dehghantanha. 2018. "Cyber Threat Intelligence: Challenges and Opportunities." In *Cyber Threat Intelligence*Advances in Information Security, Vol. 70, edited by A. Dehghantanha, M. Conti, and T. Dargahi, 1–6. Cham: Springer. doi:10.1007/978-3-319-73951-9_1.

Corredor, Pilar, Elena Ferrer, and Rafael Santamaria. 2014. "Is Cognitive Bias Really Present in Analyst Forecasts? The Role of Investor Sentiment." *International Business Review* 23 (4): 824–837.

Creswell, Antonia, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. 2018. "Generative Adversarial Networks: An Overview." *IEEE Signal Processing Magazine* 35 (1): 53–65.

cyb. n.d. "Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on Information and Communications Technology Cybersecurity Certification and Repealing Regulation (EU) No 526/2013 (Cybersecurity Act)." Accessed 15 March 2022. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32019R0881.

Dai, Wenting, Abdul Mujeeb, Marius Erdt, and Alexei Sourin. 2018. "Towards Automatic Optical Inspection of Soldering Defects." In *2018 International Conference on Cyberworlds (CW)*, 375–382. IEEE.

Das, Arun, and Paul Rad. 2020. "Opportunities and Challenges in Explainable Artificial Intelligence (Xai): A Survey." arXiv preprint arXiv:2006.11371.

dat. n.d. "Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act)." Accessed 15 March 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767.

de Melo, Celso M, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. 2022. "Next-generation Deep Learning Based on Simulators and Synthetic Data." *Trends in Cognitive Sciences* 26 (2): 174–187. doi:10.1016/j.tics.2021.11.008.

Demir, Kadir Alpaslan, Gözde Döven, and Bülent Sezen. 2019. "Industry 5.0 and Human-Robot Co-Working." *Procedia Computer Science* 158: 688–695.

Doran, Derek, Sarah Schulz, and Tarek R Besold. 2017. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives." arXiv preprint arXiv:1710.00794.

EC2. 2020, September. "European Commission, Enabling Technologies for Industry 5.0, Results of a Workshop with Europe's Technology Leaders." https://op.europa.eu/en/publication-detail/-/publication/8e5de100-2a1c-11eb-9d7e-01aa75ed71a1/language-en.

Elahi, Mehdi, Francesco Ricci, and Neil Rubens. 2016. "A Survey of Active Learning in Collaborative Filtering Recommender Systems." *Computer Science Review* 20: 29–50.

Elhabashy, Ahmad E., Lee J. Wells, and Jaime A. Camelio. 2019. "Cyber-Physical Security Research Efforts in Manufacturing–a Literature Review." *Procedia Manufacturing* 34: 921–931.

El Zaatari, Shirine, Mohamed Marei, Weidong Li, and Zahid Usman. 2019. "Cobot Programming for Collaborative

Industrial Tasks: An Overview." *Robotics and Autonomous Systems* 116: 162–180.

ePr. n.d. "Consolidated Text: Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (Directive on Privacy and Electronic Communications)." Accessed 15 March 2022. https://eur-lex.europa.eu/eli/dir/2002/58/2009-12-19.

Erol, Berat A., Conor Wallace, Patrick Benavidez, and Mo Jamshidi. 2018. "Voice Activation and Control to Improve Human Robot Interactions with IoT Perspectives." In *2018 World Automation Congress (WAC)*, 1–5.

Fildes, Robert, and Paul Goodwin. 2021. "Stability in the Inefficient Use of Forecasting Systems: A Case Study in a Supply Chain Company." *International Journal of Forecasting* 37 (2): 1031–1046.

Forkan, Abdur Rahim Mohammad, Federico Montori, Dimitrios Georgakopoulos, Prem Prakash Jayaraman, Ali Yavari, and Ahsan Morshed. 2019. "An Industrial IoT Solution for Evaluating Workers' Performance via Activity Recognition." In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 1393–1403. IEEE.

GDP. n.d. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)." Accessed 15 March 2022. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. Vol. 27, 2672–2680. Cambridge, MA, USA: MIT Press.

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'." *AI Magazine* 38 (3): 50–57.

Grieves, Michael. 2015. "Digital Twin: Manufacturing Excellence Through Virtual Factory Replication".

Grieves, Michael, and John Vickers. 2017. "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems." In *Transdisciplinary Perspectives on Complex Systems*, 85–113. Springer.

Günther, Lisa C., Susann Kärcher, and Thomas Bauernhansl. 2019. "Activity Recognition in Manual Manufacturing: Detecting Screwing Processes From Sensor Data." *Procedia CIRP* 81: 1177–1182.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Henin, Clément, and Daniel Le Métayer. 2021. "A Multi-Layered Approach for Tailored Black-Box Explanations."

Hentout, Abdelfetah, Mustapha Aouache, Abderraouf Maoudj, and Isma Akli. 2019. "Human–Robot Interaction in Industrial Collaborative Robotics: A Literature Review of the Decade 2008–2017." *Advanced Robotics* 33 (15–16): 764–799.

Hogarth, Robin M., and Spyros Makridakis. 1981. "Forecasting and Planning: An Evaluation." *Management Science* 27 (2): 115–138.

IIR. n.d. "The Industrial Internet of Things Volume G1: Reference Architecture." Version 1.9 June 19, 2019. https://www.iiconsortium.org/pdf/IIRA-v1.9.pdf.

IIS. 2016, September. "Industrial Internet of Things, Volume G4: Security Framework, Industrial Internet Consortium." IIC:PUB:G4:V1.0:PB:20160926.

Ind. n.d. "Industry 5.0: Towards More Sustainable, Resilient and Human-Centric Industry." Accessed 15 March 2022. https://op.europa.eu/en/publication-detail/-/publication/468a892a-5097-11eb-b59f-01aa75ed71a1/.

ISO. n.d. "ISO/IEC 27000:2018 Information Technology – Security Techniques – Information Security Management Systems – Overview and Vocabulary." Accessed 15 March 2022. https://www.iso.org/standard/73906.html.

Jentzsch, Sophie F., Sviatlana Höhn, and Nico Hochgeschwender. 2019. "Conversational Interfaces for Explainable AI: A Human-Centred Approach." In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 77–92. Springer.

Jeschke, Sabina, Christian Brecher, Tobias Meisen, Denis Özdemir, and Tim Eschert. 2017. "Industrial Internet of Things and Cyber Manufacturing Systems." In *Industrial Internet of Things*, 3–19. Springer.

Johannink, Tobias, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. 2019. "Residual Reinforcement Learning for Robot Control." In *2019 International Conference on Robotics and Automation (ICRA)*, 6023–6029. IEEE.

Kaasinen, Eija, Franziska Schmalfuß, Cemalettin Özturk, Susanna Aromaa, Menouer Boubekeur, Juhani Heilala, and Päivi Heikkilä, et al. 2020. "Empowering and Engaging Industrial Workers with Operator 4.0 Solutions." *Computers & Industrial Engineering* 139: Article ID 105678.

Kang, Sangseung, and Kyekyung Kim. 2018. "Motion Recognition System for Worker Safety in Manufacturing Work Cell." In *2018 18th International Conference on Control, Automation and Systems (ICCAS)*, 1774–1776. IEEE.

Keane, Mark T., Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. "If Only We had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques." arXiv preprint arXiv:2103.01035.

Klein, Jan, Sandjai Bhulai, Mark Hoogendoorn, and Rob van der Mei. 2022. "Jasmine: A New Active Learning Approach to Combat Cybercrime." *Machine Learning with Applications* 9: Article ID 100351. doi:10.1016/j.mlwa.2022.100351.

Klopfenstein, Lorenz Cuno, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. "The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms." In *Proceedings of the 2017 Conference on Designing Interactive Systems*, 555–565.

Kouroupetroglou, Christos, Dympna Casey, Massimiliano Raciti, Eva Barrett, Grazia D'Onofrio, Francesco Ricciardi, and Francesco Giuliani, et al. 2017. "Interacting With Dementia: The MARIO Approach." In *Harnessing the Power of Technology to Improve Lives, Proceedings of the 14th European Conference on the Advancement of Assistive Technology, AAATE Conf. 2017, Sheffield, UK, September 12–15, 2017*, edited by Peter Cudd and Luc P. de Witte, Vol. 242 of *Studies in Health Technology and Informatics*, 38–47. IOS Press. doi:10.3233/978-1-61499-798-6-38 .

Kumar, Punit, and Atul Gupta. 2020. “Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey.” *Journal of Computer Science and Technology* 35 (4): 913–945.

Kuo, Chu-Chi, Joseph Z. Shyu, and Kun Ding. 2019. “Industrial Revitalization Via Industry 4.0 – A Comparative Policy Analysis Among China, Germany and the USA.” *Global Transitions* 1: 3–14.

Lee, Edward A. 2006. “Cyber-Physical Systems-Are Computing Foundations Adequate.” In *Position Paper for NSF Workshop on Cyber-Physical Systems: Research Motivation, Techniques and Roadmap*, Vol. 2, 1–9. Citeseer.

Leslie, David. 2019. “Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector.” SSRN 3403301.

Li, Jian-Hua. 2018. “Cyber Security Meets Artificial Intelligence: A Survey.” *Frontiers of Information Technology & Electronic Engineering* 19 (12): 1462–1474.

Li, Chun-Liang, Chun-Sung Ferng, and Hsuan-Tien Lin. 2015. “Active Learning Using Hint Information.” *Neural Computation* 27 (8): 1738–1765.

Li, Xu, Lin Hong, Jian-Chun Wang, and Xiang Liu. 2019. “Fatigue Driving Detection Model Based on Multi-Feature Fusion and Semi-Supervised Active Learning.” *IET Intelligent Transport Systems* 13 (9): 1401–1409.

Li, Chen, Andreas Kornmaaler Hansen, Dimitrios Chrysostomou, Simon Bøgh, and Ole Madsen. 2022. “Bringing a Natural Language-Enabled Virtual Assistant to Industrial Mobile Robots for Learning, Training and Assistance of Manufacturing Tasks.” In *2022 IEEE/SICE International Symposium on System Integration (SII)*, 238–243. IEEE.

Lim, Kendrik Yan Hong, Pai Zheng, and Chun-Hsien Chen. 2020. “A State-of-the-Art Survey of Digital Twin: Techniques, Engineering Product Lifecycle Management and Business Innovation Perspectives.” *Journal of Intelligent Manufacturing* 31 (6): 1313–1337.

Liu, Kang, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2020. “Poisoning the (Data) Well in ML-Based CAD: A Case Study of Hiding Lithographic Hotspots.” In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 306–309. IEEE.

Loyola-Gonzalez, Octavio. 2019. “Black-Box vs White-Box: Understanding Their Advantages and Weaknesses from a Practical Point of View.” *IEEE Access* 7: 154096–154113.

Lu, Yuqian, Hao Zheng, Saahil Chand, Wanqing Xia, Zengkun Liu, Xun Xu, Lihui Wang, Zhaojun Qin, and Jinsong Bao. 2022. “Outlook on Human-Centric Manufacturing Towards Industry 5.0.” *Journal of Manufacturing Systems* 62: 612–627.

Lundberg, Scott, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” arXiv preprint arXiv:1705.07874.

Maddikunta, Praveen Kumar Reddy, Quoc-Viet Pham, B Prabadevi, N Deepa, Kapal Dev, Thippa Reddy Gadekallu, Rukhsana Ruby, and Madhusanka Liyanage. 2022. “Industry 5.0: A Survey on Enabling Technologies and Potential Applications.” *Journal of Industrial Information Integration* 26:100257. doi:10.1016/j.jii.2021.100257.

Mahadevan, Sridhar, and Georgios Theocharous. 1998. “Optimizing Production Manufacturing Using Reinforcement Learning.” In *FLAIRS Conference*, Vol. 372, 377.

Mahapatra, Dwarikanath, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. 2018. “Efficient Active Learning for Image Classification and Segmentation Using a Sample Selection and Conditional Generative Adversarial Network.” In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 580–588. Springer.

Mahesh, Priyanka, Akash Tiwari, Chenglu Jin, Panganamala R. Kumar, A. L. Narasimha Reddy, Satish T. S. Bukkapatanam, Nikhil Gupta, and Ramesh Karri. 2020. “A Survey of Cybersecurity of Digital Manufacturing.” *Proceedings of the IEEE* 109 (4): 495–516.

Malaisé, Adrien, Pauline Maurice, Francis Colas, François Charpillet, and Serena Ivaldi. 2018. “Activity Recognition With Multiple Wearable Sensors for Industrial Applications.” In *ACHI 2018-Eleventh International Conference on Advances in Computer-Human Interactions*.

Maman, Zahra Sedighi, Ying-Ju Chen, Amir Baghdadi, Seamus Lombardo, Lora A. Cavuoto, and Fadel M. Megahed. 2020. “A Data Analytic Framework for Physical Fatigue Management Using Wearable Sensors.” *Expert Systems with Applications* 155: Article ID 113405.

Maman, Zahra Sedighi, Mohammad Ali Alamdar Yazdi, Lora A. Cavuoto, and Fadel M. Megahed. 2017. “A Data-Driven Approach to Modeling Physical Fatigue in the Workplace Using Wearable Sensors.” *Applied Ergonomics* 65: 515–529.

Martin, Dominik, Philipp Spitzer, and Niklas Kühl. 2020. “A New Metric for Lumpy and Intermittent Demand Forecasts: Stock-Keeping-Oriented Prediction Error Costs.” arXiv preprint arXiv:2004.10537.

Maurtua, Inaki, Izaskun Fernandez, Alberto Tellaeche, Johan Kildal, Loreto Susperregi, Aitor Ibarguren, and Basilio Sierra. 2017. “Natural Multimodal Communication for Human–Robot Collaboration.” *International Journal of Advanced Robotic Systems* 14 (4): Aritcle ID 17298814177 16043.

May, Gökan, Marco Taisch, Andrea Bettoni, Omid Maghazei, Annarita Matarazzo, and Bojan Stahl. 2015. “A New Human-Centric Factory Model.” *Procedia CIRP* 26: 103–108.

Mayer, Christoph, and Radu Timofte. 2020. “Adversarial Sampling for Active Learning.” In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3071–3079.

McTear, Michael, Zoraida Callejas, and David Griol. 2016. “Conversational Interfaces: Past and Present.” In *The Conversational Interface*, 51–72. Springer.

Meister, Sebastian, Mahdieu Wermes, Jan Stüve, and Roger M. Groves. 2021. “Investigations on Explainable Artificial Intelligence Methods for the Deep Learning Classification of Fibre Layup Defect in the Automated Composite Manufacturing.” *Composites Part B: Engineering* 224: Article ID 109160.

Meng, Lingbin, Brandon McWilliams, William Jarosinski, Hye-Yeong Park, Yeon-Gil Jung, Jehyun Lee, and Jing Zhang. 2020. “Machine Learning in Additive Manufacturing: A Review.” *Jom* 72 (6): 2363–2377.

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems.” *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11 (3–4): 1–45.

Nahavandi, Saeid. 2019. “Industry 5.0–A Human-centric Solution.” *Sustainability* 11 (16): 4371.

NIS. n.d. “Proposal for a Directive of the European Parliament and of the Council on Measures for a High Common Level of Cybersecurity Across the Union, Repealing Directive (EU) 2016/1148.” Accessed 15 March 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2020:823:FIN.

Oard, Douglas W., and Jinmook Kim. 1998. “Implicit Feedback for Recommender Systems.” In *Proceedings of the AAAI Workshop on Recommender Systems*, Vol. 83, 81–83. AAAI.

Oliff, Harley, Ying Liu, Maneesh Kumar, Michael Williams, and Michael Ryan. 2020. “Reinforcement Learning for Facilitating Human-Robot-Interaction in Manufacturing.” *Journal of Manufacturing Systems* 56: 326–340.

Padmanabhan, Raghav K., Vinay H. Somasundar, Sandra D. Griffith, Jianliang Zhu, Drew Samoyedny, Kay See Tan, and Jiahao Hu, et al. 2014. “An Active Learning Approach for Rapid Characterization of Endothelial Cells in Human Tumors.” *PloS One* 9 (3): Aritcle ID e90495.

Patki, N., R. Wedge, and K. Veeramachaneni. 2016, October. “The Synthetic Data Vault.” In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410.

Pedreschi, Dino, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Luca Pappalardo, Salvatore Ruggieri, and Franco Turini. 2018. “Open the Black Box Data-Driven Explanation of Black Box Decision Systems.” arXiv preprint arXiv:1806.09936.

Peruzzini, Margherita, Fabio Grandi, and Marcello Pellicciari. 2017. “Benchmarking of Tools for User Experience Analysis in Industry 4.0.” *Procedia Manufacturing* 11: 806–813.

Preece, A., W. Webberley, D. Braines, N. Hu, T. La Porta, E. Zaroukian, and J. Z. Bakdash. 2015. “SHERLOCK: Simple Human Experiments Regarding Locally Observed Collective Knowledge.” Technical Report. US Army Research Laboratory Aberdeen Proving Ground, USA.

Rajkumar, Ragunathan, Insup Lee, Lui Sha, and John Stankovic. 2010. “Cyber-Physical Systems: The Next Computing Revolution.” In *Design Automation Conference*, 731–736. IEEE.

Ranabhat, Bikash, Joseph Clements, Jacob Gatlin, Kuang-Ting Hsiao, and Mark Yampolskiy. 2019. “Optimal Sabotage Attack on Composite Material Parts.” *International Journal of Critical Infrastructure Protection* 26: Article ID 100301.

Reñones, Anibal, Davide Dalle Carbonare, and Sergio Gusmeroli. 2018. “European Big Data Value Association Position Paper on the Smart Manufacturing Industry.” *Enterprise Interoperability: Smart Services and Business Impact of Enterprise Interoperability*, 179–185.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Romero, David, Johan Stahre, and Marco Taisch. 2020. “The Operator 4.0: Towards Socially Sustainable Factories of the Future”.

Romero, David, Johan Stahre, Thorsten Wuest, Ovidiu Noran, Peter Bernus, Åsa Fast-Berglund, and Dominic Gorecky. 2016. “Towards an Operator 4.0 Typology: A Human-Centric Perspective on the Fourth Industrial Revolution Technologies.” In *Proceedings of the International Conference on Computers and Industrial Engineering (CIE46), Tianjin, China*, 29–31.

Rozanec, Joze M. 2021. “Explainable Demand Forecasting: A Data Mining Goldmine.” In *Companion Proceedings of the Web Conference 2021*, 723–724.

Rožanec, Jože M., Luka Bizjak, Elena Trajkova, Patrik Zajec, Jelle Keizer, Blaž Fortuna, and Dunja Mladenić. 2022. “Active Learning and Approximate Model Calibration for Automated Visual Inspection in Manufacturing.” arXiv preprint arXiv:2209.05486.

Rožanec, Jože M, Blaž Fortuna, and Dunja Mladenić. 2022. “Knowledge Graph-Based Rich and Confidentiality Preserving Explainable Artificial Intelligence (XAI).” *Information Fusion* 81: 91–102.

Rožanec, Jože M, Blaž Kažič, Maja Škrjanc, Blaž Fortuna, and Dunja Mladenić. 2021. “Automotive OEM Demand Forecasting: A Comparative Study of Forecasting Algorithms and Strategies.” *Applied Sciences* 11 (15): 6787.

Rožanec, Jože M., and Dunja Mladenić. 2021. “Reframing Demand Forecasting: A Two-Fold Approach for Lumpy and Intermittent Demand.” arXiv preprint arXiv:2103.13812.

Rožanec, Jože M, Elena Trajkova, Paulien Dam, Blaž Fortuna, and Dunja Mladenić. 2021. “Streaming Machine Learning and Online Active Learning for Automated Visual Inspection.” arXiv preprint arXiv:2110.09396.

Rožanec, Jože, Elena Trajkova, Inna Novalija, Patrik Zajec, Klemen Kenda, Blaž Fortuna, and Dunja Mladenić. 2022. “Enriching Artificial Intelligence Explanations With Knowledge Fragments.” *Future Internet* 14 (5): 134.

Rožanec, Jože, Patrik Zajec, Jelle Keizer, Elena Trajkova, Blaž Fortuna, Bor Brecelj, Beno Šircelj, and Dunja Mladenić. 2022 in review. “Enhancing Manual Revision in Manufacturing With AI-Based Defect Hints”.

Rožanec, Jože M., Patrik Zajec, Klemen Kenda, Inna Novalija, Blaž Fortuna, and Dunja Mladenić. 2021. “XAI-KG: Knowledge Graph to Support XAI and Decision-Making in Manufacturing.” In *International Conference on Advanced Information Systems Engineering*, 167–172. Springer.

Rožanec, Jože M., Patrik Zajec, Klemen Kenda, Inna Novalija, Blaž Fortuna, Dunja Mladenić, and Entso Veliou, et al. 2021. “STARdom: An Architecture for Trusted and Secure Human-Centered Manufacturing Systems.” In *IFIP International Conference on Advances in Production Management Systems*, 199–207. Springer.

Rožanec, Jože M, Patrik Zajec, Spyros Theodoropoulos, Inna Novalija, Klemen Kenda, Jelle Keizer, Paulien Dam, Blaž Fortuna, and Dunja Mladenić. 2022 in review. “Synthetic Data Augmentation Using GAN for Improved Automated Visual Inspection”.

Rožanec, Jože M., Patrik Zajec, Elena Trajkova, Beno Šircelj, Bor Brecelj, Inna Novalija, Paulien Dam, Blaž Fortuna, and Dunja Mladenić. 2022. “Towards a Comprehensive Visual Quality Inspection for Industry 4.0”.

Sadeghi, Leyla, Jean-Yves Dantan, Ali Siadat, and Jacques Marsot. 2016. “Design for Human Safety in Manufacturing Systems: Applications of Design Theories, Methodologies, Tools and Techniques.” *Journal of Engineering Design* 27 (12): 844–877.

Salkind, Neil J., and Terese Rainwater. 2006. *Exploring Research*. Upper Saddle River, NJ: Pearson Prentice Hall.

Samek, Wojciech, and Klaus-Robert Müller. 2019. “Towards Explainable Artificial Intelligence.” In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 5–22. Springer.

Samsonov, Vladimir, Johannes Lipp, Philipp Noodt, Alexia Fenollar Solvay, and Tobias Meisen. 2019. “More Machine Learning for Less: Comparing Data Generation Strategies in Mechanical Engineering and Manufacturing.” In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 799–807. IEEE.

Sanchez, Manuel, Ernesto Exposito, and Jose Aguilar. 2020. “Industry 4.0: Survey From a System Integration Perspective.” *International Journal of Computer Integrated Manufacturing* 33 (10-11): 1017–1041.

Schröder, Christopher, and Andreas Niekler. 2020. “A Survey of Active Learning for Text Classification Using Deep Neural Networks.” *arXiv preprint arXiv:2008.07267*.

Schweichhart, Karsten. 2016. “Reference Architectural Model Industrie 4.0 (Rami 4.0).” An Introduction. Online: https://www.plattform-i40.de I 40.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. “Grad-Cam: Visual explanations from Deep Networks via Gradient-Based Localization.” In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Senoner, Julian, Torbjørn Netland, and Stefan Feuerriegel. 2021. “Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing.” *Management Science* 68: 5704–5723.

Serradilla, Oscar, Ekhi Zugasti, Carlos Cernuda, Andoitz Aranburu, Julian Ramirez de Okariz, and Urko Zurutuza. 2020. “Interpreting Remaining Useful Life Estimations Combining Explainable Artificial Intelligence and Domain Knowledge in Industrial Machinery.” In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. IEEE.

Settles, Burr. 2009. “Active learning literature survey”.

Seung, H. Sebastian, Manfred Opper, and Haim Sompolinsky. 1992. “Query by Committee.” In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 287–294.

Shneiderman, Ben. 2020. “Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems.” *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10 (4): 1–31.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2017. “Learning Important Features Through Propagating Activation Differences.” In *International Conference on Machine Learning*, 3145–3153. PMLR.

Silaghi, Helga, Ulrich Rohde, Viorica Spoială, Andrei Silaghi, Eugen Gergely, and Zoltan Nagy. 2014. “Voice Command of an Industrial Robot in a Noisy Environment.” In *2014 International Symposium on Fundamentals of Electrical Engineering (ISFEE)*, 1–5. IEEE.

Sinha, Samarth, Sayna Ebrahimi, and Trevor Darrell. 2019. “Variational Adversarial Active Learning.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5972–5981.

Soldatos, John, Oscar Lazaro, Franco Cavadini, Filippo Boschi, Marco Taisch, and Paola Maria Fantini, et al. 2019. *The Digital Shopfloor: Industrial Automation in the Industry 4.0 Era. Performance Analysis and Applications*, River Publishers Series in Automation, Control and Robotics.

Sovrano, Francesco, and Fabio Vitali. 2021. “An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability.” arXiv preprint arXiv:2109.05327.

Spelt, Philip F, H. E. Knee, and Charles W Glover. 1991. “Hybrid Artificial Intelligence Architecture for Diagnosis and Decision-Making in Manufacturing.” *Journal of Intelligent Manufacturing* 2 (5): 261–268.

Suh, Sungho, Vitor Fortes Fortes Rey, and Paul Lukowicz. 2022a. “Adversarial Deep Feature Extraction Network for User Independent Human Activity Recognition.” In *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 217–226. IEEE.

Suh, Sungho, Vitor Fortes Rey, and Paul Lukowicz. 2022b. “TASKED: Transformer-Based Adversarial Learning for Human Activity Recognition Using Wearable Sensors via Self-KnowledgE Distillation.” arXiv preprint arXiv:2209.09092.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. “Axiomatic Attribution for Deep Networks.” In *International Conference on Machine Learning*, 3319–3328. PMLR.

Tao, Wenjin, Ze-Hao Lai, Ming C. Leu, and Zhaozheng Yin. 2018. “Worker Activity Recognition in Smart Manufacturing Using IMU and SEMG Signals with Convolutional Neural Networks.” *Procedia Manufacturing* 26: 1159–1166.

Tao, Wenjin, Ming C. Leu, and Zhaozheng Yin. 2020. “Multi-Modal Recognition of Worker Activity for Human-Centered Intelligent Manufacturing.” *Engineering Applications of Artificial Intelligence* 95: Article ID 103868.

Tao, Fei, and Meng Zhang. 2017. “Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing.” *IEEE Access* 5: 20418–20427.

Thomopoulos, Nick T. 2015. “Demand Forecasting for Inventory Control.” In *Demand Forecasting for Inventory Control*, 1–10. Springer.

Trajkova, Elena, Jože M. Rožanec, Paulien Dam, Blaž Fortuna, and Dunja Mladenić. 2021. “Active Learning for Automated Visual Inspection of Manufactured Products.”

Vajpai, Jayashri, and Avnish Bora. 2016. “Industrial Applications of Automatic Speech Recognition Systems.” *International Journal of Engineering Research and Applications* 6 (3): 88–95.

Verma, Sahil, John Dickerson, and Keegan Hines. 2020. “Counterfactual Explanations for Machine Learning: A Review.” arXiv preprint arXiv:2010.10596.

Villani, Valeria, Massimiliano Righi, Lorenzo Sabattini, and Cristian Secchi. 2020. “Wearable Devices for the Assessment of Cognitive Effort for Human–Robot Interaction.” *IEEE Sensors Journal* 20 (21): 13047–13056.

Wan, Jiafu, Xiaomin Li Hong-Ning Dai, Andrew Kusiak, Miguel Martínez-García, and Di Li. 2020. “Artificial-Intelligence-Driven Customized Manufacturing Factory: Key Technologies, Applications, and Challenges.” *Proceedings of the IEEE* 109 (4): 377–398.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. “Image Quality Assessment: From Error Visibility to Structural Similarity.” *IEEE Transactions on Image Processing* 13 (4): 600–612.

Wang, Yichuan, Mengran Xiong, and Hossein Olya. 2020. “Toward an Understanding of Responsible Artificial Intelligence Practices.” In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 4962–4971. Hawaii International Conference on System Sciences (HICSS).

Wegner, Andre, James Graham, and Eli Ribble. 2017. “A New Approach to Cyberphysical Security in Industry 4.0.” In *Cybersecurity for Industry 4.0*, 59–72. Springer.

Wei, Taoran, Danhua Cao, Xingru Jiang, Caiyun Zheng, and Lizhe Liu. 2020. "Defective Samples Simulation Through Neural Style Transfer for Automatic Surface Defect Segment." In *2019 International Conference on Optical Instruments and Technology: Optoelectronic Measurement Technology and Systems*, Vol. 11439, 1143904. International Society for Optics and Photonics.

Wellsandta, Stefan, Zoltan Rusak, Santiago Ruiz Arenas, Doris Aschenbrenner, Karl A. Hribernik, and Klaus-Dieter Thoben. 2020. "Concept of a Voice-Enabled Digital Assistant for Predictive Maintenance in Manufacturing."

Wilson-Donnelly, Katherine A., Heather A. Priest, Eduardo Salas, and C. Shawn Burke. 2005. "The Impact of Organizational Practices on Safety in Manufacturing: A Review and Reappraisal." *Human Factors and Ergonomics in Manufacturing & Service Industries* 15 (2): 133–176.

Wu, Dongrui. 2018. "Pool-Based Sequential Active Learning for Regression." *IEEE Transactions on Neural Networks and Learning Systems* 30 (5): 1348–1359.

Wu, Mingtao, and Young Moon. 2018. "DACDI (Define, Audit, Correlate, Disclose, and Improve) Framework to Address Cyber-Manufacturing Attacks and Intrusions." *Manufacturing Letters* 15: 155–159.

Wu, Dazhong, Anqi Ren, Wenhui Zhang, Feifei Fan, Peng Liu, Xinwen Fu, and Janis Terpenny. 2018. "Cybersecurity for Digital Manufacturing." *Journal of Manufacturing Systems* 48: 3–12.

Xu, Li Da, and Lian Duan. 2019. "Big Data for Cyber Physical Systems in Industry 4.0: A Survey." *Enterprise Information Systems* 13 (2): 148–169.

Xu, Dianlei, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma, Tao Jiang, Jon Crowcroft, and Pan Hui. 2020. "Edge Intelligence: Architectures, Challenges, and Applications." arXiv preprint arXiv:2003.12172.

Xu, Xun, Yuqian Lu, Birgit Vogel-Heuser, and Lihui Wang. 2021. "Industry 4.0 and Industry 5.0–Inception, Conception and Perception." *Journal of Manufacturing Systems* 61: 530–535.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. "Modeling Tabular Data Using Conditional Gan." *Advances in Neural Information Processing Systems* 32: 7335–7345.

Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. "Federated Machine Learning: Concept and Applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (2): 1–19.

Yu, Zhenhua, Lijun Zhou, Zhiqiang Ma, and Mohammed A El-Meligy. 2017. "Trustworthiness Modeling and Analysis of Cyber-Physical Manufacturing Systems." *IEEE Access* 5: 26076–26085.

Yue, Xiaowei, Yuchen Wen, Jeffrey H. Hunt, and Jianjun Shi. 2020. "Active Learning for Gaussian Process Considering Uncertainties with Application to Shape Control of Composite Fuselage." *IEEE Transactions on Automation Science and Engineering* 18 (1): 36–46. doi:10.1109/TASE.2020.2990401.

Zajec, Patrik, Jože M. Rožanec, Elena Trajkova, Inna Novalija, Klemen Kenda, Blaž Fortuna, and Dunja Mladenić. 2021. "Help Me Learn! Architecture and Strategies to Combine Recommendations and Active Learning in Manufacturing." *Information* 12 (11): 473.

Zavrtanik, Vitjan, Matej Kristan, and Danijel Skočaj. 2021. "DRAEM-A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330 –8339.

Zeltmann, Steven Eric, Nikhil Gupta, Nektarios Georgios Tsoutsos, Michail Maniatakos, Jeyavijayan Rajendran, and Ramesh Karri. 2016. "Manufacturing and Security Challenges in 3D Printing." *Jom* 68 (7): 1872–1881.

Zeng, Xinchuan, and Tony R. Martinez. 2000. "Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation." *Journal of Experimental & Theoretical Artificial Intelligence* 12 (1): 1–12.

Zhu, Jia-Jie, and José Bento. 2017. "Generative Adversarial Active Learning." arXiv preprint arXiv:1702.07956.