Towards the application of machine learning in digital twin technology: a multi-scale review

Discover Applied Sciences

Review

Towards the application of machine learning in digital twin technology: a multi-scale review

Luigi Nele[1] · Giulio Mattera[1] · Emily W. Yap[2] · Mario Vozza[3,4] · Silvestro Vespoli[1]

Received: 15 July 2024 / Accepted: 10 September 2024
Published online: 19 September 2024

Abstract
This review article delves into the conceptual framework of digital twins and their diverse applications across research domains, highlighting the pivotal role of machine learning in shaping the development and integration of digital twin technology across multiple disciplines. Emphasising key features like multidisciplinarity and multi-scale aspects, the paper explores how data-driven techniques are employed for modelling, visualisation, monitoring, and optimisation within the digital twin framework, pinpointing the benefits introduced in the current state-of-the-art applications, and elucidates persisting challenges across various research fields, including advanced materials, smart buildings, and manufacturing systems.

Keywords Digital twin · Advanced statistics · Machine learning · Materials · Smart buildings · Manufacturing

1 Introduction

Nowadays, advancements in various technologies field such as the Internet of Things (IoT) and Artificial Intelligence (AI) facilitated the digitalisation of assets across diverse industrial sectors. In particular, Digital Twins (DTs) represent a disruptive technology that can be synthesised as an integrated multi-physics, multi-scale, probabilistic simulations of a physical asset that leverages complex physical models, sensor data and historical information is able to replicate the behaviour of their real-world counterparts [1, 2].

DTs, integrated within Cyber Physical Systems (CPS) [3, 4], can be used for different goals like feedback control, asset optimisation, visualisation and support to decision-making [5, 6].

The concept of a DT surpasses that of a digital model, as demonstrated in Fig. 1. A digital model enables a unidirectional flow of data, originating from the physical object and feeding into its digital representation. The digital model adjusts itself based on input from the physical asset, without directly intervening in the physical entity. Additionally, a digital model can be utilised offline to simulate what-if scenarios, facilitating the optimisation of the physical asset's performance. Conversely, Digital Twins possess bidirectional communication capabilities [7, 8] allowing them to communicate with the physical entity via automatic decisions—based on real-time events happening in the physical world—or via visualisation with human users. A typical example of a DT application is the predictive maintenance in manufacturing systems. In this case, the DT collect data from the physical entity and elaborate it in the digital world. If an anomalous

✉ Luigi Nele, nele@unina.it | [1]Department of Chemical, Materials and Industrial Manufacturing Engineering, University of Naples Federico II, Naples, Italy. [2]Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2522, Australia. [3]Department of Control and Computer Engineering (DAUIN), Polytechnic University of Turin, Turin, Italy. [4]DAIMON Lab, CNR-ISMN, Bologna, Italy.

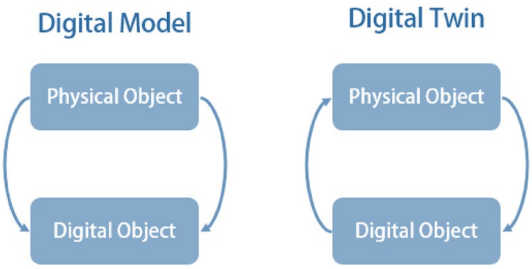

Fig. 1 Comparison between Digital Model and Digital Twin: Visual depiction highlighting unidirectional data flow and offline scenario analysis in digital models, contrasted with bidirectional communication and real-time decision-making capabilities in digital twins

event is detected by a monitoring module [9], the process can be stopped or the main parameters can be changed to adjust the process [10], interacting with the physical entity and closing the bi-directional communication in an autonomous manner. In alternative, the bidirectional communication happens with users via 3D visualisation or Graphical Unit Interface [11].

Core components of a DT comprise modelling, simulation, monitoring, data interpretation through visualisation, and decision-making processes [12]. A general DT framework inspired by the architectures proposed by Cimino et al. [13] and Qian et al. [14] is structured with distinct layers, as illustrated in Fig. 2, each serving a crucial role in its operation. The *perception layer*, for instance, acts as the initial stage where data is gathered from the physical object and undergoes pre-processing. Subsequently, the *digital object layer* utilises this processed information to forecast system states and estimate variables that may not be directly measurable. Finally, the *application analysis* layer integrates all collected and synthesised data from both perception layer the digital model, employing techniques such as visualisation, anomaly detection, classification, feedback control, and optimisation to develop the visualisation with the end-user or the decision-making on the physical entity part of the communication scheme. This comprehensive framework enables the DT to mirror and enhance the performance of its physical counterpart.

In the presented layered framework of a DT Machine Learning (ML) emerge as pivotal components, facilitating the development of data-driven or grey digital models and synthesising the acquired big data [15] collected from integrated sensors in the physical asset to provide actionable insights to operators and machine. Generally, it is possible to refer as white box model, the one grounded on physics principles, e.g. Finite Element Analysis (FEA) [16, 17] or lumped models [18]. The black box model, in contrast, relies solely on data-driven techniques, drawing conclusions from patterns and correlations in the data without requiring an understanding of the underlying system. The grey box model integrates both methods, using data

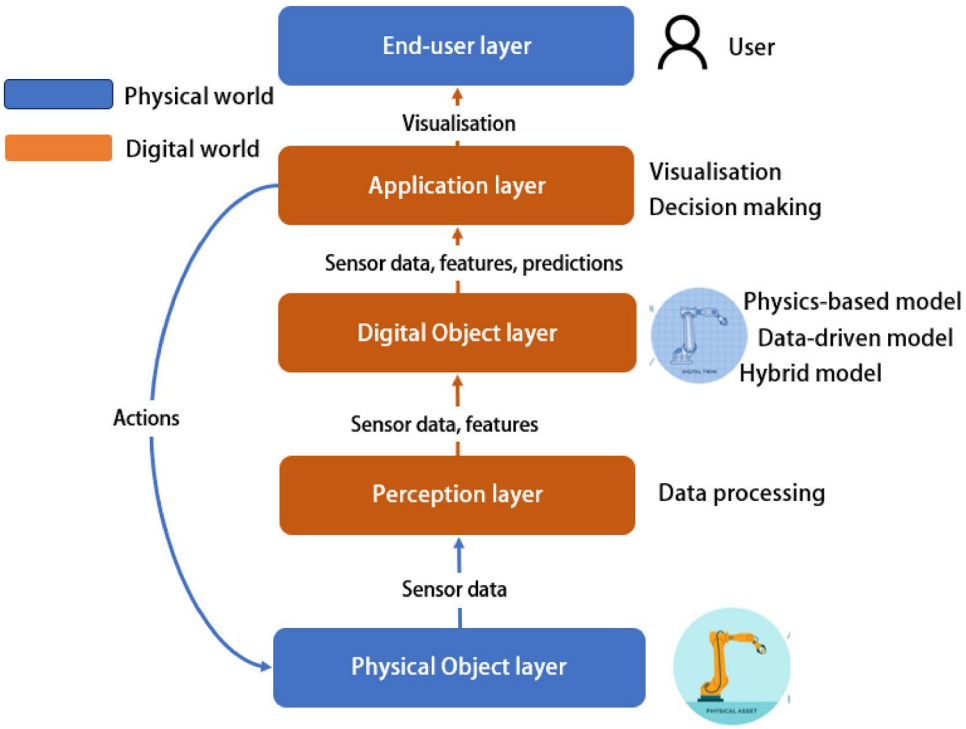

Fig. 2 The layered framework of a Digital Twin, comprising perception, digital object, and application layers, facilitates data collection, prediction, and advanced analytics for applications such as visualisation, anomaly detection, classification, feedback control, and optimisation

from FEA or other simulations to augment datasets and improve the accuracy of data-driven analyses, or combining FEA outputs with sensor data coming from the process [19].

Moreover, the software modules developed using ML methods and present in the application layer can trigger system alerts in the event of anomalies, while stochastic optimisation techniques can optimise system performance, employing actions to the physical entity. These applications are enhanced by the communication capabilities of the IoT, enabling data exchange and interaction between the digital twin and its physical counterpart. Progress in this domain promote real-time updates and synchronisation, thereby augmenting effectiveness of DT in reflecting real-world behaviour.

The applications of DTs spans across various industries, exerting a significant impact on fields including smart cities [20, 21], construction [22–24], healthcare [25] and manufacturing.

DTs encompass real-time state monitoring, energy consumption analysis and optimisation to enhance efficiency, sustainability and reducing waste. Furthermore, product failure analysis and prediction for proactive maintenance applications may be developed to further enhance the product quality. Additionally, DTs serve as valuable training tools for operational personnel, providing immersive learning experiences and simulating various scenarios for skill development and readiness [26, 27]. DT frameworks vary across domains due to their specialised applications. However, ML is universally employed as the core technology, building upon a common underlying framework, presented in its generic form in Fig. 2. This study emphasizes the multi-scale nature of digital twins, showcasing how ML principles are applied across different scales, from materials to production plants.

2 Machine learning for digital twin technology

The concept of DT was introduced by Grieves in 2003, and in recent years, owing to advancements in hardware technology and computer science fields, its implementation has become feasible. Consequently, in 2012, NASA proposed a first framework of a DT for space missions, defining it as an ultra-realistic model of a physical asset that evolves alongside it, leveraging data from sensors integrated into the physical asset [28]. However, in the last year the concept of DT evolved, and self-monitoring and self-control capabilities have introduced as additional features, especially in sector like smart buildings and manufacturing, driven by technological innovations which make this possible. Nowadays, DT technology finds application across various scientific domains to enhance product quality and reducing waste linked with producing faulty parts or sub-optimal product realisation [29, 30].

As highlighted in the introduction, modelling methods have long stood as pivotal technologies in this field. However, recent advancements in computer science and IoT have brought forth other Key Enabling Technologies in this domain, such as communication and big data analytics, with ML assuming a critical role.

Machine learning (ML) is a subset of Artificial Intelligence (AI) that involves the development of algorithms and statistical models enabling computers to perform tasks without explicit instructions. By utilizing patterns and inference derived from data, ML algorithms can improve their performance over time on a specific task. The process typically involves training on a dataset to create a predictive model, which can then make decisions or predictions based on new, unseen data. This field encompasses various approaches, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning, each tailored to different types of problems and data structures. As reported in Fig. 3**,** once data are collected from physical processes, data can be pre-processed and used to train ML algorithm that then give their output to Decision Support System (DDS). This output can come from defect detection modules [31], a predictive model [32] or general monitoring systems [33].

From a modelling standpoint, a DT entails the translation of physical entities into virtual space, with the objective of closely mirroring the behaviour of the real system through its virtual representation. In this scenario, both physics-based (white-box) [34] and data-driven (black-box) [35] modelling techniques find application as well as hybrid techniques defined as grey-model, as shown in Fig. 4. However, the computational demands inherent to DT [36] often necessitate a heavier reliance on data-driven approaches over physical ones for system modelling.

If data-driven techniques are utilised for modelling and simulating complex systems to develop DTs, statistics and ML assume a pivotal role. This is due to the diverse array of techniques available, enabling the creation of both static and dynamic models through regression analysis [37–40].

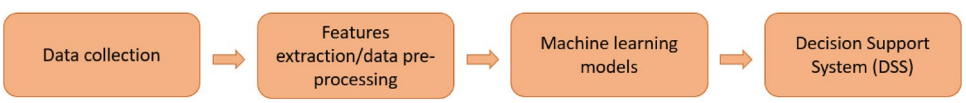

Fig. 3 Data analytics workflow typical of Machine Learning

Fig. 4 Physics-based models in which 3D geometry and first principles are used and data-driven modelling techniques may be employed to develop virtual replica of physical assets

Nevertheless, as mentioned above DTs demand additional capabilities, including real-time monitoring [41, 42] and data visualisation [43]. Also in this case ML offer a solution for processing the Big Data coming from physical asset, identifying unstable conditions or anomalies through the use of classification and clustering techniques [44]. Furthermore, dimensionality reduction techniques facilitate data visualisation, providing users with a comprehensive understanding of the outcomes derived from self-monitoring and self-control applications.

From a theoretical perspective, DTs can be developed across various scales, ranging from micro-systems, such as chemical molecules, to macro-systems like smart cities or production plants. In addition to the space-scale, a DT may contain temporal scales as well, given the different dynamic nature of the different components of a system. As an example, considering a production system in Fig. 5, a DT can be developed for the whole production plant, for the single machine in the line and for the single process. This represents several space-scale of the DT. Furthermore, additional scales like the time scale can be used to study the dynamic behaviour of the different space-scales.

It is true that complex systems exhibit a hierarchical structure from a model perspective, delineated into multiple layers based on their granularity. Similarly, in the temporal domain, models can be static, capturing a snapshot of the system at a given moment, or they can be dynamic, capturing the changes over time. Typical layers of a multi-scale DT model comprise the unit layer, system layer, and system of system (SoS) layer, each serving distinct purposes and encompassing different levels of abstraction and complexity. The general framework depicted in Fig. 2 can be adapted for each proposed time or spatial scale layer of this multi-scale framework. Despite this broad spectrum, the underlying methodologies employed to develop DTs at different scales share common foundations. In particular, ML find application across all system scales and various applications within DT development.

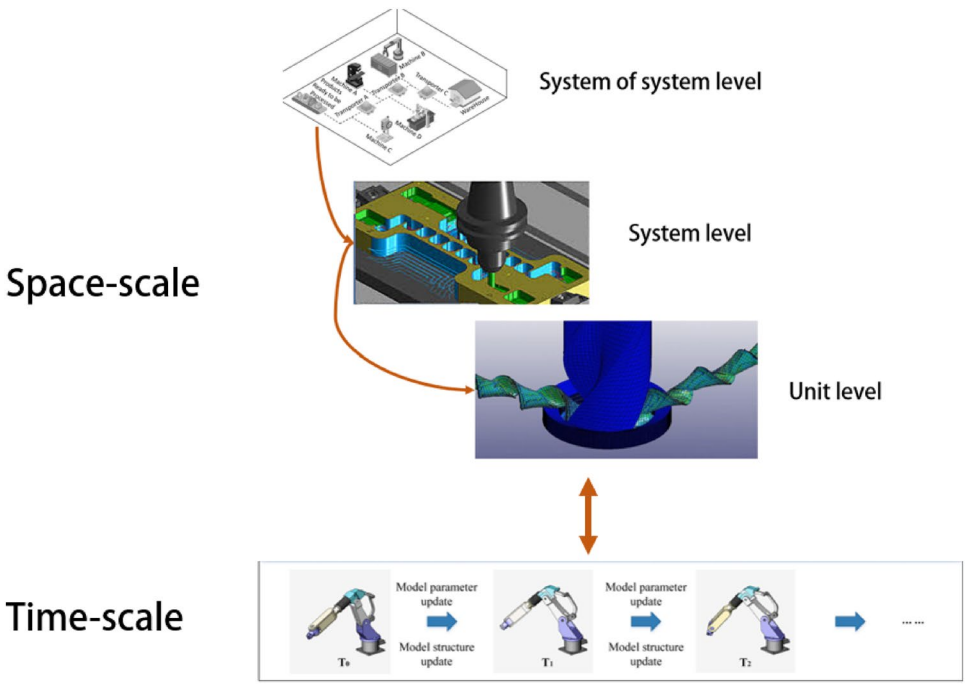

Fig. 5 The framework of multi-scale modelling inspired by Zhang et al. [45]

2.1 Perception layer

The perception layer constitutes a fundamental element within the framework of a DT, serving as the gateway for real-time data acquisition and preprocessing. At its core, the perception layer is responsible for sensing and collecting data from a multitude of sources, including sensors, IoT devices, and other data-generating entities embedded within the physical environment. This layer of the DT is responsible for capturing information regarding the state, behaviour, and performance of the physical asset or system that it represents.

Beyond data acquisition, the perception layer also undertakes essential pre-processing tasks to ensure that the incoming data is filtered, and standardised or scaled before being passed on to subsequent layers of the DT framework. This pre-processing activity may involve data fusion, noise reduction, outlier detection, and data normalisation techniques to enhance the quality and reliability of the acquired data.

2.1.1 Sensor fusion

Sensor fusion is the practice of integrating data from multiple sensors to enhance the accuracy, reliability, and resilience of estimating a system's state [46]. The effectiveness of sensor fusion is challenged by uncertainties inherent in sensor measurements, such as noise, biases, and environmental disturbances. To menage these uncertainties probabilistic methods, such as Kalman Filters (KFs) can be employed [47]. These filters iteratively update the system state estimate based on sensor measurements while considering the uncertainties stemming from both the sensors and the system dynamics. Although Kalman Filter is widely used as probabilistic method for sensor fusion analysis, also regression analysis can be a valuable tool in sensor fusion applications, particularly when integrating data from multiple sensors to estimate a target variable or system state.

In this case, considering $[x_{k1}, x_{k2}, \ldots, x_{kn}]$ sensors measurement, it is possible to weight all the sensors measurements aiming to estimate another variable y_k. If physical models are not present for this task, data-driven approaches like Ordinary Last Squares (OLS) estimator or Lp-Norm estimator can be used to estimate the optimal weights that reduce the estimation error, for the general linear regression equation in Eq. (1) and OLS method in Eq. (2) [48], where W_{kn} is the matrix of weights of dimension KxN which multiplied by the sensor measurement x_n of dimension N and sum to the bias vector of dimension K allow to obtain via linear combination the estimated variable y_k of dimension K.

$$y_k = W_{kn}x_n + \in_k \tag{1}$$

$$W_{kn} = \left(x_n^T \cdot x_n\right)^{-1} \cdot x_n^T \cdot y_k \tag{2}$$

In alternative to simple linear combination methods, more sophisticated techniques based on ML can be employed for estimate system states that exhibit a complex and nonlinear relationship with other measured variables, such as Bayesian methods[49], Support Vector Machine [50] and Neural Networks [51, 52].

2.1.2 Outlier rejection

When employing data-driven techniques to develop models for DTs, the quality of the data becomes paramount. It's crucial to ensure that the data used for training models is representative and devoid of out-of-distribution data or outliers. To achieve this, preprocessing steps are essential, involving the removal of such data points from the dataset. Additionally, outlier rejection techniques can be applied not only during model training but also within the monitoring module. By doing so, anomaly detection and fault analysis can be enhanced, leveraging the data obtained from the process to improve the overall reliability and performance of the DT system. In the field of outlier rejection several methods can be employed such as distribution based (Covariance estimation), clustering based (DBSCAN) or more complex techniques such as Isolation Forest.

A simple approach to outlier detection involves assuming that the regular data follow a known distribution, such as a Gaussian distribution. Based on this assumption, which aim to characterise the shape of the data, observations that deviate significantly from this expected shape can be identified. The minimum covariance determinant (MCD) method of Rousseeuw [53] is usually used to estimate the covariance matrix Σ of the Gaussian-shaped data, while Mahalanobis

distance (D_M) in Eq. (3), is used to define the outliers with respect to estimated covariance and the mean or median value (μ).

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{3}$$

Although this method is simple, complex distributed data can reduce the performance in discovery outliers. For this reason, other techniques like Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [54] and Isolation Forest [55, 56] can be used. DBSCAN is a clustering algorithm specialised in discover clusters of arbitrary shape in spatial datasets and it inherently excludes outliers from clusters by classifying them as noise points, which remain unassigned. Its density-based approach enhances robustness to outliers, prioritising the identification of dense regions over fitting every data point into a cluster, unlike techniques such as KMeans. Isolation Forest is a specialized anomaly detection algorithm based on Random Forest. It isolates outliers by randomly selecting features and splitting data based on random values within those features. The number of splits needed to isolate a sample is used as metrics to identify an anomaly in the dataset, since shorter path lengths are typically produced for anomalies due to random partitioning.

2.1.3 Data pre-processing

The autonomous level of DTs require automated self-learning processes, in which complex statistical learning techniques are increasing in their utilisation. Several studies have demonstrated the requirement of pre-processing the data to enhance the algorithm performances, once features have been extracted from raw data.

Especially in the field of statistical learning, closely distributed feature values facilitate faster and more effective training. Scaling methods, such as normalisation, reduces the distance between data units by making them more suitable for statistical methods which uses distance metrics to evaluate outliers or to classify samples. Another scaling method employed is standardisation, which allow to uniform the distribution of the data points, which can be beneficial for density-based methods [57].

Being more precise, Normalisation, also known as Min–Max scaling, rescales the features of a dataset to a fixed range, usually between 0 and 1. It adjusts the values proportionally so that the minimum value of the feature becomes 0, and the maximum value becomes 1.

The normalisation ensures that all features have the same scale, preventing one feature from dominating the others during training. Additionally, this helps algorithms converge faster, especially gradient-based algorithms like neural networks. On other hand, Standardisation, also known as Z-score normalisation, transforms the features of a dataset to have a mean of 0 and a standard deviation of 1. It centres the data around 0 and scales it to have a unit variance.

The standardisation makes the distribution of each feature more symmetrical and Gaussian-like, which can improve the performance of certain algorithms. Furthermore, it mitigates the effects of outliers, as the scale of the data is based on the standard deviation rather than the range of values. The results obtained using scaling techniques is shown in Fig. 6**.**

2.2 Digital object layer

Modelling and simulation are the most important components of a DT, providing the framework for creating the virtual replicas of physical entities. This component encompasses the development of mathematical models, algorithms, and computational techniques to mimic the behaviour and dynamics of real-world systems with high accuracy. Through

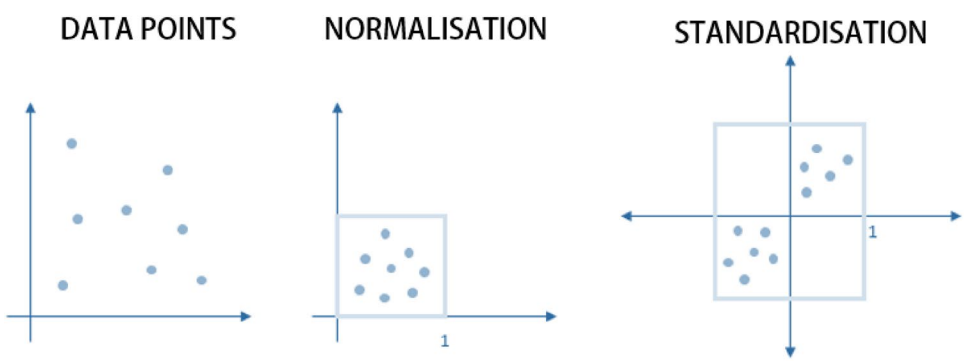

Fig. 6 Comparison between normalisation and standardisation scaling techniques

sophisticated modelling techniques, DTs can accurately capture the complexities and intricacies of physical assets, enabling predictive analysis, scenario testing, and optimisation of the physical asset. The mathematical model can be constructed using either high-fidelity physical simulations, such as Finite Element Analysis (FEA), or data-driven techniques that learn underlying system relationships from input–output data. The latter approach leverages pre-processed data generated by the perception layer in terms of both raw data coming from sensors and extracted features.

2.2.1 Physical and data-driven models

Physical models—also known as white-box model—are based on the principles of physics and are often derived from first principles and expressed through derivative equations. These equations can manifest in various forms, ranging from intricate 3D finite element models to simplified lumped models, yielding both detailed and generalised results [58]. However, despite their strengths, physical models are not immune to limitations. They necessitate a complex knowledge of system parameters, material properties, and boundary conditions, which may not always be readily available or precisely known. This reliance on specific data can lead to the creation of overly complex models that may not accurately reflect the complexities and uncertainties present in the real world. Moreover, the computational demands of physical models can present significant challenges, especially when simulating large-scale or highly dynamic systems. Balancing the need for model fidelity with computational efficiency often requires engineers and researchers to resort to simplifications or approximations.

On other hand, data-driven models are emerging as powerful tools that, unlike their physical counterparts, allow to syntheses an input–output hidden relationship using the data captured during experimental campaigns or during the life of the products/systems. At the heart of data-driven modelling lies a diverse array of techniques designed to extract insights from data such as regression, auto-regression and classification analyses. While data-driven models offer immense promise, they are not without challenges. Despite their ability to develop complex models for complex systems using the data, several additional challenges exist such as low generalisation outside observed data (or overfitting) and low interpretations of the underlying mechanisms driving the observed phenomena, which resulted in these models being categorised as black-box. Overfitting is a critical challenge in ML that occurs when a model becomes overly complex and learns the training data too closely, compromising its ability to generalise to new, unseen data. This phenomenon is characterised by high performance on the training set but poor performance on the validation set, and several solutions can be used to reduce this problem, such as regularisation or early stopping [59].

Overfitting is often a challenge in DT development due to limited test data in particular extreme or unsafe operating conditions. This data scarcity in this region often results in a dataset with low variance, increasing the risk of overfitting despite its size. In Fig. 7 is shown how low-variance dataset within boundary region can reduce the prediction performance of a DT. To address this, incorporating grey-box models can augment the dataset with simulated data generated from physical models, especially in data-sparse regions such as boundary conditions, which are complex to obtain in the real-world. This hybrid approach improves model performance and generalizability thanks to the improvement of dataset variance.

2.2.2 Polynomial regression for static model

Regression analysis is a data-driven modelling technique which allow to find a relationship between the dependent variables x and the targets y. Mathematically a regression task can be described as in Eq. (4), in which a function f_θ with parameters θ is used to map the input with the output.

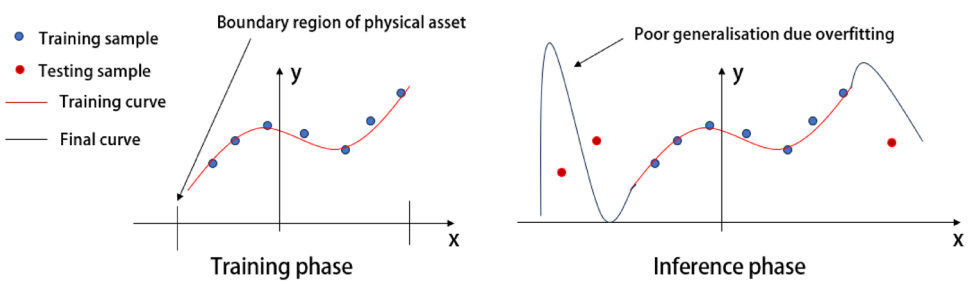

Fig. 7 The poor generalisation introduced by overfitting can drastically reduce the performance of DT model within boundary regions

$$y_k = f_\theta(x_n) \tag{4}$$

The most used regression model is the linear regression, where the relationship is assumed to be linear as reported in Eq. 1. Since the regression is a supervised learning approach, the weights W multiplied by the input vector x can be estimated using the knowledge of the desired outcome (y). As introduced in the previous sections, OLS estimator, Lp estimator methods can be used to estimate the weights W. However, also gradient descent optimisation [60] methods can be used to optimise the parameters W in a linear regression or can be used to estimate the parameters θ in a more complex form of differnetiable regression function. Although linear regression is simple to train, it is limited to capturing linear relationships between data. To address the linear relationship problem between data, Polynomial regression, as expressed in Eq. 5, can be employed, which assumes an nth degree polynomial model between the independent variables and the dependent variable, which allow this model to capture non-linearities in the model.

$$y_k = \beta_0 + \beta_1 x_n + \beta_2 x_n + \cdots + \beta_n x_n^n + \in \tag{5}$$

where y is the dependent variable, x is the independent variable, β_n are the regression coefficients, n is the degree of the polynomial, and $\in$ represents the error term. The degree of the polynomial determines the complexity of the relationship that can be captured. Higher-degree polynomials can fit the data more closely but may risk overfitting, especially with limited data, so it is essential to strike a balance between model complexity and the risk of overfitting. As in standard linear regression, several methods can be used to estimate the parameters β. If no correlation between inputs is assumed, the most popular method used is the OLS estimator.

2.2.3 Auto Regressive Polynomial model for dynamic model

An Autoregressive Regression (AR), is a statistical model used in time series analysis to model dynamic systems, in which the current state depends on previous states or observations. Mathematically, a linear AR model of order p, denoted as AR(p), can be represented as in Eq. 6.

$$y_t = c + \phi_1 y_{t-1} + \cdots \phi_p y_{t-p} + \in_t \tag{6}$$

where yt is the value of the variable at time t, c is a constant term, ϕ_p are the autoregressive parameters representing the effects of the previous p values on the current value and $\in_t$ is the error term assumed to be independent and identically distributed with mean zero and constant variance.

AR models can use both linear and polynomial regression models and the parameters of the p lags can be found using OLS or gradient-based techniques. However, most of physical dynamic systems change their behaviour subject to external control actions or eXogenous variables, which are independent by target variable y.

Mathematically, an ARX model (AutoRegressive with eXogenous variable) can be described as in Eq. 7, in which, β_q are the parameters representing the effects of the exogenous variables, $x_{q,t}$ on the current value.

$$y_t = c + \phi_1 y_{t-1} + \cdots \phi_p y_{t-p} + \in_t + \beta_1 x_{1,t} + \cdots + \beta_q x_{n,q} + \in_t \tag{7}$$

As for polynomial regression, also the auto regression requires careful consideration of polynomial degree in both input and time dimensions. To prevent overfitting and ensure good generalisation, a balance must be found between model complexity and data fit through manual tuning or specific domain knowledge.

2.2.4 Neural Networks for regression and classification

Neural networks are computational models inspired by the human brain's structure that consist of interconnected nodes, or neurons, organised into multi-layers structures. Each neuron receives inputs, processes them through a non-linear combination described as in Eq. 8, and produces an output that represent the input for the following layers.

$$y_k = f_\theta(x_n) = g(Wx) \tag{8}$$

where g is a non-linear function, or activation function, that allow to non-linearly combine the linear combination of the input of the previous layer. In this architecture each neuron is responsible for the non-linear combination of the input coming from the previous layer using the described equation. Typical activation functions are the sigmoid, the hyperbolic

tangent and Rectified Linear Unit and each of them is able to capture some degree of non-linearity in the data, leading into different results [61].

Backpropagation is a key algorithm used to train neural networks [62].

It works by iterative adjusting the weights of connections between neurons to minimise the difference between the network's output and the desired output. By continuously updating the weights based on the error, the network learns to make better predictions over time. Neural Networks models can be used to solve both regression and classification tasks and as the polynomial regression models are largely used in several industrial fields for modelling static [63–65] and dynamic [66] systems, classify defects [67, 68], optimisation [69] and discover anomalies [70]. Although the basic architecture is common for both regression and classification, what is different is the output layer. In a regression task ReLu or linear activation are generally employed in the output layer, while sigmoid or softmax are employed for classification task, converting the output layer into class-probability. Complex non-linear dynamic systems can be described by AutoRegressive or ARX Neural Network, as shown in Fig. 8, if the input are the lagged values of the physical variables.

Unlike the previously presented traditional models, neural networks, and particular some complex architectures such as Recurrent Neural Networks (RNNs) [71], Long Short-Term Memory (LSTM) [72], and Gated Recurrent Units (GRU) [73], do not necessitate predefined structural assumptions about the system in either time or input spaces. Nevertheless, tuning the substantial number of hyperparameters in these models can be challenging. However, recent advancements in automated machine learning (AutoML), facilitated by libraries like Optuna, offer efficient hyperparameter optimisation [74].

2.3 Data integration in digital twin

In the modern digital era, effective data management is crucial for streamlining industrial operations and enabling data-driven decision-making. With the advent of digital transformation, industries utilise advanced technologies to facilitate seamless data flow throughout the production cycle. However, alongside the benefits, the escalating complexity of data management and production presents significant challenges for industrial assets [75–78].

In navigating this landscape, the integration of concepts like Asset Administration Shell (AAS) and Digital Twin (DT) [79, 80] emerges as a strategic approach. By seamlessly connecting with AAS, DTs can access real-time data streams and operational parameters of physical assets, enabling predictive modelling and real-time monitoring for enhanced operational efficiency and performance optimisation. Thus, in the pursuit of digitalisation, the symbiotic relationship between efficient data management, AAS, and DT becomes increasingly indispensable for modern industries [81]. AAS and DT are closely interconnected components within the digitalization framework of industrial systems. AAS, acting as the bridge between the physical world and its digital representation, provides a structured representation of industrial assets, offering comprehensive metadata that serves as a foundation for creating and maintaining DTs.

DTs, in turn, leverage the data and functionality provided by AAS, shown in Fig. 9, to replicate the behaviour and characteristics of physical assets in a virtual environment.

By integrating with AAS, DTs can access real-time data streams, historical performance metrics, and operational parameters of corresponding physical assets. This seamless connection allows DT to continuously synchronise with their physical counterparts, enabling predictive modelling, real-time monitoring, and decision support based on accurate and up-to-date information.

To enable integration between AAS and DT, the choice of data exchange and database technology to adopt is fundamental. Historically, the industrial world has been strongly oriented towards SQL (Structured Query Language) database technology [82, 83], known primarily for its robustness. SQL databases ensure data integrity, have an efficient query system, and also support complex relationships between different data entities, making it the most efficient choice when it comes to managing structured data.

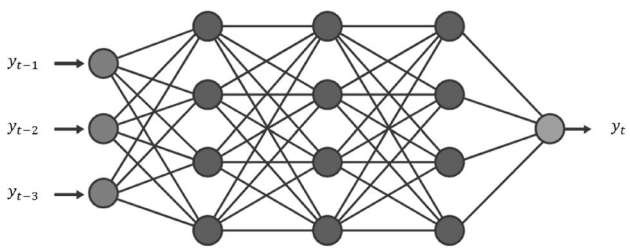

Fig. 8 An Auto Regressive Neural Network architecture

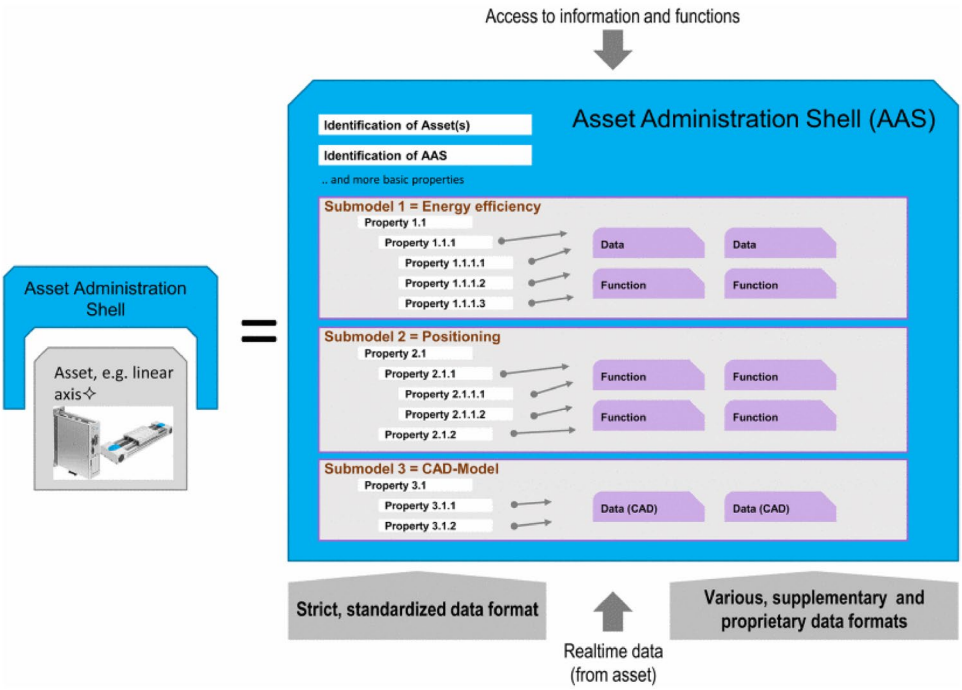

Fig. 9 An example of how an Asset Administration Shell (AAS) in the industrial context could be structured as proposed in the [81]

One of the key aspects that makes SQL databases suitable for integration between AAS and DT is the concept of ACID, an acronym that represents the fundamental properties of transactions in a database and defined in [84].

In contrast to SQL databases, with the increasing volume of data and smart devices in modern industry, there arises a need to not be strongly bound by the rigid relational structures of an SQL database. For this reason, NoSQL (Not Only SQL) data storage models have quickly gained popularity [85].

There is no precise definition for NoSQL models, but studies such as [86, 87] provide clear distinctions from traditional SQL databases. This type of database is particularly recognised for one specific characteristic: the absence of relational models [88].

Defining data management approaches is crucial, yet another significant step in handling industry-derived information is data storage, essential for supporting organisations' competitive strategies. Establishing a clear, secure storage strategy relies on understanding how data will be used and how value will be extracted from it. Components of a storage strategy may encompass data accessibility levels (real-time or periodic), standards for data quality, administrative regulations governing secure data management, privacy and confidentiality policies, and cybersecurity considerations. Depending on the organisation's strategy, data may be stored in an on-premises environment or in a data centre, which could be owned and managed by the organisation itself or a third party (such as cloud computing services). The volume of data, overall space requirements, and projected data growth rate will influence storage mode decisions.

Developing a storage strategy typically aligns with data acquisition, as there is a close link between the data to be acquired and where it will be stored [89].

2.4 Application layer

The Application layer stands as the final components in the presented DT framework, embracing key procedures like data processing, analysis, and visualisation to extract actionable insights from both collected data in the perception layer and estimates values in output to digital object.

Advances analytics methodologies such as statistical analysis and ML may be used in this layer to discover hidden patterns or anomalies, which give insights to the users to implement predictive maintenance, anomaly detection, and operational optimisations strategies efficiently via Graphical User Interface (GUI). Furthermore, in an advanced DT application, the intelligent system can use a Decision Support Systems (DSS), composed by Reinforcement Learning (RL) agent [90] or expert systems [91], to make autonomous decision and can use Large Language Models (LLMs) to communicate with the user at human-level, via vocal or text command, which are easy to interpret then visual data. These self-control and self-monitoring capabilities empower DTs to autonomously make complex decisions, thereby charge

physical assets with intelligence through their digital counterparts. Furthermore, the integration of LLMs enhance the communication with human users, whose can easily interact with the DT, enhancing the bi-directional nature of DT also from a human-centric perspective.

2.4.1 Clustering analysis

Clustering analysis involves grouping samples contained in a dataset based on their similarity. Using ML it is possible for a DT to identify clusters in a high dimensional space, which is a complex task for a human operator, especially if it is done online for high-speed dynamical systems. The similarity between clusters can be evaluated using different metrics, but usually Euclidean distance between features of the samples is used. Clustering analysis can be used in DT for process monitoring, since obtained the clusters, samples coming from the process can be assigned to each cluster and identify system states. The key advantage in this case is the absence of supervised learning techniques, such as logistic regression, to classify the system states. Among the existing different techniques in this field, k-means and hierarchical clustering are the most employed methods for clustering. Upon cluster identification, a DSS can leverage cluster labels to either propose actionable recommendations to the user or autonomously initiate system adjustments to optimise performance.

2.4.2 Dimensionality reduction

Dimensionality reduction is the process of reducing a number N of variables into a set of $M < N$ variables which describe the dataset in a low dimensional space. Techniques like Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and UMAP [92] help in simplifying models, reducing computation time, and mitigating the curse of dimensionality while preserving essential relationships in the data. This key feature associated to DTs are particularly useful to visualise high-dimensional data and speed up the computational process of cognitive modules which allow to take autonomous decision on the physical object. t-SNE is particularly powerful for preserving local similarities and revealing patterns such as clusters after the reduction and is more suitable for reducing dimensionality of non-linear data with respect to PCA. However, it is computationally intensive and primarily used for visualisation rather than for downstream machine learning task, in opposition to PCA. For this reason, t-SNE can be used in slow-dynamic systems and for visualisation purposes rather than feedback control or real-time monitoring applications. Therefore, the dimensionality reduction techniques is particularly useful for data visualisation. Once data have been extracted from the physical asset in the perception layer, reducing the dimensionality allow for data visualisation in 2D or 3D space to the user, which can leverage scatter plot, time series plot and CAD replica to visualise and monitor in the digital space the physical asset and its performance.

2.4.3 Fuzzy logic

Fuzzy logic is a form of many-valued logic derived from fuzzy set theory to handle reasoning that is approximate rather than precise. Unlike binary sets, fuzzy logic variables may have a truth value ranging between 0 and 1. In Fig. 10 is shown an example of 1D—Fuzzy chart. Using expert knowledge, the system can be described in different discrete states associated to the input level, as well the control action of a Fuzzy controller. The fuzzification depends by the design of fuzzy roles or membership, which are typically trapezoidal and triangular. The trapezoidal membership function consists of a flat top (where the membership value is 1) and two sloped sides where the membership values increase or decrease linearly, while the triangular membership function is characterised by a peak (where the membership value is 1) and two linear slopes that decrease to 0 on either side of the peak.

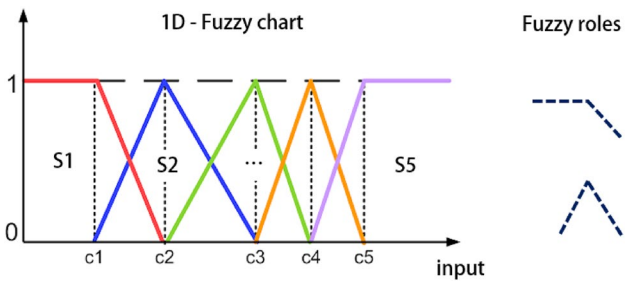

Fig. 10 An example of 1D Fuzzy chart with 1 input and 5 different states

Fuzzy logic can be used to take autonomous action on the physical asset or to suggest different strategy of intervention to the user, which can display this suggestion in the proposed GUI, which guarantee the simplest communication between users and DTs in the application layer.

2.4.4 Reinforcement learning

Reinforcement Learning (RL) is a rapidly growing subfield of ML that focuses on training agents to make sequential decisions in an environment to maximise a cumulative reward signal [93]. RL has emerged as a powerful approach for solving complex decision-making problems in various domains, including robotics, game play, and autonomous systems [94, 95]. In the context of DTs, RL offers a promising framework to enable intelligent decision-making and control in manufacturing systems [96–98].

The core idea behind RL, shown in Fig. 11, is to learn optimal policies through trial-and-error interactions with the environment. An RL agent learns by taking actions in the environment and observing the resulting rewards and state transitions. The goal is to find a policy π, which maps states to actions, that maximises the expected cumulative reward over time.

One of the most popular RL algorithms is Q-learning [99], that can be used when the action space is discrete. It learns an action value function Q(s, a) that estimates the expected cumulative reward for taking action a in state s and following the optimal policy thereafter.

Another prominent RL approach is the policy gradient method [100, 101] used for continuous action space. This algorithm directly learns a parameterised policy $\pi_\theta(a, s)$ that maps states to action probabilities using a function approximator, such as polynomial regression or neural networks.

In the context of manufacturing systems, RL has been applied to various problems, such as production scheduling [102], process control [103], and maintenance optimization [104]. By integrating RL with DT, manufacturers can create intelligent agents that learn optimal policies to control and optimise manufacturing processes in real time [105].

As for Fuzzy Logic, RL agents can leverage a rich data environment encompassing raw sensor data, extracted features, cluster labels, and DT predictions to autonomously make decisions or provide user-oriented recommendations through a DSS. Furthermore, RL agents are emerging as optimal online feedback controllers in the realm of complex, nonlinear, or poorly understood dynamic systems, where precise target tracking is paramount [106].

2.4.5 Large Language Models to enhance bi-directional communication with human user

In recent years, Large Language Models (LLMs) have achieved significant advances in Natural Language Processing (NLP), empowering computers with unprecedented abilities to understand, analyse, and generate human-quality text. These models, boosted by Transformer architectures like GPT and LLAMA, are trained on massive datasets to acquire their remarkable language proficiency [107]. Through accessible application programming interfaces (APIs), LLMs are being seamlessly integrated into a wide range of industries, including chemistry and advanced materials [108], smart buildings [109], and manufacturing [110], revolutionizing operations and decision-making processes.

The Transformer architecture employs self-attention [111] to weigh the importance of different input parts, capturing complex dependencies, determining the importance of each word in relation to others via attention score, leading to a contextual understanding.

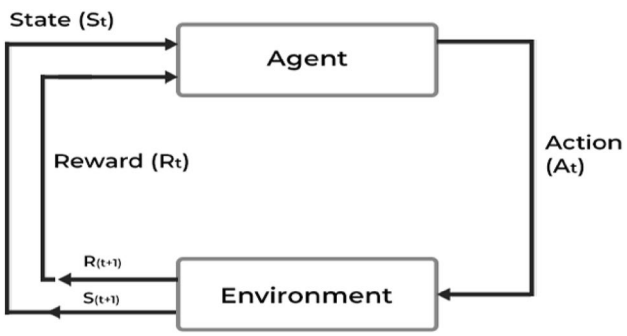

Fig. 11 A general scheme of Reinforcement Learning

LLMs are fundamentally transforming the way humans interact with and derive value from complex systems. Users can now communicate with machines using both textual and vocal commands, requesting specific tasks or insights. In response, LLMs process these requests, interact with underlying ML models and data, and deliver clear, human-understandable feedback, often in natural language format. This bidirectional communication loop, shown in Fig. 12, empowers users to engage in dynamic, iterative exploration and decision-making processes, enhancing the visualisation and decision support capabilities of DTs by providing a conversational interface. Essentially, LLMs act as the bridge between human intent and machine intelligence, enabling a more fluid, efficient, and effective collaboration between users and complex systems, a topic that is at the centre of transition to a more human-centric Industry 5.0 paradigm.

3 Application of machine learning in digital twin technology

As discussed in the introduction the earliest definition of DTs is that it is a virtual replica of a physical object, process or system. With the virtual replica, there are two streams of interpretation where one sees it as a mathematical model, while the other focuses on the visualisation. Nowadays, in addition to physical and data-driven based models, DTs can encompass visual elements through the addition of 2D or 3D models to bring contextual meaning to the data, whether real-time or predicted. However, its scope slight changes for different research fields. For example, in relation to the Industry 4.0 principles, DTs have a broader scope and is neither the mathematical analytical model nor the visualisation. DTs and simulations are also not alike. Unlike simulations that operate offline, DTs have online integration with data streams acquired from real equipment or processes and the simulation is a component of a DT to enable analysis and produce desired results such as predictions or fault detection. While the specific requirements for DTs vary across different scientific domains, the proposed framework, comprising perception, digital object, and application layers, offers a versatile foundation. As detailed in Sect. 2, ML is an indispensable component of this framework and its application across diverse fields. This section provides a comprehensive overview of ML applications in DT development, categorised by scale (unit, system, and system of systems levels). By examining existing literature, we aim to explain the diverse ways ML is employed to construct DTs in various scientific contexts.

3.1 Digital twins for advanced materials

The development of DTs in materials science represents a cutting-edge frontier, integrating advanced digital technologies to model, simulate, and optimise material properties. A DT serves as a virtual counterpart of a physical material, enabling researchers to investigate material behaviour under diverse conditions without the constraints of physical experimentation. This approach offers substantial benefits, including optimised design and accelerated testing.. As a foundational component of any physical system, a material DT occupies the unit level within the DT hierarchy (see Fig. 5). Global initiatives such as ICME, MGI, MGE, HCPS, and Industry 4.0 have significantly advanced both data and design infrastructures, propelling the field of materials science forward [112, 113]. Despite the established prominence of physics-based models like Density Functional Theory (DFT) and Molecular Dynamics (MD) in materials science, there is a growing interest on data-driven ML approaches to generate digital models for materials in the field of DTs. This shift is driven by the potential for ML models to outperform traditional methods, especially when trained on extensive datasets incorporating accurate quantum or experimental data [114]. Such advancements are crucial for the design of novel materials with targeted properties. Kadupitiya et al. [115] proposed a ML surrogate regression model to estimate the output features of MD simulations that allow for a 10,000 times smaller simulation time with high fidelity. This innovation allows to visualise and simulate the system in a fast way, enabling optimisation in design of material with specific material proprieties. Similarly, Shanks et al. [116] proposed the usage of local Gaussian process (LGP) surrogate models to accelerate Bayesian optimisation of materials thermophysical properties.

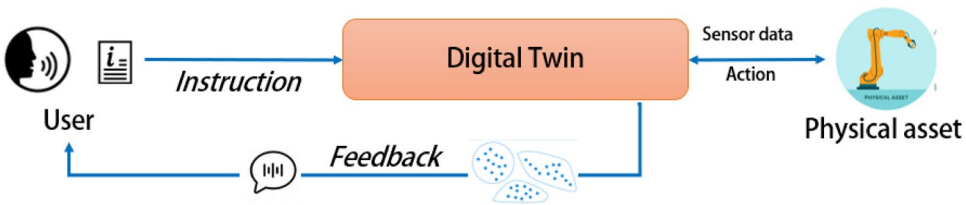

Fig. 12 The bidirectional communication loop to the user enabled by LLM

As proposed by Vozza et al. [117] by leveraging Electronic Structure simulations and synthetic Scanning Tunnelling Microscopy (STM) images (see Fig. 13), the atomic-scale structures can be related to the materials properties employing state-of-art ML techniques such as Yolo and XGBoost classification model.

While experimental data is commonly used, incorporating high-fidelity simulation data from computational models can enhance dataset completeness and reveal hidden relationships. This hybrid approach, referred as grey box, combines experimental observations with simulated data, leveraging techniques like diffraction measurements and generative models (GANs) to create comprehensive datasets for advanced analysis [118–120]. As mentioned, ML is largely employed for materials optimisation, and several studies have demonstrated that Reinforcement Learning overperform similar methodologies in optimisation, especially in cases in which a model for the system is not available [121, 122].

Furthermore, one of the most sophisticated approaches in the construction of DTs for materials is the use of multiscale models [123–125]. Materials exhibit behaviours and properties that vary across different length and time scales, from atomic and molecular levels to macroscopic levels. Multiscale models aim to integrate information and phenomena across these scales, providing a comprehensive and coherent view of the material. At the atomic and molecular scale, technologies such as molecular dynamics (MD) and Monte Carlo (MC) [126] methods are used to study atomic interactions, crystalline structure, atomic defects, and diffusion processes, which are crucial for designing alloys, semiconductors, and polymeric materials. At the microscopic scale, finite element methods (FEM) and discrete grid methods (DEM) are employed to model the microstructure of the material, including grains, phases, and dislocations, thus predicting properties like strength, ductility, and toughness [127, 128].

At the macroscopic scale, finite element analysis (FEA) and continuous methods simulate the overall behaviour of the material under mechanical, thermal, and chemical loads, essential for structural design, life cycle analysis, and performance evaluation in real operational conditions.

This study underscores the emergence of DTs as virtual representations of physical systems in advanced materials. Empowering ML, DTs excel in generating high-fidelity, computationally efficient multiscale simulation models and optimising material composition and structure thanks SOA methods like generative ML and RL. Despite these advancements, challenges persist, including the integration of multiscale and multiphysics models, as well as the scarcity of experimental data. Additionally, the absence of standardized DT definitions delays progress. However, recent trends, such as the application of generative models to synthesize data, offer promising solutions to some of these obstacles.

3.2 Manufacturing processes

DT technology represent one of the Key Enabling Technologies (KETs) of the fourth industrial revolution and is assuming a crucial role in enhancing both manufacturing processes (system level) and manufacturing systems (system of system level) [129]. With growing emphasis on sustainability and the widespread accessibility of ML, the integration of DTs into manufacturing processes has led to a proliferation of innovative applications [130–134]. A DT of a manufacturing process is a system where a digital counterpart developed on the machine software communicates and interacts with physical assets, composed by the hardware of the machine and the physical process, e.g. welding, machining, casting, additive manufacturing or plastic deformation. This digital counterpart possesses supervisory, diagnostic, predictive, and control capabilities which allow to give to the machine self-monitoring and self-control capabilities [135] when advanced statistical and ML methods are employed to process the big data acquired by the perception layer. A DT framework for Additive Manufacturing process based on the proposed general framework is presented in Fig. 14. Data acquired from the perception layer is utilised to forecast system performance using a digital model within the digital object layer. Subsequently, this information is leveraged in the application layer to monitor, control, and optimise the

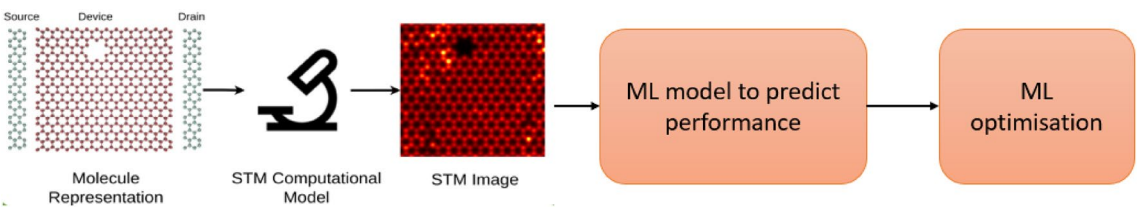

Fig. 13 From molecule computational representation to synthetic image generation of scanning tunnelling microscopy of defected grafene flake [117] which can be used to train ML models useful for ML optimisation of the material properties

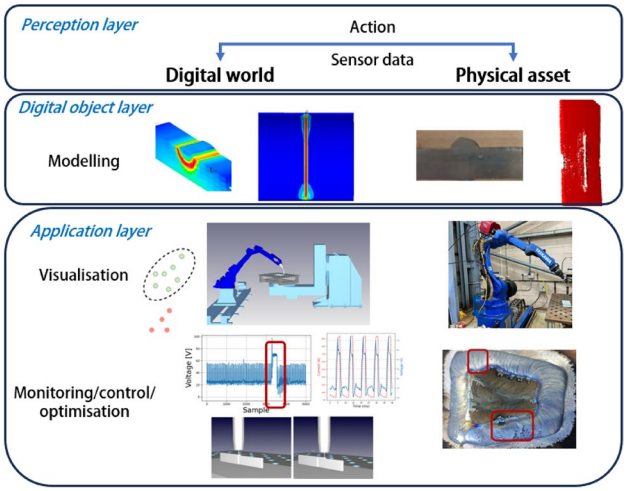

Fig. 14 A Digital Twin contain a digital model which allow to monitor, control and visualise the physical asset behaviour within the proposed framework

manufacturing process, enabling comprehensive visualisation of the physical asset from both kinematic and process perspectives through dedicated applications.

Unlike the previously discussed DTs for advanced materials, a standardised framework for manufacturing processes has been established through ISO 23247 [136–138]. Based on ISO 23247 standard, a DT has to contain both physical and process components that are interconnected and synchronised through device communications. Therefore, building a DT requires a digital representation of the Manufacturing Entities (ME), incorporating both static and dynamic information attributes. Static information remains constant throughout the process, while dynamic information changes over time. Once the data are collected in the perception layer, several applications of ML can be found in the literature.

Regarding sensor fusion applications, Caggiano et al. [139] proposed an Artificial Neural Network (ANN) based method to estimate the tool wear during the drilling process of titanium alloy. This method involves creating a Sensor Fusion Pattern Vector (SFPV) using PCA-derived statistical measures of sensor signals. This SFPV is then used to estimate tool wear with ANN. Integrating this model with a DSS within a DT can potentially generate predictive models for tool changes during drilling operation. Additionally, some authors proposed ML approaches based on image processing to estimate online the width and penetration depth of welding processes [140–142]. In this case the estimation of this state of the system is used to create feedback loop to the system controller. A similar approach was proposed by Chabot et al. [143], in which frequency domain features of sound signal during Wire Arc Additive Manufacturing (WAAM) process is used in a linear regression to estimate the torch distance from the substrate, enabling for an indirect feedback loop to the control system.

As mentioned above, ML regression techniques can be employed in the perception layer to estimate system states online with different purposes. However, similar techniques can be used offline aiming to generate digital models for the digital object layer. Li et al. [144] proposed a long-term dynamic model of the WAAM process via LSTM neural network which allow to forecast the system states in terms of layer geometry based on input parameters and past values. This digital object is then used into this research to develop a Model Predictive Controller (MPC) for the process. However, a digital object model can be used also for anomaly detection purposes as shown by Reisch et al. [145]. In particular, if a digital object model is obtained using only defect-free deposition data, any excessive error between predictions and actual values obtained from the system may suggest the presence of an anomaly. Additionally, Chen et al. [146] used ML regression models to generate predictive models of the thrust force and delamination during a drilling process. This predictive model generated maps that allow to optimise the process parameters of the next drill reducing delamination in composite material drilling. All the presented models are used online and are able to forecast system states based on the past values. However, digital models can be static and can be used for optimisation purposes. For example, Xia et al. [147] developed ML regression models which allow to estimate the surface roughness of an additively manufactured parts based on process parameters employed, allowing to use this model in an optimisation loop, as suggested by Tomaz et al. [148].

Finally, ML is largely employed to develop the monitoring application of DT, as an alternative way to the discussed employment of the digital model. In particular, Caggiano et al. [149, 150] used statistics to obtain a Sensor Fusion Pattern Vector (SFPV) and a Machine Learning Hierarchical clustering to cluster the process state based on multi-sensor data statistics during an Electrical Discharge Machine process. The proposed system allowed to enable the supervisory

and diagnostic capabilities of a DT, which can be used to suggest action to the user or take autonomous decision. As additional example, Mattera et al. [151] proposed ML methods to monitor online the process using image processing and cluster analysis, enabling DT to detect potential anomalies during the deposition process.

This study emphasizes the research efforts of the last years dedicated to integrating ML applications into manufacturing processes to improve system performance. As presented, these studies align well with the proposed DT framework. Furthermore, the existence of international standards in this domain facilitates the realisation of DTs for manufacturing, which are now feasible. However, the shift towards human-centric and sustainable production necessitates new research directions. Applications focusing on sustainability rather than performance optimisation will become increasingly important. This evolving paradigm calls for innovative frameworks and architectures prioritising sustainability.

3.3 Manufacturing systems

As discussed for the manufacturing processes, DTs have become a foundation of Industry 4.0, particularly in optimising manufacturing operations and enabling data-driven decision-making. In the context of manufacturing system, DTs operate at a system-of-systems (SoS) level, where individual manufacturing processes communicate and interact with the others. As illustrated in Fig. 15, a multi-scale approach is applicable here as well. The perception layer comprises a machine layer, collecting data from individual machines, which is then aggregated in a data centre layer serving as the central hub for data communication and distribution to the application layer.

DTs provide a virtual representation of the physical manufacturing system, encompassing various elements such as machines, conveyors, robots, and production lines [152, 153]. By leveraging real-time data from sensors, IoT devices, and Manufacturing Execution Systems (MES), DTs can accurately replicate the behaviour and performance of the actual manufacturing system [154].

As for the single machine. typical applications of ML in DTs in manufacturing systems include real-time monitoring and control, predictive maintenance, production planning and scheduling optimisation, and virtual commissioning. [155]. These applications leverage the data-driven nature of DTs and the statistical/machine learning methods discussed in the previous sections to improve system performance and support decision making.

One notable application of DTs in manufacturing systems is the automated discovery and generation of simulation models. Lugaresi and Matta [156] proposed a method to automatically discover manufacturing system structures and generate Discrete Event Simulation (DES) models from event logs. Their approach utilises process mining techniques to identify the topology of the system, estimate the parameters of the model, and generate a simulation model that accurately represents the manufacturing system. This automated approach enables the rapid development of DTs digital object model that can estimate system performance and support real-time decision making, both automatically employing RL techniques or via DSS [157, 158].

Furthermore, DTs facilitate the implementation of flexible and reconfigurable manufacturing systems through virtual commissioning. Zhang [96, 103] introduced a reconfigurable modelling approach based on DT for smart manufacturing systems. They employed statistical methods, including data analysis, clustering, and similarity matching, to identify

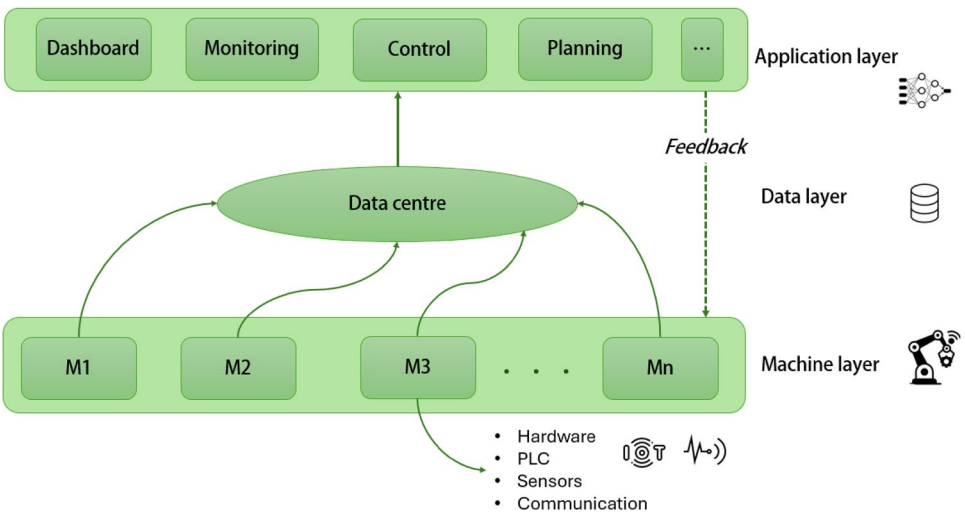

Fig. 15 A general scheme of an Industry 4.0 manufacturing system. Each machine acts as a site for different data sources, utilising Big Data technologies such as databases and ML for the elaboration in the application layer. The system includes a graphical user interface for user interaction and a monitoring module that extracts insightful data from raw data. This data is then used to plan production and control each machine in the network, ensuring optimised and efficient manufacturing processes

reconfigurable modules and generate alternative system configurations. DT enables virtual testing and validation of these configurations, allowing manufacturers to evaluate the feasibility and performance of new production layouts and control strategies. Using statistical methods and ML within the DT framework, manufacturers can optimise the reconfiguration of the system and quickly adapt to changing product requirements and market demands.

Production planning and scheduling optimisation are another area where DTs and statistical methods converge. Zhang et al. [103] developed a hybrid optimisation algorithm that combines Genetic Algorithms (GA) and RL to generate optimal production schedules. The DT serves as a virtual replica of the shop floor, enabling the simulation and evaluation of various scheduling scenarios. The proposed approach utilises statistical methods, such as data pre-processing, feature selection, and model training, to enhance the performance of the optimisation algorithm and improve the quality of the generated schedules.

RL has shown great potential to enhance the decision-making capabilities of DTs in manufacturing systems. Waschneck et al. [102] proposed a deep RL approach to optimise production scheduling in a flexible manufacturing system. They trained a deep Q-network (DQN) agent to make scheduling decisions based on the current state of the system, considering factors such as machine availability, job due dates, and setup times. The DQN agent learns to minimise tardiness and maximise throughput by exploring different scheduling strategies and learning from the resulting rewards. In the area of maintenance optimisation, Siraskar et al. [104] proposed an RL-based approach for predictive maintenance in manufacturing systems. They trained a DQN agent to make maintenance decisions based on the condition monitoring data of the machines. The RL agent learns to schedule maintenance actions to minimise downtime and maintenance costs while ensuring machine reliability. However, the application of RL in manufacturing systems also presents several challenges, such as the need for large amounts of training data and the scalability and interpretability of RL algorithms. Researchers are exploring various approaches to address these challenges, including transfer learning [159] hierarchical RL [160] and explainable RL [161]. These techniques aim to accelerate learning, improve scalability, and enhance the interpretability of RL-based DTs in manufacturing systems.

The integration of DTs and ML has demonstrated significant potential for optimising manufacturing resources through both visualisation and ML optimisation techniques such as RL and GA. This combination empowers data-driven decision-making across various manufacturing operations, from automated simulation model generation to production scheduling and maintenance optimisation. Furthermore, the multi-scale nature of DTs enables the utilisation of machine-level monitoring information to assess performance and detect anomalies within manufacturing networks. However, advancements in ML for network monitoring can be leveraged to identify anomalies both within and across factories, paving the way for system-of-systems level monitoring applications.

3.4 Built Environment and Clean Energy

In the effort to improve energy efficiency in buildings while ensuring occupancy comfort, there is a push for energy flexibility of buildings to manage energy demand and generation that varies with the climate, occupancy comfort and load on the distribution grid, and transition from fossil fuels to renewable energy sources [162]. Energy flexibility can be obtained by testing various control strategies on a building and the operation of their systems. This requires a building model representing the thermal interactions within the building from internal heat gains, occupancy patterns and climate conditions, and equipment such as heat pumps, HVAC systems, solar photovoltaic systems and electric vehicles to estimate energy consumption. In this sense, Digital Twin technology can be utilised in order to create a virtual replica of a physical system allowing for real-time monitoring, analysis, and optimisation.

For what concern the digital object layer, researchers have applied a white-box to black-box and grey-box (hybrid) approaches to achieve energy flexibility [163]. The white-box method being traditional physics-based models accounts for physical processes and can be created using building energy simulation programs such as EnergyPlus, TRNSYS, and Modelica. On the other hand, the black-box method is data-driven and utilises statistics and machine learning algorithms such as regression analysis.

The grey-box or hybrid model, utilises the advantageous characteristics of both the white-box and black-box models, and can be developed in different ways. For example, Lin et al. [164] created a model using Modelica to describe the dynamic behaviour of a HVAC system. However, some elements in the Modelica model used data-driven techniques and advanced statistical methods to approximate key parameters of the building such as physical properties of walls. Another method presented to develop hybridized models is to create physics-based models and test control strategies using Model Predictive Control or machine learning algorithms [165, 166]. Also in this case RL and other ML optimisation

algorithm can be used to optimise resources within the application layer of the proposed DT framework thanks to digital model of the system, hosted in the digital object layer [167, 168].

However, these models only represent the physical behaviour of a building and are one component that make up a DT to produce desired output information such as key performance indexes to assist with decision making. Returning to the definition of digital twins, other investigative works focus on the acquisition of real data from equipment or systems using Internet of Things and the mapping of the data from the physical entity to the virtual replica [169].

Work undertaken by [170] describe a platform called the Data Clearing House. It acts as a unified repository for time-series data and metadata and maps those data points to a semantic model that follows the building-related ontology called Brick Schema [171]. This platform enables the ingestion of data from Building Management Systems (BMS) that oversee the operation and management of building assets, and from Internet of Things devices which are more common in modern buildings. It also overcomes inconsistencies with data point labelling conventions from different BMS service operators through the use of the ontologies. This process enables greater interoperability and scalability by allowing seamless addition of other models such as the analytical simulation models described earlier.

Monitoring applications represent another common use case for ML in this domain. Talei et al. [172] exemplify this by demonstrating how time series analysis and ML can identify inefficiencies in HVAC consumption management. Their method pinpoints time slots for reduced HVAC usage, given that these systems are significant electricity consumers in buildings. The insights from this monitoring module can be integrated into DTs with advanced control strategies or DSS to enhance building performance and sustainability. As previously discussed, both grey system modelling and monitoring applications, combined with ML optimisation techniques, contribute to advancements in the field of DTs.

4 Conclusions

This study presented and discussed a generic framework of digital twins and its applications across various research domains. Through a review analysis, it was revealed that statistics and machine learning methods play a pivotal role in the development and integration of digital twin technology in multiple disciplines. Key characteristics such as multidisciplinary and multi-scale aspects of digital twins were emphasised, along with the applications of data-driven techniques enabled by machine learning for modelling, visualisation, monitoring and optimisation. Both traditional and innovative techniques were examined, with references to relevant ISO standards in the fields where present. Additionally, the introduced advantages of digital twins, the state-of-the-art of various applications, and the ongoing challenges in the analysed research fields were presented, including advanced materials, smart buildings, manufacturing processes, and manufacturing systems. This work provides a comprehensive overview of the potentials and challenges of digital twins, laying a solid foundation for future research and advancements in this continually evolving field, in which statistics, probability and machine learning plays a crucial role.

Acknowledgements The INVITALIA Project NEMESI is gratefully acknowledged for their support to this research work.

Author Contributions Conceptualization, GM, EY, SV and MV; methodology, GM, EY, SV, LN and MV; investigation, GM, EY, SV, LN, and MV; writing-original draft preparation, GM, EY, SV and MV; writing-review and editing, GM, EY, SV, LN and MV; All authors have read and agreed to the published version of the manuscript.

Funding This research received no external funding.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Tao F, Qi Q, Wang L, Nee AYC. Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: correlation and comparison. Engineering. 2019;5(4):653–61.
2. Wu J, Yang Y, Cheng XUN, Zuo H, Cheng Z. The development of digital twin technology review. Chin Autom Congress. 2020;2020:4901–6.
3. Josifovska K, Yigitbas E, Engels G, Reference framework for digital twins within cyber-physical systems, in *2019 IEEE/ACM 5th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)*, 2019, pp. 25–31.
4. Monostori L, et al. Cyber-physical systems in manufacturing. CIRP Ann. 2016;65(2):621–41. https://doi.org/10.1016/j.cirp.2016.06.005.
5. Javaid M, Haleem A, Suman R. Digital twin applications toward industry 4.0: A review. Cogn Robot. 2023;3:71–92.
6. Yao J-F, Yang Y, Wang X-C, Zhang X-P. Systematic review of digital twin technology and applications. Vis Comput Ind Biomed Art. 2023;6(1):10.
7. Nwogu C, Lugaresi G, Anagnostou A, Matta A, Taylor SJE. Towards a requirement-driven digital twin architecture. Procedia CIRP. 2022;107:758–63.
8. Protic A, Jin Z, Marian R, Abd K, Campbell D, Chahl J. Implementation of a bi-directional digital twin for industry 4 labs in academia: a solution based on OPC UA. IEEE Int Conf Ind Eng Eng Manag. 2020;2020:979–83.
9. G. Mattera, J. Polden, A. Caggiano, L. Nele, Z. Pan, and J. Norrish, "Semi-supervised Learning for Real-Time Anomaly Detection in Pulsed Transfer Wire Arc Additive Manufacturing," *J Manuf Process*, 2024.
10. Lee J, Bagheri B, Kao H-A. A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. Manuf Lett. 2015;3:18–23. https://doi.org/10.1016/j.mfglet.2014.12.001.
11. Zhu Z, Liu C, Xu X. Visualisation of the digital twin data in manufacturing by using augmented reality. Procedia CIRP. 2019;81:898–903.
12. Liu M, Fang S, Dong H, Xu C. Review of digital twin about concepts, technologies, and industrial applications. J Manuf Syst. 2021;58:346–61.
13. Cimino C, Negri E, Fumagalli L. Review of digital twin applications in manufacturing. Comput Ind. 2019;113:103130. https://doi.org/10.1016/j.compind.2019.103130.
14. Qian C, Liu X, Ripley C, Qian M, Liang F, Yu W. Digital twin—cyber replica of physical things: architecture, applications and future research directions. Fut Internet. 2022;14(2):64. https://doi.org/10.3390/fi14020064.
15. Khan M, Wu X, Xu X, Dou W, Big data challenges and opportunities in the hype of Industry 4.0, in *2017 IEEE International Conference on Communications (ICC)*, IEEE, May 2017, pp. 1–6. https://doi.org/10.1109/ICC.2017.7996801.
16. Guo X, et al. A digital twin modeling method for array antenna assembly performance real-time analysis. Int J Adv Manuf Technol. 2023;126(7–8):3765–81. https://doi.org/10.1007/s00170-023-11324-1.
17. Eghbalian M, Ansari R, Haghighi S. A combined molecular dynamics-finite element multiscale modeling to analyze the mechanical properties of randomly dispersed, chemisorbed carbon nanotubes/polymer nanocomposites. Mech Adv Mater Struct. 2023;30(24):5159–75. https://doi.org/10.1080/15376494.2022.2114038.
18. Doodman Tipi AR, Pariz N, and others, Improving the dynamic metal transfer model of gas metal arc welding (GMAW) process, in *The International Journal of Advanced Manufacturing Technology*, vol. 76, no. 1, pp. 657–668, 2015.
19. Xiao J, Liu N, Lua J, Saathoff C, Seneviratne WP, Data-Driven and Reduced-Order Modeling of Composite Drilling, in *AIAA Scitech 2020 Forum*, Reston, Virginia: American Institute of Aeronautics and Astronautics, 2020. https://doi.org/10.2514/6.2020-1859.
20. Deng T, Zhang K, Shen Z-JM. A systematic review of a digital twin city: a new pattern of urban governance toward smart cities. J Manag Sci Eng. 2021;6(2):125–34.
21. Farsi M, Daneshkhah A, Hosseinian-Far A, Jahankhani H, and others, *Digital twin technologies and smart cities*, vol. 1134. Springer, 2020.
22. Tuhaise VV, Tah JHM, Abanda FH. Technologies for digital twin applications in construction. Autom Constr. 2023;152:104931.
23. Opoku D-GJ, Perera S, Osei-Kyei R, Rashidi M. Digital twin application in the construction industry: a literature review. J Build Eng. 2021;40:102726.
24. Yang B, Lv Z, Wang F. Digital twins for intelligent green buildings. Buildings. 2022;12(6):856.
25. Vallée A. Digital twin for healthcare systems. Front Digit Health. 2023;5:1253050.
26. Rabah S, et al. Towards improving the future of manufacturing through digital twin and augmented reality technologies. Procedia Manuf. 2018;17:460–7.
27. Yin Y, Zheng P, Li C, Wang L. A state-of-the-art survey on Augmented Reality-assisted Digital Twin for futuristic human-centric industry transformation. Robot Comput Integr Manuf. 2023;81:102515.
28. Glaessgen E, and Stargel D, The digital twin paradigm for future NASA and US Air Force vehicles, in *53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA*, 2012, p. 1818.
29. Converso G, Gallo M, Murino T, Vespoli S. Predicting failure probability in Industry 4.0 production systems: a workload-based prognostic model for maintenance planning. Appl Sci. 2023;13(3):1938.
30. Caggiano A, Mattera G, Nele L. Smart tool wear monitoring of CFRP/CFRP stack drilling using autoencoders and memory-based neural networks. Appl Sci. 2023;13(5):3307.
31. Nele L, Mattera G, Vozza M. Deep neural networks for defects detection in gas metal arc welding. Appl Sci. 2022;12(7):3615.
32. Tao F, Xiao B, Qi Q, Cheng J, Ji P. Digital twin modeling. J Manuf Syst. 2022;64:372–89.
33. Caggiano A, Perez R, Segreto T, Teti R, Xirouchakis P. Advanced sensor signal feature extraction and pattern recognition for wire EDM process monitoring. Procedia CIRP. 2016;42:34–9. https://doi.org/10.1016/j.procir.2016.02.181.
34. Zhao S, Qiu X, Burnett I, Rigby M, Lele A. A lumped-parameter model for sound generation in gas metal arc welding. Mech Syst Signal Process. 2021;147:107085.
35. Xia C, et al. Model predictive control of layer width in wire arc additive manufacturing. J Manuf Process. 2020;58:179–86.
36. He R, Chen G, Dong C, Sun S, Shen X. Data-driven digital twin technology for optimized control in process systems. ISA Trans. 2019;95:221–34.
37. Bikas H, Stavropoulos P, Chryssolouris G. Additive manufacturing methods and modelling approaches: a critical review. Int J Adv Manuf Technol. 2016;83:389–405.

38. Jung TJ, Jeong YH, Shin Y. Simulation of directional drilling by dynamic finite element method. J Mech Sci Technol. 2022;36(7):3239–50.
39. Li JY, Yao XX, Zhang Z. Physical model based on data-driven analysis of chemical composition effects of friction stir welding. J Mater Eng Perform. 2020;29:6591–604.
40. Dhar R, Krishna A, and Muhammed B, Physics and data driven model for prediction of residual stresses in machining, arXiv preprint arXiv:2403.18441, 2024.
41. Li H, Shi X, Wu B, Corradi DR, Pan Z, Li H. Wire arc additive manufacturing: a review on digital twinning and visualization process. J Manuf Process. 2024;116:293–305. https://doi.org/10.1016/j.jmapro.2024.03.001.
42. G. MATTERA, J. POLDEN, and L. NELE, "Monitoring Wire Arc Additive Manufacturing process of Inconel 718 thin-walled structure using wavelet decomposition and clustering analysis of welding signal," *Journal of Advanced Manufacturing Science and Technology*, vol. 0, no. 0, pp. 2025006–0, 2024, https://doi.org/10.51393/j.jamst.2025006.
43. Farhadi A, Lee SKH, Hinchy EP, O'Dowd NP, McCarthy CT. The development of a digital twin framework for an industrial robotic drilling process. Sensors. 2022;22(19):7232.
44. C. Gao, H. Park, and A. Easwaran, "An anomaly detection framework for digital twin driven cyber-physical systems," in *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, 2021, pp. 44–54.
45. Zhang H, Qi Q, Tao F. A multi-scale modeling method for digital twin shop-floor. J Manuf Syst. 2022;62:417–28.
46. Sasiadek JZ. Sensor fusion. Annu Rev Control. 2002;26(2):203–28.
47. S. Yazdkhasti and J. Z. Sasiadek, "Multi sensor fusion based on adaptive Kalman filtering," in *Advances in Aerospace Guidance, Navigation and Control: Selected Papers of the Fourth CEAS Specialist Conference on Guidance, Navigation and Control Held in Warsaw, Poland, April 2017*, 2018, pp. 317–333.
48. Giacalone M, Panarello D, Mattera R. Multicollinearity in regression: an efficiency comparison between Lp-norm and least squares estimators. Qual Quant. 2018;52(4):1831–59. https://doi.org/10.1007/s11135-017-0571-y.
49. Subrahmanya N, Shin YC, Meckl PH. A Bayesian machine learning method for sensor selection and fusion with application to on-board fault diagnostics. Mech Syst Signal Process. 2010;24(1):182–92.
50. Ding D, He F, Yuan L, Pan Z, Wang L, Ros M. The first step towards intelligent wire arc additive manufacturing: an automatic bead modelling system using machine learning through industrial information integration. J Ind Inf Integr. 2021;23:100218. https://doi.org/10.1016/j.jii.2021.100218.
51. Mattera G, Caggiano A, Nele L. Reinforcement learning as data-driven optimization technique for GMAW process. Weld World. 2023. https://doi.org/10.1007/s40194-023-01641-0.
52. Xiong J, Zhang G, Hu J, Wu L. Bead geometry prediction for robotic GMAW-based rapid manufacturing through a neural network and a second-order regression analysis. J Intell Manuf. 2014;25(1):157–63.
53. Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. Technometrics. 1999;41(3):212–23.
54. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst. 2017;42(3):1–21.
55. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, 2008, pp. 413–422.
56. G. Mattera, J. Polden, A. Caggiano, P. Commins, L. Nele, and Z. Pan, "Anomaly Detection of Wire Arc Additively Manufactured Parts via Surface Tension Transfer through Unsupervised Machine Learning Techniques," in *17th CIRP Conference on Intelligent Computation in Manufacturing Engineering*, Naples: Procedia CIRP, 2023.
57. Sharma V. A study on data scaling methods for machine learning. Int J Glob Acad Sci Res. 2022;1(1):31–42.
58. Mu H, He F, Yuan L, Hatamian H, Commins P, Pan Z. Online distortion simulation using generative machine learning models: a step toward digital twin of metallic additive manufacturing. J Ind Inf Integr. 2024;38:100563.
59. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser. 2019;1168:022022. https://doi.org/10.1088/1742-6596/1168/2/022022.
60. Haji SH, Abdulazeez AM. Comparison of optimization techniques based on gradient descent algorithm: a review. PalArch's J Archaeol Egypt Egyptol. 2021;18(4):2715–43.
61. L. Datta, "A survey on activation functions and their relation with xavier and he normal initialization," *arXiv preprint* arXiv:2004.06632, *2020.*
62. M. Cilimkovic, "Neural networks and back propagation algorithm," *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, vol. 15, no. 1, 2015.
63. Vahabli E, Rahmati S. Application of an RBF neural network for FDM parts' surface roughness prediction for enhancing surface quality. Int J Precis Eng Manuf. 2016;17:1589–603.
64. X. Jiang, L. Zhou, and P. Li, Maximum thinning rate prediction of friction heat single point incremental forming for AZ31B magnesium alloy based on BP neural network, *J Adv Manuf Syst* 2023.
65. C. L. Poornima, C. S. Rao, and D. N. Varma, Predicting weld quality in duplex stainless steel butt joints during laser beam welding: a hybrid DNN-HEVA approach, *J Adv Manuf Syst* 2024.
66. I. Sülo, S. R. Keskin, G. Dogan, and T. Brown, Energy efficient smart buildings: LSTM neural networks for time series prediction, in *2019 International conference on deep learning and machine learning in emerging applications (Deep-ML)*, 2019, pp. 18–22.
67. Caggiano A, Zhang J, Alfieri V, Caiazzo F, Gao R, Teti R. Machine learning-based image processing for on-line defect recognition in additive manufacturing. CIRP Ann. 2019;68(1):451–4.
68. A. D'Alterio, G. Mattera, and A. Caggiano, "Development of a vision system enhanced by deep learning to support robotic laser cleaning," in *Procedia CIRP, 18th CIRP Conference on Intelligent Computation in Manufacturing Engineering*, Proceedia CIRP, Ed., 2024.
69. Tafarroj MM, Moghaddam MA, Dalir H, Kolahan F. Using hybrid artificial neural network and particle swarm optimization algorithm for modeling and optimization of welding process. J Adv Manuf Syst. 2021;20(04):783–99.
70. N. Yousef and A. Sata, "Intelligent Inspection for Evaluating Severity of Surface Defects in Investment Casting," *Journal of Advanced Manufacturing Systems*, pp. 1–11, 2023.
71. V. Nigam, "Natural Language Processing: From Basics, to using RNN and LSTM," *Towards Data Science*, 2019.

72. M. Said Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "Network Anomaly Detection Using LSTM Based Autoencoder," in *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, New York, NY, USA: ACM, Nov. 2020, pp. 37–45. https://doi.org/10.1145/3416013.3426457.
73. J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Gated Feedback Recurrent Neural Networks," *CoRR*, vol. abs/1502.02367, 2015, [Online]. Available: http://arxiv.org/abs/1502.02367
74. T. Agrawal, "Optuna and AutoML," in *Hyperparameter Optimization in Machine Learning*, Berkeley, CA: Apress, 2021, pp. 109–129. https://doi.org/10.1007/978-1-4842-6579-6_5.
75. Raptis TP, Passarella A, Conti M. Data management in industry 4.0: State of the art and open challenges. IEEE Access. 2019;7:97052–93.
76. Daki H, El Hannani A, Aqqal A, Haidine A, Dahbi A. Big Data management in smart grid: concepts, requirements and implementation. J Big Data. 2017;4:1–19.
77. Diène B, Rodrigues JJPC, Diallo O, Ndoye ELHM, Korotaev VV. Data management techniques for Internet of Things. Mech Syst Signal Process. 2020;138:106564.
78. T. Forni, M. Vozza, F. Le Piane, A. Lorenzoni, M. Baldoni, and F. Mercuri, "AI and data-driven infrastructures for workflow automation and integration in advanced research and industrial applications," in *Ital-IA Thematic Workshops*, 2023. [Online]. Available: http://ceur-ws.org
79. P. Adolphs *et al.*, "Struktur der verwaltungsschale: Fortentwicklung des referenzmodells für die Industrie 4.0-komponente," *Bundesministerium für Wirtschaft und Energie (BMW), Berlin*, pp. 345–361, 2016.
80. Z. Bradac, P. Marcon, F. Zezulka, J. Arm, and T. Benesl, "Digital twin and AAS in the industry 4.0 framework," in *IOP Conference Series: Materials Science and Engineering*, 2019, p. 12001.
81. C. Wagner *et al.*, "The role of the Industry 4.0 asset administration shell and the digital twin during the life cycle of a plant," in *2017 22nd IEEE international conference on emerging technologies and factory automation (ETFA)*, 2017, pp. 1–8.
82. Date CJ. A Guide to the SQL Standard. Inc: Addison-Wesley Longman Publishing Co.; 1989.
83. V. F. de Oliveira, M. A. de O. Pessoa, F. Junqueira, and P. E. Miyagi, "SQL and NoSQL Databases in the Context of Industry 4.0," *Machines*, vol. 10, no. 1, p. 20, 2021.
84. R. Elmasri and S. B. Navathe, "Fundamentals of Database Systems 7th ed.," 2016, *Pearson*.
85. Shareef T, Sharif K, Rashid B. A survey of comparison different cloud database performance: SQL and NoSQL. Passer J Bas Appl Sci. 2022;4(1):45–57. https://doi.org/10.24271/psr.2022.301247.1104.
86. P. J. Sadalage and M. Fowler, *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education, 2013.
87. M. T. Özsu, P. Valduriez, and others, *Principles of distributed database systems*, vol. 2. Springer, 1999.
88. A. B. M. Moniruzzaman and S. A. Hossain, "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison," *arXiv preprint* arXiv:1307.0191, *2013*.
89. Qi Q, et al. Enabling technologies and tools for digital twin. J Manuf Syst. 2021;58:3–21.
90. Mattera G, Mattera R. Shrinkage estimation with reinforcement learning of large variance matrices for portfolio selection. Intell Syst Appl. 2023;17:200181.
91. Tan CF, Wahidin LS, Khalil SN, Tamaldin N, Hu J, Rauterberg GWM. The application of expert system: a review of research and applications. ARPN J Eng Appl Sci. 2016;11(4):2448–53.
92. M. Vozza *et al.*, "Advanced clustering technique for automatic labelling of welding signals," in *Procedia CIRP, 18th CIRP Conference on Intelligent Computation in Manufacturing Engineering*, 2024.
93. R. S. Sutton and A. Barto, *Reinforcement learning*, Second edition. in Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2018.
94. Y. Li, "Deep Reinforcement Learning: An Overview," 2017, *arXiv*. https://doi.org/10.48550/ARXIV.1701.07274.
95. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: a brief survey. IEEE Signal Process Mag. 2017;34(6):26–38. https://doi.org/10.1109/msp.2017.2743240.
96. Zhang C, Xu W, Liu J, Liu Z, Zhou Z, Pham DT. Digital twin-enabled reconfigurable modeling for smart manufacturing systems. Int J Comput Integr Manuf. 2019;34(7–8):709–33. https://doi.org/10.1080/0951192x.2019.1699256.
97. Nian R, Liu J, Huang B. A review On reinforcement learning: Introduction and applications in industrial process control. Comput Chem Eng. 2020;139:106886. https://doi.org/10.1016/j.compchemeng.2020.106886.
98. S. P. K. Spielberg, R. B. Gopaluni, and P. D. Loewen, "Deep reinforcement learning approaches for process control," in *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*, IEEE, May 2017, pp. 201–206. https://doi.org/10.1109/ADCONIP.2017.7983780.
99. Watkins CJCH, Dayan P. Q-learning. Mach Learn. 1992;8(3–4):279–92. https://doi.org/10.1007/bf00992698.
100. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint* arXiv:1707.06347, *2017*.
101. R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in *Advances in Neural Information Processing Systems 12* , 1999.
102. B. Waschneck *et al.*, "Deep reinforcement learning for semiconductor production scheduling," in *2018 29th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, IEEE, 2018, pp. 301–306. https://doi.org/10.1109/asmc.2018.8373191.
103. Zhang H, Zhang G, Yan Q. Digital twin-driven cyber-physical production system towards smart shop-floor. J Ambient Intell Humaniz Comput. 2018;10(11):4439–53. https://doi.org/10.1007/s12652-018-1125-4.
104. Siraskar R, Kumar S, Patil S, Bongale A, Kotecha K. Reinforcement learning for predictive maintenance: a systematic technical review. Artif Intell Rev. 2023;56(11):12885–947. https://doi.org/10.1007/s10462-023-10468-6.
105. G. Mattera, A. Caggiano, and L. Nele, "Optimal data-driven control of manufacturing processes using reinforcement learning: an application to wire arc additive manufacturing," *J Intell Manuf*, pp. 1–20, 2024.
106. N. Nievas, A. Pagès-Bernaus, F. Bonada, L. Echeverria, and X. Domingo, "Reinforcement Learning for Autonomous Process Control in Industry 4.0: Advantages and Challenges," *Applied Artificial Intelligence*, vol. 38, no. 1, Dec. 2024, https://doi.org/10.1080/08839514.2024.2383101.

107. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570–8. https://doi.org/10.1038/s41586-023-06792-0.
108. C. Liao, Y. Yu, Y. Mei, and Y. Wei, "From Words to Molecules: A Survey of Large Language Models in Chemistry," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.01439
109. Waterworth D, Sethuvenkatraman S, Sheng QZ. Advancing smart building readiness: automated metadata extraction using neural language processing methods. Adv Appl Energy. 2021;3:100041. https://doi.org/10.1016/j.adapen.2021.100041.
110. Javaid M, Haleem A, Singh RP. A study on ChatGPT for Industry 4.0: background, potentials, challenges, and eventualities. J Econ Technol. 2023;1:127–43. https://doi.org/10.1016/j.ject.2023.08.001.
111. A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
112. Zhong RY, Xu X, Klotz E, Newman ST. Intelligent manufacturing in the context of industry 4.0: a review. Engineering. 2017;3(5):616–30. https://doi.org/10.1016/J.ENG.2017.05.015.
113. W. Yi Wang, J. Li, W. Liu, and Z.-K. Liu, Integrated computational materials engineering for advanced materials: a brief review, *Comput Mater Sci*, vol. 158, pp. 42–48, 2019, https://doi.org/10.1016/j.commatsci.2018.11.001.
114. Faber FA, et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. J Chem Theory Comput. 2017;13(11):5255–64. https://doi.org/10.1021/acs.jctc.7b00577.
115. Kadupitiya JCS, Sun F, Fox G, Jadhao V. Machine learning surrogates for molecular dynamics simulations of soft materials. J Comput Sci. 2020;42:101107.
116. Shanks BL, Sullivan HW, Shazed AR, Hoepfner MP. Accelerated bayesian inference for molecular simulations using local gaussian process surrogate models. J Chem Theory Comput. 2024;20(9):3798–808.
117. M. Vozza, T. Forni, F. Le Piane, and F. Mercuri, "Efficient Workflow Automation for Materials Modelling: Towards Predictive AI Systems Using High Throughput Synthetic Dataset Generation," 2024.
118. Banko L, Lysogorskiy Y, Grochla D, Naujoks D, Drautz R, Ludwig A. Predicting structure zone diagrams for thin film synthesis by generative machine learning. Commun Mater. 2020;1(1):15. https://doi.org/10.1038/s43246-020-0017-2.
119. Kunka C, Shanker A, Chen EY, Kalidindi SR, Dingreville R. Decoding defect statistics from diffractograms via machine learning. NPJ Comput Mater. 2021;7(1):67. https://doi.org/10.1038/s41524-021-00539-z.
120. Coleman SP, Sichani MM, Spearot DE. A computational algorithm to produce virtual x-ray and electron diffraction patterns from atomistic simulations. JOM. 2014;66(3):408–16. https://doi.org/10.1007/s11837-013-0829-3.
121. Zhou Z, Kearnes S, Li L, Zare RN, Riley P. Optimization of molecules via deep reinforcement learning. Sci Rep. 2019;9(1):10752. https://doi.org/10.1038/s41598-019-47148-x.
122. Sui F, Guo R, Zhang Z, Gu GX, Lin L. Deep reinforcement learning for digital materials design. ACS Mater Lett. 2021;3(10):1433–9. https://doi.org/10.1021/acsmaterialslett.1c00390.
123. Matouš K, Geers MGD, Kouznetsova VG, Gillman A. A review of predictive nonlinear theories for multiscale modeling of heterogeneous materials. J Comput Phys. 2017;330:192–220.
124. M. F. Horstemeyer, "Multiscale modeling: a review," *Practical aspects of computational chemistry: methods, concepts and applications*, pp. 87–135, 2010.
125. Schmauder S, Schäfer I. Multiscale materials modeling. Mater Today. 2016;19:130–1.
126. Gerold V, Kern J. The determination of atomic interaction energies in solid solutions from short range order coefficients—an inverse monte-carlo method. Acta Metall. 1987;35(2):393–9.
127. Gu X, Huang M, Qian J. DEM investigation on the evolution of microstructure in granular soils under shearing. Granul Matter. 2014;16(1):91–106.
128. Chawla N, et al. Microstructure-based simulation of thermomechanical behavior of composite materials by object-oriented finite element analysis. Mater Charact. 2002;49(5):395–407.
129. Leng J, Wang D, Shen W, Li X, Liu Q, Chen X. Digital twins-based smart manufacturing system design in Industry 4.0: a review. J Manuf Syst. 2021;60:119–37. https://doi.org/10.1016/j.jmsy.2021.05.011.
130. Ghobakhloo M. Industry 4.0, digitization, and opportunities for sustainability. J Clean Prod. 2020;252:119869. https://doi.org/10.1016/j.jclepro.2019.119869.
131. Liu S, Bao J, Zheng P. A review of digital twin-driven machining: From digitization to intellectualization. J Manuf Syst. 2023;67:361–78.
132. Chen J, et al. Digital twin-driven real-time suppression of delamination damage in CFRP drilling. J Intell Manuf. 2024. https://doi.org/10.1007/s10845-023-02315-w.
133. He Y, Guo J, Zheng X. From surveillance to digital twin: challenges and recent advances of signal processing for industrial Internet of Things. IEEE Signal Process Mag. 2018;35(5):120–9. https://doi.org/10.1109/MSP.2018.2842228.
134. Mattera G, Nele L, Paolella D. Monitoring and control the Wire Arc Additive Manufacturing process using artificial intelligence techniques: a review. J Intell Manuf. 2023. https://doi.org/10.1007/s10845-023-02085-5.
135. Mu H, He F, Yuan L, Commins P, Wang H, Pan Z. Toward a smart wire arc additive manufacturing system: a review on current developments and a framework of digital twin. J Manuf Syst. 2023;67:174–89.
136. Kim DB, Shao G, Jo G. A digital twin implementation architecture for wire+ arc additive manufacturing based on ISO 23247. Manuf Lett. 2022;34:1–5.
137. Cabral JVA, Gasca EAR, Alvares AJ. Digital twin implementation for machining center based on ISO 23247 standard. IEEE Lat Am Trans. 2023;21(5):628–35.
138. G. Shao and others, "Use case scenarios for digital twin implementation based on ISO 23247," *National institute of standards: Gaithersburg, MD, USA*, 2021.
139. Caggiano A. Tool wear prediction in Ti-6Al-4V machining through multiple sensor monitoring and PCA features pattern recognition. Sensors. 2018;18(3):823.
140. Yu R, Cao Y, Chen H, Ye Q, Zhang Y. Deep learning based real-time and in-situ monitoring of weld penetration: Where we are and what are needed revolutionary solutions? J Manuf Process. 2023;93:15–46. https://doi.org/10.1016/j.jmapro.2023.03.011.

141. Cheng Y, Yu R, Zhou Q, Chen H, Yuan W, Zhang Y. Real-time sensing of gas metal arc welding process–a literature review and analysis. J Manuf Process. 2021;70:452–69. https://doi.org/10.1016/j.jmapro.2021.08.058.
142. Kershaw J, Yu R, Zhang Y, Wang P. Hybrid machine learning-enabled adaptive welding speed control. J Manuf Process. 2021;71:374–83. https://doi.org/10.1016/j.jmapro.2021.09.023.
143. Chabot A, Rauch M, Hascoët J-Y. Novel control model of Contact-Tip-to-Work Distance (CTWD) for sound monitoring of arc-based DED processes based on spectral analysis. Int J Adv Manuf Technol. 2021;116(11–12):3463–72. https://doi.org/10.1007/s00170-021-07621-2.
144. Z. Li, Z. Hou, Z. Pan, D. Wu, and J. Xu, "A Non-autoregressive Dynamic Model based Welding Parameter Planning Method for Varying Geometry Beads in WAAM," *IEEE Transactions on Industrial Electronics*, 2022.
145. R. Reisch, T. Hauser, B. Lutz, M. Pantano, T. Kamps, and A. Knoll, "Distance-based multivariate anomaly detection in wire arc additive manufacturing," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020, pp. 659–664.
146. J. Chen *et al.*, "Digital twin-driven real-time suppression of delamination damage in CFRP drilling," *J Intell Manuf*, pp. 1–18, 2024.
147. Xia C, Pan Z, Polden J, Li H, Xu Y, Chen S. Modelling and prediction of surface roughness in wire arc additive manufacturing using machine learning. J Intell Manuf. 2022;33(5):1467–82. https://doi.org/10.1007/s10845-020-01725-4.
148. I. do V. Tomaz, F. H. G. Colaço, S. Sarfraz, D. Yu. Pimenov, M. K. Gupta, and G. Pintaude, "Investigations on quality characteristics in gas tungsten arc welding process using artificial neural network integrated with genetic algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 113, no. 11–12, pp. 3569–3583, Apr. 2021, https://doi.org/10.1007/s00170-021-06846-5.
149. Caggiano A, Napolitano F, Teti R, Bonini S, Maradia U. Advanced die sinking EDM process monitoring based on anomaly detection for online identification of improper process conditions. Proc CIRP. 2020;88:381–6.
150. Caggiano A, Napolitano F, Teti R. Hierarchical cluster analysis for pattern recognition of process conditions in die sinking EDM process monitoring. Proc CIRP. 2021;99:514–9.
151. G. Mattera, J. Polden, and L. Nele, "A Time-Frequency domain features extraction approach enhanced by computer vision for Wire Arc Additive Manufacturing monitoring using Fourier and Wavelet transform," *J Adv Manuf Syst*. 2024.
152. Grassi A, Guizzi G, Santillo LC, Vespoli S. Assessing the performances of a novel decentralised scheduling approach in Industry 4.0 and cloud manufacturing contexts. Int J Prod Res. 2021;59(20):6034–53. https://doi.org/10.1080/00207543.2020.1799105.
153. Grassi A, Guizzi G, Santillo LC, Vespoli S. A semi-heterarchical production control architecture for industry 4.0-based manufacturing systems. Manuf Lett. 2020;24:43–6. https://doi.org/10.1016/j.mfglet.2020.03.007.
154. Lu Y, Liu C, Wang KI-K, Huang H, Xu X. Digital Twin-driven smart manufacturing: connotation, reference model, applications and research issues. Robot Comput Integr Manuf. 2020;61:101837. https://doi.org/10.1016/j.rcim.2019.101837.
155. Guizzi G, Revetria R, Vanacore G, Vespoli S. On the open job-shop scheduling problem: a decentralized multi-agent approach for the manufacturing system performance optimization. Proc CIRP. 2019;79:192–7. https://doi.org/10.1016/j.procir.2019.02.045.
156. Lugaresi G, Matta A. Automated manufacturing system discovery and digital twin generation. J Manuf Syst. 2021;59:51–66. https://doi.org/10.1016/j.jmsy.2021.01.005.
157. Marchesano MG, Guizzi G, Santillo LC, Vespoli S. A deep reinforcement learning approach for the throughput control of a flow-shop production system. IFAC-PapersOnLine. 2021;54(1):61–6. https://doi.org/10.1016/j.ifacol.2021.08.006.
158. M. G. Marchesano, G. Guizzi, L. C. Santillo, and S. Vespoli, "Dynamic Scheduling in a Flow Shop Using Deep Reinforcement Learning," 2021, pp. 152–160. https://doi.org/10.1007/978-3-030-85874-2_16.
159. Chang J, Yu D, Hu Y, He W, Yu H. Deep reinforcement learning for dynamic flexible job shop scheduling with random job arrival. Processes. 2022;10(4):760. https://doi.org/10.3390/pr10040760.
160. Chang J, Yu D, Zhou Z, He W, Zhang L. Hierarchical reinforcement learning for multi-objective real-time flexible scheduling in a smart shop floor. Machines. 2022;10(12):1195. https://doi.org/10.3390/machines10121195.
161. Madumal P, Miller T, Sonenberg L, Vetere F. Explainable reinforcement learning through a causal lens. Proc AAAI Conf Artif Intell. 2020;34(03):2493–500. https://doi.org/10.1609/aaai.v34i03.5631.
162. Jensen SØ, et al. IEA EBC annex 67 energy flexible buildings. Energy Build. 2017;155:25–34. https://doi.org/10.1016/j.enbuild.2017.08.044.
163. Li H, Wang Z, Hong T, Piette MA. Energy flexibility of residential buildings: a systematic review of characterization and quantification methods and applications. Adv Appl Energy. 2021;3:100054. https://doi.org/10.1016/j.adapen.2021.100054.
164. Y.-W. Lin, T. L. E. Tang, and C. J. Spanos, "Hybrid Approach for Digital Twins in the Built Environment," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, New York, NY, USA: ACM, Jun. 2021, pp. 450–457. https://doi.org/10.1145/3447555.3466585.
165. Pean T, Costa-Castello R, Fuentes E, Salom J. Experimental testing of variable speed heat pump control strategies for enhancing energy flexibility in buildings. IEEE Access. 2019;7:37071–87. https://doi.org/10.1109/ACCESS.2019.2903084.
166. Zhou X, Du H, Sun Y, Ren H, Cui P, Ma Z. A new framework integrating reinforcement learning, a rule-based expert system, and decision tree analysis to improve building energy flexibility. J Build Eng. 2023;71:106536. https://doi.org/10.1016/j.jobe.2023.106536.
167. Yu L, Qin S, Zhang M, Shen C, Jiang T, Guan X. A review of deep reinforcement learning for smart building energy management. IEEE Internet Things J. 2021;8(15):12046–63. https://doi.org/10.1109/JIOT.2021.3078462.
168. Fu Q, Han Z, Chen J, Lu Y, Wu H, Wang Y. Applications of reinforcement learning for building energy efficiency control: a review. J Build Eng. 2022;50:104165. https://doi.org/10.1016/j.jobe.2022.104165.
169. Sifat MdMH, et al. Towards electric digital twin grid: technology and framework review. Energy and AI. 2023;11:100213. https://doi.org/10.1016/j.egyai.2022.100213.
170. D. Hugo *et al.*, "A smart building semantic platform to enable data re-use in energy analytics applications: the Data Clearing House," Nov. 2023.
171. Balaji B, et al. Brick: metadata schema for portable smart building applications. Appl Energy. 2018;226:1273–92. https://doi.org/10.1016/j.apenergy.2018.02.091.
172. Talei H, Benhaddou D, Gamarra C, Benbrahim H, Essaaidi M. Smart building energy inefficiencies detection through time series analysis and unsupervised machine learning. Energies (Basel). 2021;14(19):6042. https://doi.org/10.3390/en14196042.