

Article

A Framework for Integrating Vision Transformers with Digital Twins in Industry 5.0 Context

Attila Kovari ^{1,2,3,4} 

¹ Institute of Digital Technology, Faculty of Informatics, Eszterházy Károly Catholic University, 3300 Eger, Hungary; kovari.attila@uni-eszterhazy.hu

² Department of Computer Systems and Control Engineering, Institute of Computer Engineering, University of Dunaujvaros, 2400 Dunaujvaros, Hungary

³ Institute of Electronics and Communication Systems, Kandó Kálmán Faculty of Electrical Engineering, Óbuda University, 1034 Budapest, Hungary

⁴ GAMF Faculty of Engineering and Computer Science, John von Neumann University, 6000 Kecskemet, Hungary

Abstract: The transition from Industry 4.0 to Industry 5.0 gives more prominence to human-centered and sustainable manufacturing practices. This paper proposes a conceptual design framework based on Vision Transformers (ViTs) and digital twins, to meet the demands of Industry 5.0. ViTs, known for their advanced visual data analysis capabilities, complement the simulation and optimization capabilities of digital twins, which in turn can enhance predictive maintenance, quality control, and human–machine symbiosis. The applied framework is capable of analyzing multidimensional data, integrating operational and visual streams for real-time tracking and application in decision making. Its main characteristics are anomaly detection, predictive analytics, and adaptive optimization, which are in line with the objectives of Industry 5.0 for sustainability, resilience, and personalization. Use cases, including predictive maintenance and quality control, demonstrate higher efficiency, waste reduction, and reliable operator interaction. In this work, the emergent role of ViTs and digital twins in the development of intelligent, dynamic, and human-centric industrial ecosystems is discussed.

Keywords: Industry 5.0; digital twin; Vision Transformer; predictive maintenance; quality control; human-centric manufacturing; sustainability; real-time analytics; human–machine collaboration



Academic Editors: Stefano Marrone,
Laura Verde and Lelio Campanile

Received: 12 December 2024

Revised: 1 January 2025

Accepted: 4 January 2025

Published: 7 January 2025

Citation: Kovari, A. A Framework for Integrating Vision Transformers with Digital Twins in Industry 5.0 Context. *Machines* **2025**, *13*, 36.
<https://doi.org/10.3390/machines13010036>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the continuous development of industrial paradigms, technologies and manufacturing procedures have made great strides. The current transition from Industry 4.0 to Industry 5.0 is the reorientation of systems from being predominantly automation-based to more human-centric and environmentally friendly. This transition is also backed by emerging policy and research frameworks; for example, the European Commission’s vision of Industry 5.0, which focuses not only on further technological development, but also on increased human-oriented design, on sustainability, and on resilience of industrial systems. Although Industry 4.0 attempted to connect cyber–physical systems with industrial information and communication technologies and artificial intelligence to increase efficiency and automation [1], Industry 5.0 seeks to connect humans and machines in order to achieve higher productivity, personalization, and resilience in manufacturing processes [2]. To this end, emerging technologies are being designed not only with this in mind, but also to link industrial systems to the priorities of society and the environment.

Digital twin technology is one of the core technologies in current manufacturing. A digital twin is a virtual model of an analog physical system that allows features of real-time surveillance, simulation, and optimization [3]. Manufacturers can predict future performance, identify early problems, and improve system performance all without physical interaction by establishing a digital equivalent of the physical process. This approach is critical for improving the malleability, efficiency, and sustainability of production processes in the face of growing complexity [4].

On the other hand, progress in the field of artificial intelligence (AI), especially in computer vision, has resulted in the creation of highly sophisticated models such as Vision Transformers (ViTs). These Transformer-based models, extending the Transformer architecture to visual data [5], have shown better performance in object detection, image classification, and segmentation tasks than conventional convolutional neural networks (CNNs) [6]. ViTs possess the benefit of providing more global context understanding in an image, and thus are powerful for industrial applications in which careful visual analysis plays a key role [7].

Although both digital twins and Vision Transformers are increasingly applied in industry, much work remains to bridge the gap between these two technologies, especially in the context of Industry 5.0. Present digital twin implementations frequently utilize classical data analytics and machine learning models, which can neglect to make efficient use of the powerful information encoded in visual data [8]. Substantially, although ViTs have been successfully applied to visual data, their use in real-time adaptive systems like digital twins is still under development. This discrepancy restricts the capability of industries to react with real-time, data-based decisions from operational data and image-based information, which restricts possibilities for applications of predictive maintenance and quality control, and human–machine cooperation [9].

Among the advanced technologies driving the evolution from Industry 4.0 to Industry 5.0, ViTs and digital twins offer a particularly promising way to create multidimensional architectures that enhance operational capabilities. Unlike basic digital twins, traditional digital twins are mostly based on quantitative operational data; however, the incorporation of ViTs allows the analysis of qualitative visual data, therefore providing a systematic overview of system states. This synergy contributes to predictive maintenance by detecting, via visual cues, underlying subtle visual disarrangements that can flag equipment failure and enable early, effective intervention with reduced downtime. In addition, the strength of ViTs in image classification and object detection enables high-level quality control through the continuous monitoring of products' standards, consistent with Industry 5.0's personalization and customization. Integration also enables human–machine integration, delivering intricate data in easily understandable forms to allow operators to take evidence-based decisions, a hallmark of the human-centric approach of Industry 5.0 [9]. In addition, the complete data analysis not only performs process optimization, but also enables sustainable manufacturing through the reduction of both resources and waste. Ultimately, the massively scalable and flexible architecture of this integrated approach guarantees it remains applicable to a wide variety of industrial settings and continually adapts to technological progress, solidifying its role in the evolving terrain of Industry 5.0.

The aim of this paper is the definition of a conceptual framework that combines Vision Transformers and digital twins for Industry 5.0. This framework aims to maximize the benefits of Vision Transformers in visual data processing/analysis, while improving the predictive/simulation abilities of digital twins. Through the fusion of these technologies, the framework aims for better decision making, human–machine collaboration, and supports the creation of greener and more robust manufacturing systems. The proposed

methodology corresponds with the Industry 5.0 agenda by combining technical progress and human ingenuity in such a way as to lead to more personalized and intelligent industrial applications.

2. Background

The integration of powerful technologies is changing the industrial paradigm and inducing deep shifts in manufacturing processes and systems. In this part, a deep analysis of digital twins, ViTs, and the changing characteristics of Industry 5.0 is presented. The process of learning these concepts is essential to developing an understanding of the utility of integrating ViTs with digital twins in current industry.

2.1. Digital Twins

Digital twins are computerized representations of physical objects, processes, or systems, allowing real-time monitoring, simulation, and optimization. Originating from aerospace, the concept has been adopted more and more over the years due to its potential to enhance decision making as well as operational performance [10]. An industry can digitally recreate a real-world object and use it to simulate, analyze, and implement performance improvement without the need to interrupt live production, with the physical object in the physical world, the virtual form in the virtual world, and the data links between the two [11]. Sensors and IoT devices enable real-time data transfer, which makes a digital twin able to represent its physical counterpart very accurately [8]. Such synchronization allows organizations to derive useful information from system behavior and performance.

The functionalities of digital twins are multifaceted. Real-time monitoring makes available in real time the latest values of a number of performance parameters, which allows rapid recognition of anomalies, which in turn allows for preventive maintenance [11]. Their simulation and prediction capabilities provide engineers with the ability to test various options and evaluate the possible effect of changes without affecting the real production [12]. Optimization and control are realized by processing data patterns and making changes to increase efficiency, decrease down time, and decrease operating costs [13].

The benefits of using digital twins are numerous and extend to all branches of industrial operation. Improved operational efficiency comes by way of optimized processes and decreased waste [14]. Maintenance expenses are reduced by migrating from the reactive to predictive type of maintenance, reducing unforeseen halts [15]. Enhanced product quality is achieved due to the capability to simulate and test products in the virtual world prior to their physical production [16]. Moreover, the rapid innovation cycle is further enhanced by fast prototyping and testing in the digital space [17].

Digital twins are currently applied and accordingly show their transformative possibilities. In manufacturing, companies like Siemens and General Electric utilize digital twins to optimize production lines, enhance product design, and implement predictive maintenance strategies [18]. In the healthcare sector, digital twins are used to represent human organs for personalized medicine and surgical planning with beneficial effects on patients' outcomes [19]. Digital twins for traffic flow, energy use, and emergency planning strategies are used by urban planners in smart cities to model and optimize [20]. These instances highlight the multifunctionality of digital twins and the essential role they play in emerging fields.

2.2. Vision Transformers

Vision Transformers are an important breakthrough in computer vision. Extending the use of Transformer models in the field of natural language processing (NLP) to process images, ViTs adapt these models for image processing [21] to provide a new solution

for image recognition problems. Traditional convolution neural networks (CNNs) have served as the workhorse for computer vision, using convolutional layers to capture local features in images. Nevertheless, CNNs are frequently not able to encode global context and information because of their receptive local fields. This constraint is overcome by self-attention features, that are used in order to represent relations between all parts of an image at the same time [22].

The advantages of ViTs over traditional CNNs are manifold. In the first place, the global context awareness generated through self-attention mechanisms has provided the models with the capability to learn hierarchical visual structure and rectified the shortcoming of only focusing on the local context in that the models improve their performance for image comprehension tasks that need to understand the whole image [23]. Second, ViTs can achieve similar performance with reduced parameters (MMVIO) when using K-fold training on data of a large size, consequently reducing the computation load and resource consumption [24]. Third, the easy extension of the Transformer architecture is leveraged by ViTs to perform diverse computer vision tasks such as object detection, segmentation, and video analysis [25]. ViTs also reduce the reliance on inductive biases in CNNs such as locality and translation invariance, so that the model learns these constraints directly from the data [26].

The application of ViTs to computer vision problems has been shown in multiple studies. For image classification, ViTs have produced state-of-the-art results on benchmarks such as the ImageNet, and outperformed classical CNNs in accuracy [27]. In object detection, models such as DETR (Detection Transformer) have made the detection pipeline simpler by dispensing with the requirement for hand-engineered parts and have shown competitive performance [28]. ViTs have also been successfully used in semantic segmentation, enhancing the accuracy of pixel-level classification tasks [29]. In addition, the generalization of ViTs has expanded their use in the video context for video understanding, whereby ViTs process the sequential frames to capture the spatial and temporal information to achieve goals like action recognition [30].

The appearance of ViTs represents a new paradigm in computer vision and provides opportunities to process intricate visual information in industrial sectors. Their capacities to find global context and simulate long-range dependencies make them well suited for applications involving complex visual patterns and structures.

2.3. Industry 5.0 Requirements

Industry 5.0 is the next stage of industrial evolution; it is characterized by a human-in-the-loop model, in which technology design is implemented in tandem with public and ecological needs. In contrast to Industry 4.0, in which automation and digitization are the focus, Industry 5.0 oversees the coupled human-machine interaction and development and innovation based on sustainability and resilience in the manufacturing and connected domains.

Personalization and customization are among the priority requirements of Industry 5.0. There is a growing demand for products able to satisfy the needs and preferences of the consumer. Therefore, Industry 5.0 provides the solution by allowing mass customization at the expense of efficiency [31]. Flexible manufacturing systems and agile supply chains are the key to creating personalized items on a large scale [32]. Advanced data analytics and artificial intelligence technologies are also essential to consumer-preference exploration and to the transformation into production processes [33].

Collaboration between humans and machines is another key issue. Industry 5.0 is regarded as a symbiotic partnership between humans and intelligent machines [34]. Collaborative robots (cobots) are designed to live and work side by side with humans

to improve them and perform repetitive/hazardous tasks. Synergetic effects improve work efficiency, creativity, and work satisfaction because of the freedom to be creative and strategic [35].

Sustainability and resilience are also central to Industry 5.0. The greatest emphasis is placed on environmental sustainability, trying to minimize waste, emissions, and resource use [36]. Industries are invited to implement circular economic practices, i.e., products, and materials are re-used and recycled to reduce the environmental footprint [37]. Adaptive systems that can restore themselves from shocks, including stock chain network disruptions or market chills, are essential for Industry 5.0's survival long term [38]. This includes the adoption of effective risk management mechanisms and use of digital technologies to promote transparency and responsiveness [39].

Another Industry 5.0 requirement is data analytics-based enhanced decision making. The effective use of big data and deep analytics enables industries to gain meaningful information from massive big data, which can enhance productivity and drive innovation [40]. Embedding artificial intelligence and machine learning allows predictive maintenance, supply/demand forecasting, and process control that can be used to gain a sense of what the consequences may be [41].

Social responsibility is an integral component of Industry 5.0. There is awareness of the value of social sustainability such as fair labor, involvement of the community, and the provision of value to society [42]. It is anticipated that industries should uphold ethical standards, protect workers, and practice in a way that is mutually beneficial for all stakeholders. This integrated perspective guarantees that technological advancement is in correspondence with social values and leads to inclusive growth [43].

To satisfy these requirements, it is necessary for industries to take up technologies which allow flexibility, collaboration, and sustainability. The integration of digital twins and ViTs aligns with these goals by enhancing real-time monitoring, predictive capabilities, and the ability to process complex visual data. Digital twins enable simulation and optimization of processes towards efficiency and sustainability. ViTs improve the processing of visual information and support quality control, as well as usher in the next generation of human-machine interactions.

2.4. Integration

The combined use of ViTs with digital twins is a major leap forward in the solution to the complex needs of Industry 5.0. The current contribution aims at overcoming the gap between what has been available so far through the combination of the powerful visual data processing of ViTs with the simulation and optimization abilities of digital twins. Such integration is not just a technology convergence, but a strategic fit with a purpose to nurture an intelligent, timely, and humanized manufacturing ecosystem.

In the current implementations, digital twins and ViTs are treated as two independent entities, which also restricts the possibilities offered by synergy between them. The absence of a unified structure leads to industries not being able to fully leverage the abundance of visual information contained in ViTs; in the digital twin context, these data are not contributing to the development of real-time decision making or predictive ability.

By integrating ViTs with digital twins, certain issues that industries are encountering in the transition from Industry 4.0 to Industry 5.0 can be alleviated. For example, advanced visual data processing may lead to better predictive maintenance through the correct detection of wear and failure in machinery using live image analysis. It also has the potential to facilitate more natural human-machine interaction through the generation of detailed visual feedback and the ability of operators to communicate with systems in a visual language with no need for symbolic manipulation. Moreover, this fusion may also

help to take sustainability action such as process efficiency by analyzing profound data, waste reduction, and resource optimization.

Nevertheless, at the time of writing, there is no well-defined architecture that holistically combines ViTs with digital twins in the Industry 5.0 paradigm. The lack of such a framework is a missed chance to provide the opportunity to leverage the full power of these technologies to answer the intricate requirements of contemporary manufacturing. To achieve a well-defined scheme, complementary to the powerful visual processing capability of ViTs, a reactive simulator, and the optimization capability of digital twins, industries require a structured multidisciplinary method to merge the advanced visual processing of ViTs with the real-time digital simulation and optimization of a digital twin.

Developing an integrated architecture is critical to fill this gap. Such a framework will serve as the structure and workflow for the application of ViTs across digital twin worlds to guarantee the pipeline, interoperability, and scalability of the data. It would overcome technical issues including data integration, computation demands, and system design, and, in turn, would be compatible with the human-centered and sustainable goals of Industry 5.0.

2.5. Introduction to a New Framework

Recognizing the need for an integrated approach, this paper proposes a novel framework that brings together ViTs and digital twins within the Industry 5.0 paradigm. The aim of the proposed framework is to exploit the benefits of both the technologies for user friendliness and enhanced manufacturing processes, supporting human–machine cooperation, and for sustainable and resilient manufacturing.

Through integration of ViTs into digital twins, it is possible to obtain more accurate and more complete system models, which include operational and visual data. This integration enables advanced analytics, predictive maintenance, and real-time optimization, contributing to improved efficiency and decision making. In addition, it is compatible with the characteristics of Industry 5.0 through a personalization process, better human–machine interaction, and sustainable practice.

The following subsections describe the proposed framework (i.e., its architecture, components, and applications). Hence, the presented framework focuses on filling the gap in technology integration and provides a basis for further research and development in industrial environments, thus supporting access to the next generation of smart manufacturing systems.

3. Methods

To meet the requirements of high-technology industries, an integrative framework for embedding ViTs into digital twins was developed in a systematic manner. System components, including data acquisition, preprocessing, ViT integration, digital twin components, and user interaction, were established at the beginning of the process. For scalability and adaptability in a variety of industrial environments, a modular architecture was designed. To clarify the process that led to our final framework, a multi-level process was performed, starting with a broad literature review. At first, the current literature on Vision Transformers, digital twins, and critical aspects of Industry 5.0 was analyzed, including human factors, ecology, and real-time analysis. This review established a base set of principles regarding viable data streams, model structures, and operational limits. Then, the knowledge from these sources was combined into a conceptual design by identifying the main modules—data collection, preprocessing, ViT embedding, digital twin core, and user interface—and how they operate in an industrial context. Finally, the work streamlined and packaged these building blocks into a uniform, step-by-step workflow, made so that the

framework could be adapted to many different use cases, including predictive maintenance and quality control. Through the above steps, a scalable, versatile architecture was reached in line with the emerging needs of Industry 5.0.

ViTs were selected because of their ability to leverage self-attention mechanisms to capture the visual information with high accuracy. For the robustness of the model, the visual data were preprocessed by patching and augmentation. To further improve the digital twin and allow for real-time replication, predictive simulations, and system optimization, these emerging insights were integrated with operational data.

In order to verify the applicability of the framework, some use case scenarios such as quality control and predictive maintenance were implemented and tested in simulated environments. To achieve accordance with the goals of Industry 5.0, these scenarios presented the ability of the system to recognize equipment abnormalities, to model the impact of defects, and to support real-time interventions.

This methodology guarantees the appropriateness of the framework to the goals of Industry 5.0 and provides a valid, reproducible, and human-centric solution to predictive analytics and real-time monitoring.

4. Proposed Framework

In this section, the paper presents the core outcome of this work, an integrated framework which couples Vision Transformers and digital twins, designed to meet the needs of Industry 5.0. The following chapters describe individual aspects and workflows within this general framework, demonstrating the coupling between data acquisition, model integration, and real-time analysis to facilitate human-centered and long-term operations.

The integration of ViTs with digital twins offers a novel approach to addressing the challenges and requirements of Industry 5.0. The aim of this proposed framework is to enhance real-time data analysis, predictive ability, and human-machine interaction, stemming from the high-level visual data processing provided by ViTs, as well as the simulation and optimization capabilities of digital twins. This part presents the framework architecture and the main parts of the framework, as well as the process of the framework that converts the raw data into a valuable insight.

4.1. Architecture Overview

The structure of the proposed framework is designed to promote natural interaction between the real world and the digital image. All elements of the framework consist of a stack of layers that are interconnected and can facilitate a natural information flow. They consist of data acquisition, preprocessing, ViT integration, digital twin upgrade, and user interaction. The modularity and scalability of the top-level design are exploited both to allow generalization of the system to other industrial settings and to be combined with future technological advances. The modularity and scalability of the highest level of design are harnessed both from a generalization of the system to other industrial domains and in order to be integrated with future technological developments. Figure 1, illustrating the high-level architecture, consists of the following modules that are interconnected components:

- Data acquisition layer;
- Preprocessing module;
- Vision Transformer integration;
- Digital twin core;
- User interface and interaction layer.

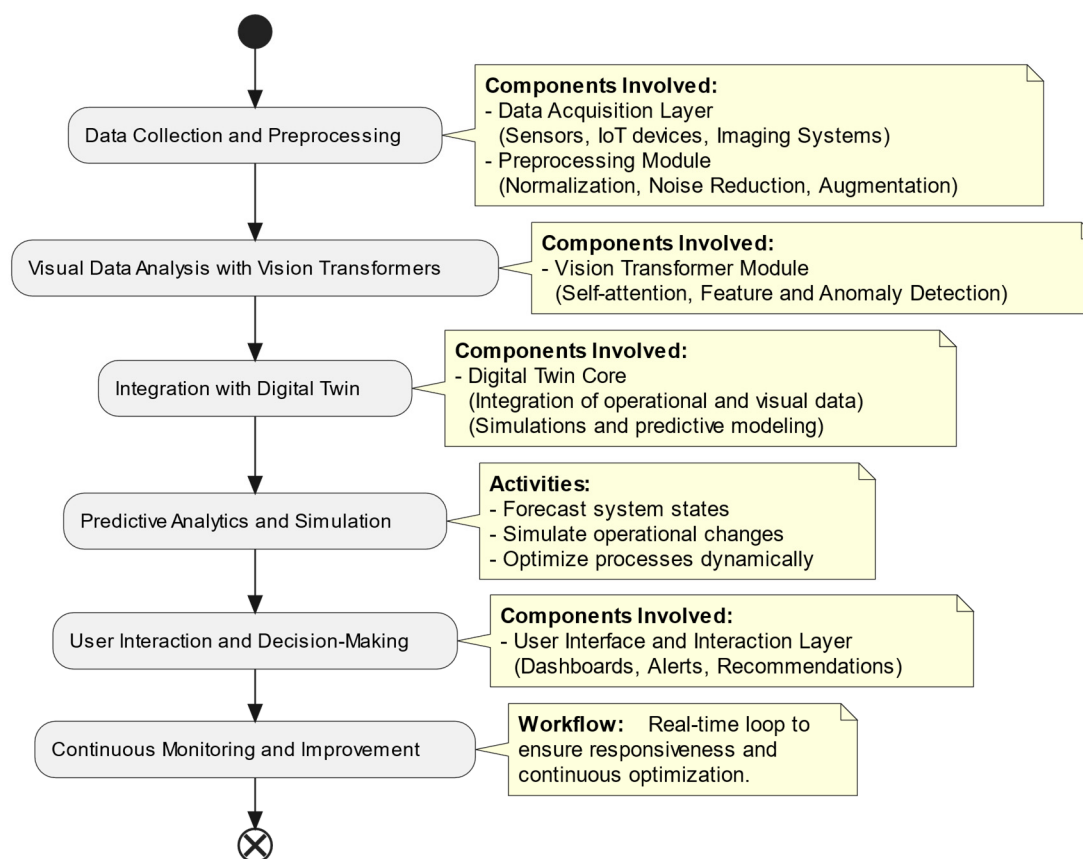


Figure 1. High-level workflow of the proposed framework integrating ViTs with digital twins.

All of these components cooperate in order to provide a stream of information that fits the system to the challenges of the industrial environment, which continuously change.

For the purpose of providing more useful suggestions to the practitioners in industry, I briefly outline in this section the major constituents of the proposed framework as well as how they can be applied in practice:

Data acquisition: Practitioners can combine standard industrial IoT sensors (e.g., temperature, vibration) with high-resolution or infrared cameras. These can be implemented by using OPC UA protocols or lightweight MQTT brokers at the edge, based on factory limitations.

Preprocessing: In this stage, the process is generally to filter the raw sensor stream, to apply image enhancement (noise reduction/augmentation), and to align time-series data. Instead, libraries (e.g., OpenCV or scikit-image, which are readily available) can be readily ported here with little to no modifications.

Vision Transformer integration: While ViTs can be computationally expensive, small, fine-tuned models can be executed on GPUs or high-end edge devices. As with PyTorch's model quantization, there are tools that can be applied to speed up real-time inference.

Digital twin core: Practitioners can use either a commercial digital twin application or an open-source simulation software (e.g., robotics ROS 2 or Gazebo Harmonic or Ionic). Synchronization between the sensor and vision inputs with the virtual model should be maintained in real time.

User interaction: The implementation details are dictated by the complexity of the user interface. Industrial SCADA systems, web dashboards (using frameworks such as Node-RED), or AR headsets can be incorporated to provide intuitive operator control and monitoring.

4.2. Components Description

This section provides a detailed breakdown of the framework's architecture, highlighting the roles and functionalities of each component. By systematically defining each stage from data acquisition to user interaction, the section gives a holistic view on how the system combines various technologies to realize real-time monitoring, analysis, and decision making.

4.2.1. Data Acquisition Layer

The basis of the framework is in the multidimensional data of the physical environment. An Internet of Things (IoT) system and sensors are installed in the industrial setting to collect operational data (e.g., temperature, pressure, vibration, and other important parameters) [44,45]. Such devices allow the real-time control of equipment and processes, hence constituting the infrastructure of the Industrial Internet of Things (IIoT). At the same time, high-performance cameras and imaging systems record visual information and offer high-resolution images and videos of machines, products, and processes. Edge computing nodes perform local computations of the data stream, helping to mitigate latency and central system load, a factor for real-time behavior [46].

4.2.2. Preprocessing Module

The raw data are then preprocessed to ensure quality and compatibility with subsequent analytical models. Standardization of the operational data is achieved by scale and unit normalization to allow comparison and analysis that is not confounded by scale and unit variability. In the case of visual data, noise reduction algorithms increase the quality of images through the filtering of spurious information or distorted information. The data augmentation techniques (e.g., rotation and scaling of data, and flipping of the data) in the dataset [47] not only help to increase the diversity of the data, but also can help to avoid over-fitting to make the ViTs more robust. The images are then rescaled to fixed-size patches, a necessary step for processing by ViTs.

4.2.3. Vision Transformer Integration

The Vision Transformer module is the core analytical component of the framework, leveraging the advanced capabilities of ViTs to process and interpret visual data. In contrast to conventional CNNs, ViTs incorporate self-attention methods to learn the global context as well as long-range relationships in images, thereby achieving a more efficient, complete understanding of visual statistics. Dividing images into patches and computing their positional information, ViTs are able to analyze the sophisticated visual patterns and anomalies that are important for tasks such as defect detection and quality evaluation in manufacturing.

4.2.4. Digital Twin Core

Insights captured by the ViTs are combined into the digital twin and this dramatically improves the accuracy and predictive capability. The digital twin constantly updates a continuous stream of real-time operational and visual information to guarantee the validity of the virtual representation as an image of the real-world counterpart [48]. This enhanced digital model can be used for simulating an array of operational conditions, making failure predictions, and optimizing processes in a dynamic way. Through the addition of detailed graphical information, the digital twin has the capacity to model more accurately and improve the prediction of possible problems, and to optimize processes by processing both qualitative and quantitative data. Predictive analytics models use this augmented dataset to predict future states, for example, for predictive maintenance and for decision

making [49]. A number of researchers have explored the methods and support possibilities of different digital user tools and robotics in different contexts [50–52].

4.2.5. User Interface and Interaction Layer

Central to the framework is the user interface and interaction layer, which provides the translation between the human operator and the integrated system. This layer offers intuitive dashboards and visualization capabilities, which deliver the most relevant performance metrics and alerts, as well as actionable information from the digital twin [49]. Operators can communicate with the system through control panels by programming parameters, by starting simulations, and by responding to alarms according to the data-driven suggestions. This interaction creates a collaborative situation, where the experience of human expertise is coupled with the capabilities of machine intelligence and in which the human-centered goals of Industry 5.0 are to the fore.

4.3. General Workflow Explanation for the Proposed Framework

The system operational workflow of the proposed framework is continuously iterative data transformation toward actionable and utilizable information, which results in improved system performance and resiliency. The workflow can be delineated into the following stages:

- Data collection and preprocessing;
- Visual data analysis with ViTs;
- Integration with the digital twin;
- Predictive analytics and simulation;
- User interaction and decision making;
- Continuous monitoring and improvement.

Figure 2 shows the data flow diagram and Figure 3 shows component interactions in the proposed framework.

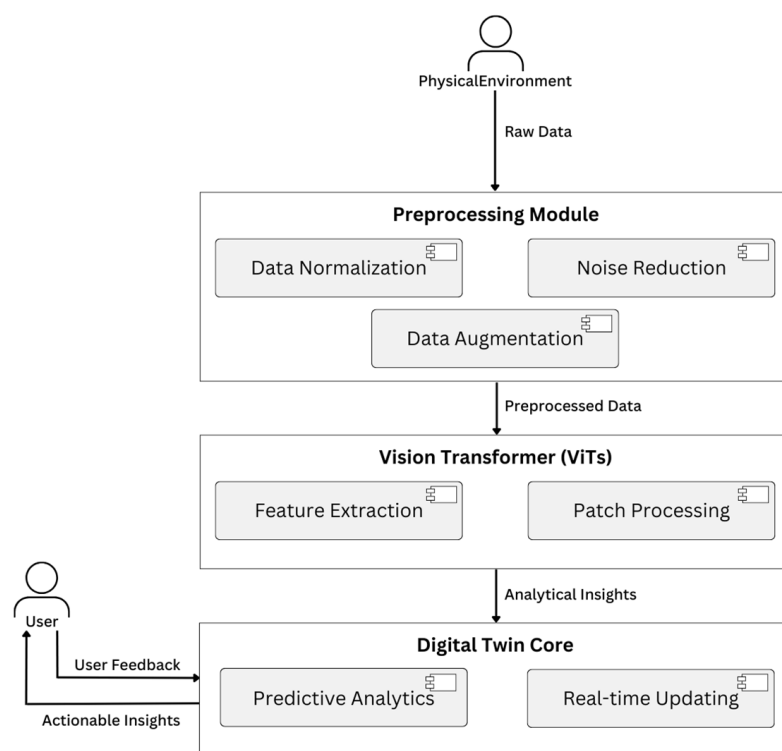


Figure 2. Data flow diagram of the proposed framework integrating ViTs with digital twins.

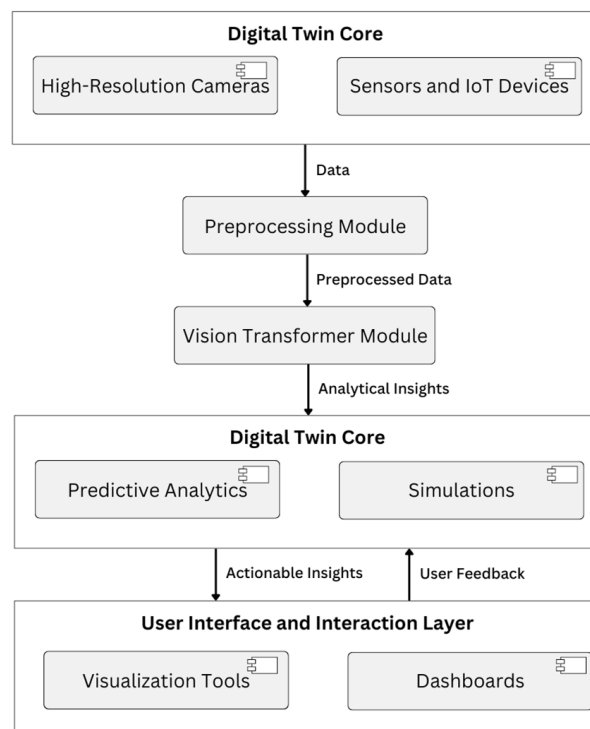


Figure 3. Component interactions of the proposed framework integrating ViTs with digital twins.

4.3.1. Data Collection and Preprocessing

The collection of operational and visual data from the physical world is started by synchronous acquisition of the data. Sensors and IoT devices transmit real-time measurements, while imaging systems capture high-quality visual data. These data are then passed to the preprocessing module, where they are normalized, cleaned, and augmented to ensure quality and consistency.

4.3.2. Visual Data Analysis with Vision Transformers

Processed visual information is input into the Vision Transformer module in which ViTs extract features, patterns, and outliers by the application of self-attention mechanisms. From this analysis, defects can be identified, equipment health can be monitored, and product quality can be evaluated, offering invaluable information to guide further decision making.

4.3.3. Integration with the Digital Twin

Insights computed by the ViTs are fed into the digital twin core. This integration adds detailed visual information to the virtual model, improving the performance of the model in reproducing and predicting system behavior accurately. The enhanced digital twin enables the incorporation of both quantitative and qualitative information in the form of operational metrics and visual information, consequently resulting in more holistic simulations and predictions.

4.3.4. Predictive Analytics and Simulation

Based on unified data, the digital twin carries out predictive analysis to predict future states of the system. It is capable of modeling the effects of many operational modifications, forecasting the likelihood of equipment failure, and refining processes to yield improved efficiency and reduced waste. These kinds of predictive functions allow for preventive maintenance and wise courses of action, in line with the sustainability and resilience objectives of Industry 5.0.

4.3.5. User Interaction and Decision Making

Operators communicate with the system through the user interface and interaction layer, enabling the system to receive real-time dynamic notifications and recommendations because of the analysis of the digital twin. The simple visualization of complex data by means of instinctive dashboards and visualization tools enables the easy comprehension of the data and subsequent rapid decisions. Operators can modify system parameters, create work orders, and implement process optimization based on the actionable intelligence that is presented to them.

4.3.6. Continuous Monitoring and Improvement

The architecture works in a recurrent cycle, where the data are collected and analyzed and decisions are made on the fly. When new data are acquired, they are processed and loaded into both the ViTs and the digital twin, thus keeping the system current and feed-forward to changes in the physical world. This real-time monitoring allows continuous evolution and adaptation, keeping the manufacturing system robust, resilient, and in tune with the goals of Industry 5.0.

4.4. Implementation Strategy and Technological Details

Comprehensive collection, processing, and integration of data from multiple sources, sophisticated analytical techniques, and smooth communication between digital and physical systems are all necessary for the effective implementation of the suggested framework. This section describes the framework's multi-step implementation process, highlighting the interaction of operational, visual, and simulated data streams.

Step 1: Multi-Source Data Collection

The process starts with the collection of data from various sources that are divided into operational, visual, and process-related categories. IoT sensors that track vital parameters like temperature, pressure, vibration, energy usage, and machine health are used to collect operational data in real time. Continuous insight into the performance of the physical system is ensured by these data. Edge computing devices preprocess the operational data on site to minimize latency and facilitate quick decision making. In parallel, visual data are acquired using high-resolution cameras, which are deployed to receive multi-viewpoint images of machinery and work processes. Including thermal and infrared cameras, these cameras sense anomalies (i.e., overheating or abnormal heat dissipation). Process and product data are also integrated, including static engineering models like CAD files, manufacturing blueprints, and technical specifications. These static datasets can be used as a reference to detect discrepancies between predicted and observed results, allowing them to detect potential problems at an early stage.

- Operational data: A network of Internet of Things sensors continuously tracks important physical system parameters like vibration, temperature, pressure, machine health, and energy usage. Process optimization and real-time diagnostics depend on these data. In order to reduce latency and facilitate quick on-site decision making, edge computing devices are strategically placed to preprocess the data locally.
- Visual data: Real-time visual data are captured by high-resolution cameras and imaging systems, giving granular insights into machinery and industrial processes. Comprehensive situational awareness is ensured by multi-angle imaging. By identifying abnormalities like overheating or problems with heat dissipation, specialized systems such as thermal imaging and infrared sensors provide an additional layer of detail.
- Process and product data: The system incorporates static data sources like technical blueprints, manufacturing specifications, and computer-aided design (CAD) files. To

ensure adherence to technical standards and early detection of discrepancies, these static models enable comparisons between expected designs and real-world deviations.

Step 2: Data Ingestion, Preprocessing, and Layered Fusion

In the second step, the recorded data are processed and combined such that the results are calibrated and accurate for subsequent processing steps. Operational and visual data are rescaled within and among domains and time-series data across sensors are aligned to preserve temporal coherence. Data cleaning techniques identify and rectify inaccuracies or missing points, thereby improving data quality. The visual data are further subjected to advanced augmentation, such as random cropping, rotation, and noise injection, which increase the range of generalization of the Vision Transformer. Pictures are then cropped into a set of small patches, where salient spatial and temporal information exists, to be put into ViT analysis. Multi-source data streams are further aggregated in different regions after preprocessing. One form of feature-level fusion is fusing raw sensor data and images directly and permits the combined use of operational and visual data in simultaneous processing. The global model understanding, at the decision level, is obtained from the aggregated individual models (i.e., anomaly detection and defect classification).

- **Data normalization and cleaning:** To address variations in units and scales, the gathered data are standardized. Synchronization methods guarantee consistent temporal alignment for sensor time-series data. A high-quality dataset is ensured by employing statistical error-handling techniques to identify and eliminate anomalous or incomplete data points.
- **Data augmentation for visual analysis:** Techniques like noise injection, color jittering, rotations, and random cropping are used to improve visual data. These techniques enhance ViTs' generalization and resilience. In order to capture the spatial and temporal features necessary for further analysis, images are separated into smaller, fixed-size patches.
- **Layered multi-source data fusion:** A multi-layered method of merging data streams from operational, visual, and process sources.
- **Feature-level fusion:** In order to find connections between observed patterns and physical conditions, feature-level fusion directly integrates visual inputs with raw sensor data.
- **Decision-level fusion:** Creates a thorough understanding of the system by combining the results of several analytical models, such as anomaly detection and defect classification.

Step 3: Hierarchical Data Analysis Using Vision Transformers

This step involves the application of ViTs to the preprocessed data in a hierarchical analysis, resulting in detailed information at different abstraction levels. Visual data patches are processed by ViTs, where a lower layer learns fine-grained information (i.e., surface defects or subtle wear), and another layer collects high-level patterns (i.e., general system performance trends or serious failures). ViTs also learn cross-sensor mappings by correlating operational deviations (e.g., temperature spikes) and visual analogues (e.g., mechanical wear). This hierarchical scheme enables multi-level anomaly detection; that is, between minor variations, i.e., small visual defects, and serious ones, i.e., equipment breakdown, or serious production outage. Such an analysis is carried out in predictive tasks, where ViTs defect-classify and predict the likelihood of system-level failure from long-term behavior (proactive maintenance and operational stability).

- **Multi-stage feature extraction:** The Vision Transformer module processes image patches layer by layer, with deeper layers capturing higher-level patterns like system-

wide malfunctions or operational trends and initial layers detecting fine-grained details like micro-defects.

- Cross-sensor correlation: ViTs create links between sensor readings and visual abnormalities. To verify equipment health issues, for example, visual indicators of wear found in images are correlated with spikes in vibration or temperature data.
- Hierarchical anomaly detection: The analysis allows for precise and prioritized responses by differentiating between critical system-wide disruptions and minor deviations, such as minor defects or slight operational fluctuations.

Step 4: Advanced Integration with the Digital Twin Core

The information acquired during data processing is stored in the form of the digital twin, which is further optimized for real-time monitoring, simulation, and optimization. The digital twin is in steady synchronization with the operation and visual information, such that it always maintains a valid representation of the physical system. Through deep data fusion, data integration of operational measures, for instance, vibration and temperature, is combined with other corresponding data, such as visual intelligence, defect identification, etc., and a fused model is produced. The enhanced digital twin offers functionalities towards more complex simulations thanks to the capability to ingest multi-modal data, which may, for an operator, not only visualize complicated situations, but also allow increasingly accurate prediction of more problems. The digital twin is also flexible enough to adapt to a wide range of industrial use cases because it can be customized to fit various production settings, machine configurations, or product requirements.

- Continuous synchronization: The digital twin is updated in real time to reflect the most recent operational states and visual observations of the physical system, ensuring continuous synchronization. An accurate and dynamic virtual representation is ensured by this synchronization.
- Deep data fusion: The integration creates a unified model for improved simulations by combining ViT-driven visual insights (like defect localization) with operational metrics (like temperature and vibration).
- Personalization and use-case adaptation: The system enables operators to modify simulations according to particular equipment configurations or production settings. This flexibility accommodates a range of industrial settings and product iterations.

Step 5: Multi-Layered Predictive Analytics, Simulations, and Optimization

Short-term predictive analytics uses real-time operational and visual degradation to predict and address immediate maintenance requirements, whereas long-term forecasting relies on trend analysis to predict future problems and reconfigure maintenance schedules. A digital twin can also be used for case-study simulations, where operators can test-drive “what-if” scenarios including factor variations, equipment failures, and environmental variations. Such simulations are compatible with adaptive changes that optimize energy, reduce waste, and provide robustness against perturbations, which are in line with the sustainability objectives of Industry 5.0. Multi-objective optimization also fine-tunes system performance by iteratively modifying the parameters of machines and production processes to find the optimal trade-off between efficiency, quality, and exploitation of resources.

- Predictive maintenance: Long-term forecasting anticipates future trends to optimize maintenance schedules and cut costs, while short-term analyses pinpoint urgent maintenance requirements based on observed degradation.
- Scenario simulation: In order to enable proactive adjustments and ensure system resilience, the digital twin simulates “what-if” scenarios, such as equipment failures or environmental changes.

- **Multi-objective optimization:** The system can concurrently optimize resource consumption, quality, and efficiency thanks to real-time data feeds. Alignment with Industry 5.0 objectives is ensured by dynamic adjustments to production lines or machine parameters.

Step 6: Multi-Modal User Interaction and Decision Support

The framework emphasizes a human-in-the-loop approach as it includes the tools that can potentially improve operator decision making. This is about designing an intuitive interface, which is used to cross-over between operators and the entire system. Real-time data visualizations and alerts are presented on a set of customizable dashboards that give operators real-world actionable information about predictive analytics, process optimization, and anomaly detection. The use of AI-driven recommendations improves the pros and cons through the suggestion of maintenance actions or operational changes as a function of data analysis. Augmented reality applications, in turn, continue to enhance the operator's control by superimposing visual instructions for maintenance tasks or repairs by real-time information gathered from the digital twin and ViTs. AR-based training simulations enable the virtual modeling of workspaces, where operators can execute maintenance procedures or responses to disasters, as a means of enhancing workforce preparedness and system accuracy.

- **Interactive dashboards:** Operators can view real-time system performance, predictive analytics, and alert visualizations, facilitating prompt and well-informed decision making.
- **AI-assisted recommendations:** By utilizing its analytical capabilities, the system produces practical recommendations, like process modifications or maintenance scheduling.
- **Augmented reality (AR) integration:** AR tools enhance operator efficiency and skill development by superimposing visual guides for maintenance tasks or training simulations.

Step 7: Continuous Monitoring and Adaptive Learning

The architecture works under a continuous feedback loop, so that it becomes adaptive to changes in the operational environment and to new data inputs. Automated feedback loops keep the system in real time by updating the system with current operational conditions, and consequently both the ViTs and the digital twin represent current operational conditions. Machine learning models are refitted periodically with data, and better results are achieved in analyzing the anomalies, predictive analysis, and optimization. The system is also built to be self-repairing, by which one can perform autonomous corrections, for example, modifying the production lines or adjusting the machine parameters in order to avoid the devastating failure. The scalability of the framework architecture enables its deployment in a wide range of industrial settings, and its adaptable nature allows it to leverage the latest technologies, such as 6G communication, quantum computing, and the next generation of AI models, maintaining its relevance over time and its adaptability.

- **Adaptive machine learning:** Models are retrained using new data, increasing the accuracy of process optimization, anomaly detection, and predictive analytics.
- **Self-healing mechanisms:** To ensure fewer interruptions and increased resilience, the system automatically adapts to deviations by rerouting tasks or adjusting machine settings.
- **Scalability and future integration:** The modular architecture can handle growing technologies like 6G, quantum computing, or next-generation AI models, and it supports scaling across various manufacturing environments.

5. Use Case Scenarios

The proposed framework integrating ViTs and digital twins has potential as a tool for many Industry 5.0 applications. Three use cases are presented below, all of which are based on information and process descriptions, and each is illustrated by an accompanying flowchart that joins these scenarios into a consistent structure. The three scenarios presented here—predictive maintenance, quality control, and human–robot collaboration—were selected based on the key research gaps identified in the Introduction and Background sections. In particular, we are concerned with the paucity of effective, vision-based predictive analytics in industrial applications, the deficit of visual inspection and real-time control integration for achieving higher product quality, and the increasing focus on human-centric automation, wherein machines and operators work together synergistically. Through the association of each case study with these two aforementioned gaps, we present how the proposed framework can be seen as a structured answer for the theoretical and practical problems discussed previously, thus making our solution pathway directly respond to the fundamental questions within Industry 5.0 studies.

5.1. Predictive Maintenance Scenario

Unplanned downtime caused by equipment failure is an important problem in manufacturing. This framework offers an anticipatory solution by way of predictive maintenance through the merging of operational and visual data streams. Predictive maintenance is a key application of the proposed framework, which utilizes ViTs and digital twins to detect failures proactively (Figure 4).

The process begins with a multi-sensor dataset, including operational datatypes collected from sensors and IoT devices (e.g., temperature, pressure, and vibrational), to work with and simultaneous high-resolution and thermal snapshots recorded with high-resolution and thermal cameras. This bi-stream method guarantees a high-resolution description of the condition of the equipment, containing not only quantitative but also qualitative information.

The preprocessing phase ensures data quality and consistency. Operational data are normalized to reflect sensor-dependent differences and outliers are censored to deal with points that are deemed invalid. Visual data are improved by noise reduction and are extended by means of cropping, rotation, and scaling to make them ready for ViT analysis. Images are segmented into patches, maintaining spatial associations that are essential to anomaly detection. These preprocessed data are convolved into multiple levels, i.e., feature-level fusions interleave raw operational metrics with extracted visual features, and decision-level fusions interleave the outputs of independent analysis models to obtain a unified view of the equipment health.

The processing phase is based on Vision Transformers, that provide a deep analysis of visual information. ViTs use self-attention motifs to analyze image patches and to detect unusual things (e.g., surface cracks, wear patterns, material discolorations). These visual results are cross-compared to functional irregularities, e.g., spike in vibrations or temperature, to verify possible problems. This correlation improves the accuracy of anomaly detection and reduces the rate of false positives and false negatives.

Anomalies detected and their associated correlations are integrated into the digital twin, a continuously evolving virtual representation of the physical equipment. The digital twin replicates the possible effects of these disturbances on the operation of the equipment, predicting failure time and chain reaction effects on coupled systems. For instance, a crack in a turbine blade is simulated in order to predict its effect on vibration and the energy yield. The simulation in this example can explain in detail the failure prediction and the generation of maintenance prescriptions.

At the predictive analytics stage, the system provides actionable outcomes, including at what time interval maintenance should occur and what resources should be used. Maintenance schedules are optimized to align with planned downtime, minimizing disruptions to production. Alerts are sent to operators through dynamic dashboards, providing details on detected anomalies, predicted failures, and specific maintenance actions.

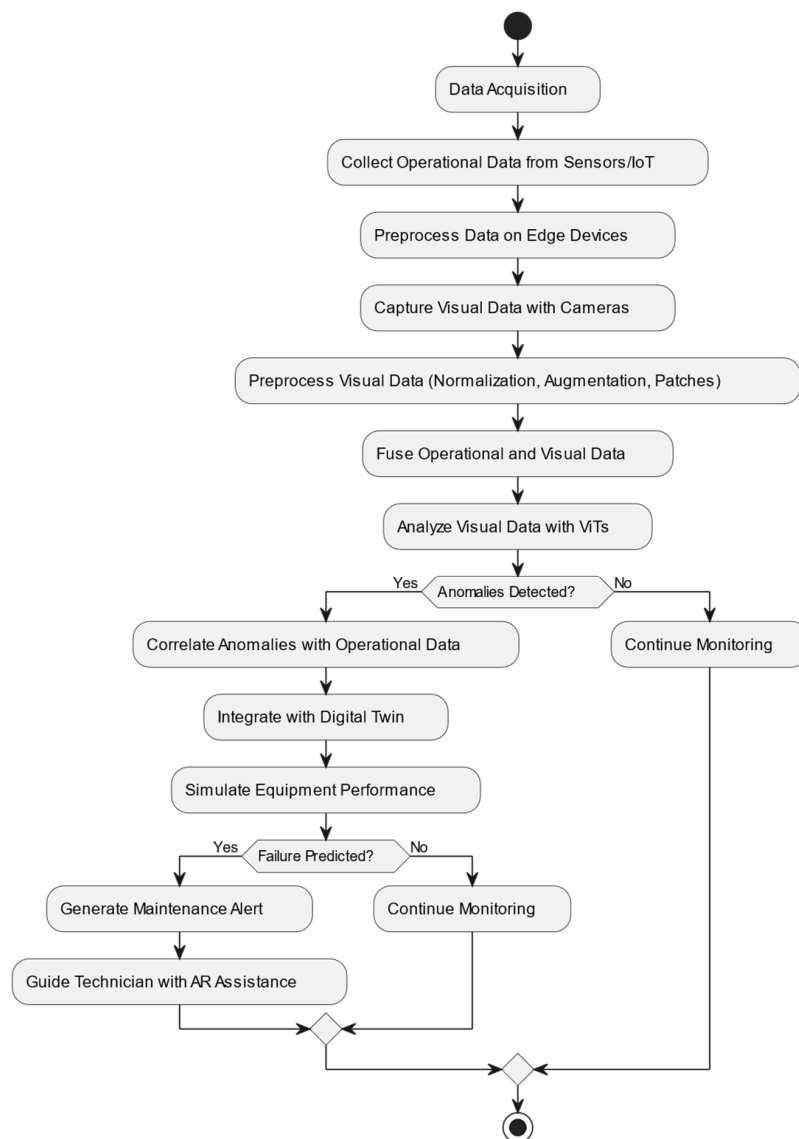


Figure 4. Process flowchart for predictive maintenance based on the framework.

Operator interaction and augmented reality (AR) are the last steps in the process. Operators see the generated insights on dashboards and are shown how-to instructions using AR overlays which present serial visually guided instructions for repair or substitution of components. AR could, for example, annotate the exact position of a crack in a component and superimpose diagrams of the required tools and methods on the view. Following maintenance, the operator's feedback is fed into the system to improve future predictions and simulations.

This ongoing feedback mechanism guarantees that the framework is able to learn about the new data and changing operational environments, and thus it continuously upgrades its accuracy and performance with time. The outcome is an effective predictive maintenance pipeline that saves on downtime, extends equipment's life, and optimizes resource utilization.

Even without an explicit simulation now, we have already begun to define common industrial sensor arrays (e.g., 1 kHz temperature and vibration sensor arrays, one-frame-per-second integrated thermal cameras) for which the ViT–digital twin proposed integration would provide near-real-time machine health monitoring. During a probable mission, after the Vision Transformer identifies developing microcracks or abnormal heat signatures on rotating elements, the digital twin automatically inverts the time series of the anomaly and can forecast whether and when a catastrophic failure will likely occur. This predictive ability enables operators to prepare interventions better, especially in facilities in which downtime is fitted to be as low as possible. Despite the absence of numeric constraints now, our conceptualization is consistent with conventional reliability-centered maintenance approaches, thus setting the ground for future empirical research.

Expected Outcomes

The integration of ViTs with digital twins for predictive maintenance delivers transformative benefits:

- **Efficiency:** Early detection and fixing of problems prevents unscheduled downtime and reduces repair costs. Optimal maintenance scheduling leads to the most efficient use of resources and less waste.
- **Accuracy:** Through fusion of visual and operational data, the framework improves anomaly detection and failure prediction by decreasing false positives and false negatives.
- **Decision making:** Real-time simulations and AI-enabled insights equip operators with actionable suggestions, allowing them to make informed decisions that enhance system reliability and productivity.

5.2. Quality Control Scenario

High quality standards are important in current manufacturing in order to guarantee customer satisfaction and minimize production waste. As proposed by the design, ViTs are combined with digital twins in the proposed framework to improve QC processes and allow real-time defect identification, impact simulations, and real-time tailoring of production systems. The process is started by collecting detailed data from visual and operational sources (Figure 5).

Acquisition of data is the basis of the quality control process. High-resolution cameras take images of products on the assembly line, focusing on surface quality, alignment, and structural integrity. Advanced imaging systems, like infrared/thermal cameras, support these views by identifying material irregularities or heat irritability that alone may not be discernible in regular image data. Simultaneously, sensors measure key operational parameters, including product dimensions, weight, and environmental factors such as temperature and humidity. The combined results of these data sources can be considered a single, holistic product quality view.

During preprocessing, the acquired data are subjected to thorough cleaning and preparation to be ready for analysis. The visual information is improved by noise reduction algorithms, which eliminate artifacts arising from environmental conditions in the image, such as lighting or dust. Methods of augmentation, including cropping, rotating and scaling, provide the visual dataset with enough diversity and robustness, and image patching divides large images into patches to be more accurately analyzed by ViTs. Operational data are normalized, to allow comparisons of sensor measurements from different sensors, and cleaned, to eliminate spurious or incomplete records. These preprocessing steps guarantee the data stream's integrity and compatibility.

The next stage involves defect detection using Vision Transformers. The ViT module works for preprocessed visual data, using its self-attention structure to detect surface defects as well as alignment problems. Faults including scratches, discoloration, or cracks

are detected with high accuracy, whereas component misalignments or dimensional error are detected as outliers. The identified defects are then categorized according to their severity, i.e., minor (form defects) or critical (performance defects). This classification helps prioritize the response to the detected anomalies.

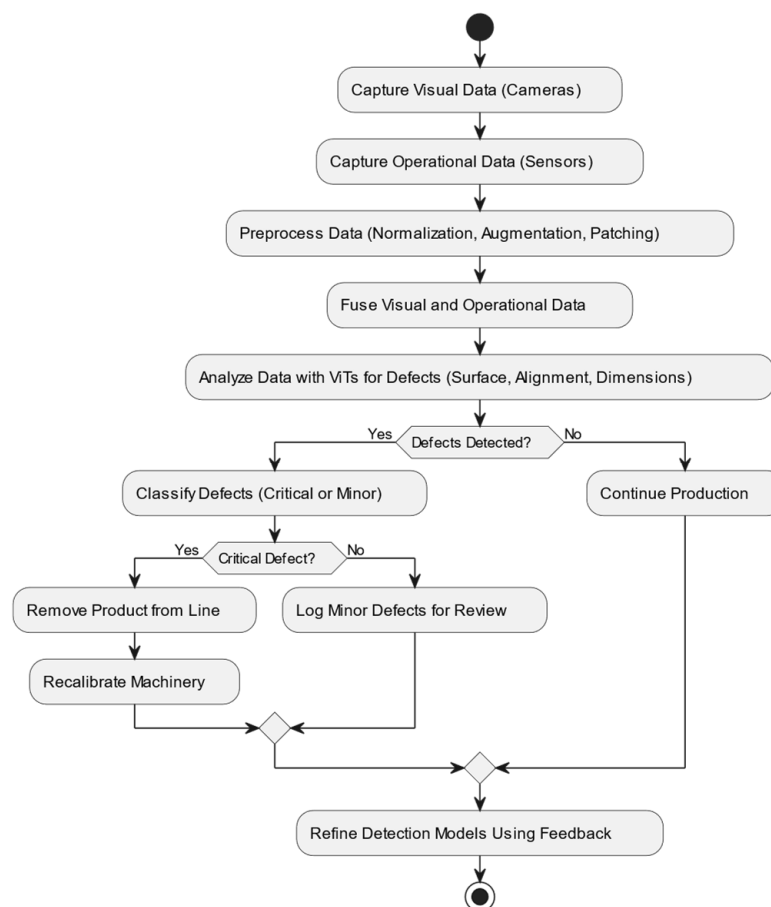


Figure 5. Process flowchart for quality control based on the framework.

After defects are identified, they are loaded into the digital twin for further analysis. The digital twin, as a virtual model of the product and its production environment, simulates the effect of the identified defects on the product's performance and durability. For example, a mechanical component out of alignment can be simulated to tell us how it impacts operational effectiveness or structural integrity. In this simulation, a deep understanding of the implication of the defect can be obtained, which is used for any subsequent corrective actions.

During the corrective action phase, the framework allows online fine-grained control of the production process. Machine parameters are recalibrated in real time to compensate for detected misalignments, so that subsequent goods comply with the specified quality requirements. Products with major defects are automatically purged from the production flow for resoldering or disposal. At the same time, routine problems or anomalies that need operator action are flagged to the operator. Process modification or machine maintenance recommendations are presented through intuitive dashboards.

The framework includes a feedback mechanism to allow for ongoing improvement. Operators may be able to give their comments on the identified defects, their categorizations, and suggested repairs. Feedback is applied to update the machine learning models such as ViTs, increasing the accuracy and the flexibility of the models over time. Also, the historical

data on defects found and the reasons for them are processed, in order to find any pattern and improve the production process, which can decrease the recurrence of the issue.

Through integrating these steps in a continuous workflow, the current proposed framework improves the throughput and outcome of quality control procedures. It guarantees that defects can be detected in time and accordingly rectified, waste can be avoided, and quality standards can be kept high in manufacturing.

Expected Outcomes

- **Enhanced defect detection:** Due to the state-of-the-art performance of ViTs, it is possible to achieve accurate detection of both visible and subtle defects, e.g., scratches, misalignments, or structural defects.
- **Improved efficiency:** Automated defect detection and classification accelerate the quality control workflow by minimizing manual inspection work and improving decision-making speed.
- **Optimized production:** Real-time adaptation guarantees that production lines remain of high quality, minimizes the quantities of waste, and maximizes production.
- **Informed decision making:** Digital twin simulation integration offers practical information on mechanisms responsible for defects as well as consequences, enabling more rational process adjustments.

5.3. Human–Robot Collaboration Scenario

Human–robot cooperation (HRC) is a key factor in Industry 5.0 to ensure harmonious cohabitation of human operators and cobots in the factory environment. The proposed framework leverages ViTs and digital twins to monitor workspace dynamics, interpret operator gestures, and adapt cobot actions in real time, ensuring safety, efficiency, and flexibility (Figure 6).

The whole process starts with data acquisition, such that cameras observe the workspace, recording operators' actions and gestures. Such cameras may consist of high-resolution, depth or stereo image sensing systems for achieving an overall comprehension of spatial relations in the workspace. Meanwhile, cobots continuously send operational data, like position, velocity, and task status, over the embedded sensor and IoT devices. This double data stream presents an interactive representation of human and cobot actions.

During the preprocessing stage, visual data are reduced to noise in order to remove distortions due to illumination or motion artifacts, etc. Data augmentation methods improve the robustness of gesture recognition by generating scale, orientation, and resolution variations in images. For operational data, normalization provides uniformity between cobots and tasks, allowing seamless integration with visual data. These preprocessed datasets are then combined at the feature level to enable joint analysis and at the decision level to fuse gesture recognition results and cobot operational states.

The gesture recognition phase is driven by ViTs. Through a self-attention mechanism, ViTs learn the visual data and extract not only low-level gestures, e.g., stop, pause, resume, but also high-level gestures, e.g., gesturing by pointing at a particular location or gesturing with other gestures to make requests, etc. At the same time, ViTs track spatial relationships to guarantee that operators do not come dangerously close to cobots, activating safety protocols if thresholds are crossed.

After gestures are identified, they are fused in the digital twin that approximates possible cobot behavior. For example, if a user points at the stop button, the digital twin simulates the deceleration of the cobot and ensures that it comes to a safe stop position with respect to the user. Similarly, if the operator requests that the task is modified, the digital twin rearranges the cobot's task order, and the new order is confirmed to solve the goals of safety and efficiency.

Based on the digital twin's simulation, the cobot executes dynamic adjustments in real time. For instance, if the operator requests repositioning, the cobot moves to the required location without a safety risk. Safety protocols are automatically enforced, e.g., reducing speed or terminating a task when a restricted zone is reached by the operator. These real-time updates improve the workflow ecology, enabling operators and cobots to interact in a seamless manner.

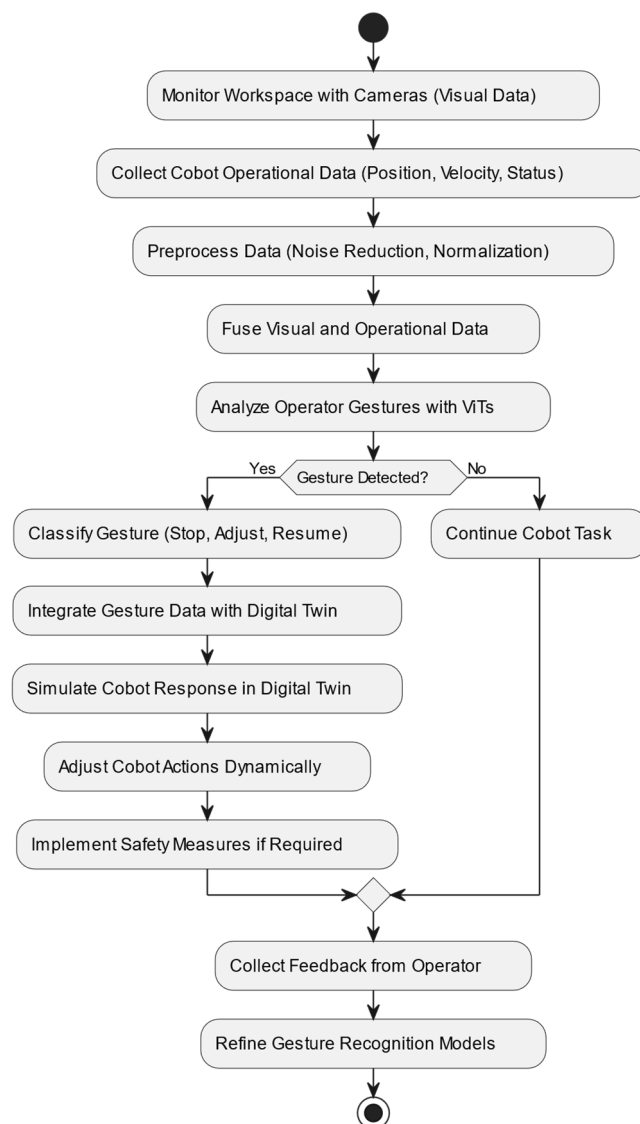


Figure 6. Process flowchart for human–robot collaboration based on the framework.

Finally, operator feedback is collected to refine the system. Operators give feedback on gesture recognition robustness and responsiveness, as well as on cobot performance, using dashboards or AR interfaces. This feedback is used to further train gesture recognition models, such that the system can continuously learn and adapt to new gestures or operating conditions. Historical records of operator–cobot interaction are also derived and investigated to detect patterns and to facilitate better future interactions.

This iterative feedback loop guarantees that the architecture continues adapting and is robust, creating a safe and effective human–robot cooperating environment. Through dynamic interaction with operator cues and respecting safety regulations, the framework conforms with the human-centered approach of Industry 5.0.

Expected Outcomes

- **Enhanced safety:** The proactive monitoring of the workspace dynamics guarantees operators and cobots stay at safe distances, thereby avoiding accidents.
- **Improved efficiency:** Real-time gesture recognition and cobot task adaptation reduce latency and improve cooperative workflows.
- **Operator empowerment:** Intuitive gesture-based controls provide operators with the ability to concentrate on higher-level tasks, while cobots are in charge of routine or physically taxing activities.

6. Advantages of the Framework

The proposed framework of combining ViTs with digital twins in Industry 5.0 provides several breakthrough benefits. Among them, improvements in data analysis, real-time processing features, and scalability to different fields can be achieved. Each one of these advantages is there for the framework's capacity to address, in an integrated way, the specific challenges posed by contemporary industrial spaces, achieving operational excellence and human-centered flexibility.

6.1. Enhanced Data Analysis

One of the most important advantages of the framework is that it is in a position to re-shape data interpretation, especially visual interpretation. Traditional image analysis methods, such as those based on CNNs, often focus on local features within an image, making it challenging to detect subtle anomalies or understand complex spatial relationships. These disadvantages are overcome by using self-attention mechanisms in ViTs, which enable them to interact with both local and global visual information at the same time.

This ability allows the framework to locate high-level details, for example, microcracks on equipment surfaces, as well as high-level patterns, for example, systemic deviations in production quality. The framework is strengthened by the addition of these findings to the digital twin, so that simulations are more accurate, leading to better predictions of the behavior of the system. The enhanced predictive accuracy of this classifier lowers the possibility of unplanned downtime, guarantees a uniform product quality, and improves the overall system reliability.

Furthermore, the architecture combines operational and visual data streams, resulting in a holistic picture of industrial processes. This integrated view enables more detailed data-based insights, and thus more sophisticated optimization paradigms and anticipatory (preventive) interventions.

6.2. Real-Time Processing Is at the Forefront

Another key benefit of the framework is the ability to collect, process, and analyze data in real time. Contemporary industrial environments produce large quantities of data that need to be analyzed and acted upon on the spot to ensure the operational efficacy and robustness of production processes. The architecture of the framework is such to be able to address this problem by implementing edge computing and parallel processing methods, thus guaranteeing low-latency data processing.

Real-time computation allows the digital twin to continuously recreate the virtual version of the physical system, updating the virtual version according to real physical changes occurring. This continuous adaptation enables a fast detection of failure, so that the reaction system can be set up to prevent failures from turning into catastrophic failures. On the one hand, a small increase in the vibration level (which is combined with a visual anomaly) can be used to immediately simulate a prediction of potential failure of equipment in the digital twin. With these predictions, operators can take diverse decisions, e.g., one can plan maintenance or change the operational parameters.

The real-time functionalities of the framework also include dynamic simulations, which enable a manufacturer to simulate operating task conditions in real time. This agility provides the ability to maintain the responsiveness of the systems to dynamic conditions (e.g., demand variability and failures), thereby increasing their robustness.

6.3. Scalability and Flexibility

Due to the modular nature of the framework, it can scale simply and seamlessly in different industrial contexts, ranging from small-scale production plants all the way to large-scale, complex manufacturing systems. Such scalability is realized by the plug-and-play nature of the framework, with its capability to participate in a wide range of data sources and to adapt to a broad spectrum of operational needs.

For instance, in a low-scale production line a framework can be used to track a small number of machines, with an emphasis on quality control or maintenance tasks. On the other hand, in a large industrial plant, the same framework can be extended to control hundreds of machines and processes, and additional modules can be modified for operational requirements.

The framework's reusability across different AI models and data modalities further increases its versatility. For example, it can be combined with the latest technologies such as generative adversarial networks (GANs) for performing synthetic data generation processes or reinforcement learning for fine-tuning decision-making procedures. This flexibility guarantees that the framework is applicable to the introduction of new technologies and industrial processes alike.

In addition to manufacturing, the design of the framework permits its use in different fields, e.g., logistics, medicine, smart city, etc. In logistics, it could be applied to trace the state of goods in transport, facilitating early corrective actions in case of anomalies such as temperature or vibrations. It may be used in healthcare to analyze medical imaging data to enhance diagnostic performance and patient results. This cross-domain applicability highlights the potential of the framework to move beyond traditional manufacturing industries into other sectors.

7. Challenges and Considerations

Although the presented framework provides unprecedented benefits, there are some challenges in its implementation. The solutions to these problems are of the utmost importance for the successful application of its concepts in Industry 5.0 environments. These issues can be categorized as technical issues and ethical and social issues.

7.1. Technical Challenges

ViTs are computationally demanding, with the use of self-attention mechanisms and large-scale parallel processing. The need for this type of analysis in real time from visual data and operational data integration is increasing. In industrial settings, in which high throughput is crucial, latency due to lack of computational power can impair the ability of the system to respond to user actions. To overcome this difficulty, taking advantage of edge computing allows data to be processed at their origin, i.e., at the edge, thereby minimizing latency and bandwidth consumption. Further, fine-tuning ViT models with methods such as model compression, quantization, or distillation has been shown to cut down the computational cost without substantial impact on the performance. Scaling infrastructure, namely, the deployment of high-performance GPUs/TPUs, will also be crucial for large-scale applications.

In order to overcome some potential latency and synchronicity problems, we propose a multi-tier architecture which aggregates edge computing to perform fast anomaly checks

together with the centralized cloud computing for advanced modeling and data archiving. The computational complexity of the Vision Transformer has previously been lowered by model compression methods, such as instance quantization, without significant degradation of accuracy. In particular, by strictly applying existing communication standards (e.g., OPC UA) and by enforcing strong cybersecurity measures, data integrity and privacy are maintained throughout the system. In terms of real-time data flow, edge computing nodes can be used for preprocessing local sensor streams and partially performing the ViT inference, thus they help to reduce data volume traffic to central servers. This method reduces latency and alleviates bandwidth limits, especially in the case of cameras recording high-resolution frames at a high frame rate. Furthermore, quantizing/pruning ViT models can further reduce inference latency while still not decreasing accuracy too much. On the data management side, adherence to consistent industrial standards (i.e., OPC UA) facilitates the integration of sensors and protects data transfer. For efficient management of big datasets, which are ever increasing in size, the introduction of two-tiered storage and the use of incremental training or fine-tuning has been proposed in the case of ViT models, so that the digital twin and AI chains are both correct and scalable over time.

The pipeline gathers and combines large quantities of operational and visual data, a considerable portion of which are sensitive and confidential. Securing these data is paramount to prevent intrusions, intellectual property theft, and abuse. Furthermore, data from multiple sources (e.g., Internet of Things devices and image systems) can be integrated, which expands the attack surface. The countermeasures to these risks include using strong cryptographic protocols for data to be stored and transmitted as well as enforcing strict access control policies to minimize system weaknesses. International data privacy regulation compliance is also required, especially in global supply chains. Regular audits and penetration testing can further bolster the framework's resilience against cyber threats.

Combining synchronous multi-source data streams and seamless exchange of knowledge between ViTs and digital twins is a major integration bottleneck. A natural time-space data resolution can be different from operational or visual data, necessitating complex preprocessing and fusion operations. One solution exists in using middleware platforms that ensure data format standardization and real-time integration of data. Developing a module-based architecture for the framework can also amount to a reduction in integration complexity, enabling piecemeal deployment and testing of discrete parts.

7.2. Ethical and Social Implications

The focus of Industry 5.0 is to retain a human-centered approach, but automation based on this framework might generate concerns about job displacement. With machines ever more skilled in quality control, predictive maintenance, and decision making, there is a danger that some jobs traditionally occupied by humans may become redundant. On the contrary, the focus should be shifted to upskilling and reskilling programmers, to equip people to work with new technologies. Focusing on the human oversight and experience inherent in systematic operation can enhance the framework's congruence with the human-centered approach of Industry 5.0.

Both ViTs and other AI components are trained on ground truth data, and thus the data may be biased or non-representative, leading to undesirable and discriminatory decision making. For instance, a contaminated dataset might not be able to identify defects in some materials, or to identify some types of operational anomalies. Collecting diverse and representative training datasets is a critical first step in reducing bias. Periodic audits of the AI models for fairness and accuracy can allow ecosystems to keep trust in the decision making of the framework.

AI-based systems, such as systems using ViTs, sometimes behave as “black boxes”, making it difficult to understand the decision-making process. That same lack of transparency can have a harmful effect on the trust between operators and managers. Adding explainable AI (XAI) techniques allows us to gain understanding of how ViTs classify anomalies, perform predictions, or suggest actions. Transparent reporting and the use of visualization tools on the user interface can also help promote trust and adoption in the framework.

The technical and ethical difficulties regarding the realization of this framework are significant but not insurmountable. Adequately investing in computational infrastructure, ensuring data security, and social implications will help industries adopt this transformative framework. Most importantly, the coming together of technologists, policy makers, and workers will ensure that this framework is in tune with the larger goals of Industry 5.0: sustainability, resilience, and human-centricity.

8. Conclusions

The combination of ViTs with digital twins represents a transformative advancement for Industry 5.0 by solving data-rich, human-centric, and sustainable industrial systems. This framework brings together the high-end data analysis capabilities of ViTs and the forecasting and simulation capabilities of digital twins to provide tangible solutions for applications in, e.g., predictive maintenance, quality control, and human–robot cooperation. Through its real-time data processing capability, adaptive decision making, and smooth human–machine interaction, the framework is compatible with Industry 5.0’s aims for efficiency, robustness, and sustainability. This integrative approach echoes methodologies seen in other domains, such as the use of convolutional transfer learning models for specialized applications, e.g., skin lesion classification on iOS devices, which optimizes machine learning models for practical and impactful deployment [53], as well as in educational contexts, where AI-enhanced adaptive learning systems have been explored for improving user engagement and learning outcomes [54].

Future work will focus on translating the framework into actionable prototypes and testing its efficiency in simulated and real-world conditions. The main issues are the proper assessment of the ViT module’s effectiveness, ensuring a seamless embedding using digital twins, and determining the scalability of the system across various industrial contexts. By incorporating other AI-based models, such as GANs for data augmentation or reinforcement learning for adaptive optimization, it is evolutionary to augment its application to other areas. Cross-domain applications, such as healthcare, logistics, and agriculture, also have great potential for leveraging the versatility of the framework.

Addressing technical challenges, including computational load, data privacy, and complexity of integration, will be critical to its viability. Equally, it should be ensured that the framework is based on ethical guidelines by encouraging trust through open decision making and by reducing AI model bias. Interdisciplinary collaboration among researchers, industry stakeholders, and policymakers will play a key role in detailing and scaling up the framework.

This framework provides the basis for intelligent and responsive industrial systems that will incorporate and value both technological advancement and human-centered factors. Its ability to increase efficiency, decrease waste, and promote interdisciplinary collaboration represents a major advance toward fulfilling the Industry 5.0 requirements. Continued development and research will make sure that this integration is not only driving technological advancement, but also sustainable and inclusive growth.

Through incorporating the methodological clarifications, and through better matching the case studies with the key research gaps in the Introduction, this framework

now provides a more solid foundation upon which to build practical implementations in predictive maintenance, quality control, and human–robot interaction. The refinements presented—particularly in outlining actionable steps for each framework module—underscore the practicality and adaptability of our approach. We thus support our conclusion that Vision Transformers combined with digital twins is an effective way of responding to the new importance placed on human-centeredness and sustainability in Industry 5.0. Future work will concentrate on verifying these components by a more intensive experimental program and by greater empirical involvement on the part of industrial collaborators so as to maintain the integrity of the framework in light of the continuing technological and ethical challenges raised by future breakthroughs.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lasi, H.; Fettke, P.; Kemper, H.-G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [\[CrossRef\]](#)
2. Nahavandi, S. Industry 5.0—A Human-Centric Solution. *Sustainability* **2019**, *11*, 4371. [\[CrossRef\]](#)
3. Grieves, M.; Vickers, J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In *Transdisciplinary Perspectives on Complex Systems*; Kahlen, F.-J., Flumerfelt, S., Alves, A., Eds.; Springer: Cham, Switzerland, 2017; pp. 85–113.
4. Tao, F.; Zhang, M. Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing. *IEEE Access* **2017**, *5*, 20418–20427. [\[CrossRef\]](#)
5. Chen, C.-F.R.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
6. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [\[CrossRef\]](#)
7. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002.
8. Qi, Q.; Tao, F. Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison. *IEEE Access* **2018**, *6*, 3585–3593. [\[CrossRef\]](#)
9. Zheng, Y.; Yang, S.; Cheng, H. An Application Framework of Digital Twin and Its Case Study. *J. Ambient Intell. Humaniz. Comput.* **2019**, *10*, 1141–1153. [\[CrossRef\]](#)
10. Leng, J.; Ruan, G.; Jiang, P.; Liu, Q.; Zhou, X.; Chen, X. Digital twins-based smart manufacturing system design in Industry 4.0: A review. *J. Manuf. Syst.* **2021**, *60*, 119–137. [\[CrossRef\]](#)
11. Fuller, A.; Fan, Z.; Day, C.; Barlow, C. Digital twin: Enabling technologies, challenges and open research. *IEEE Access* **2020**, *8*, 108952–108971. [\[CrossRef\]](#)
12. Shangguan, D.; Chen, L.; Ding, J. A Digital Twin-Based Approach for the Fault Diagnosis and Health Monitoring of a Complex Satellite System. *Symmetry* **2020**, *12*, 1307. [\[CrossRef\]](#)
13. Fang, Y.; Peng, C.; Lou, P.; Zhou, Z.; Hu, J.; Yan, J. Digital-Twin-Based Job Shop Scheduling Toward Smart Manufacturing. *IEEE Trans. Ind. Inform.* **2019**, *15*, 6425–6435. [\[CrossRef\]](#)
14. Kineber, A.F.; Singh, A.K.; Fazeli, A.; Mohandes, S.R.; Cheung, C.; Arashpour, M.; Ejohwomu, O.; Zayed, T. Modelling the relationship between digital twins implementation barriers and sustainability pillars: Insights from building and construction sector. *Sustain. Cities Soc.* **2023**, *99*, 104930. [\[CrossRef\]](#)
15. Wang, W.; Zaheer, Q.; Qiu, S.; Wang, W.; Ai, C.; Wang, J.; Wang, S.; Hu, W. Digital Twins in Operation and Maintenance (O&P). In *Digital Twin Technologies in Transportation Infrastructure Management*; Springer: Singapore, 2024; pp. 179–203.
16. Abisset-Chavanne, E.; Coupaye, T.; Golra, F.R.; Lamy, D.; Piel, A.; Scart, O.; Vicat-Blanc, P. A Digital Twin Use Cases Classification and Definition Framework Based on Industrial Feedback. *Comput. Ind.* **2024**, *161*, 104113. [\[CrossRef\]](#)
17. Iliuță, M.-E.; Moisesescu, M.-A.; Pop, E.; Ionita, A.-D.; Caramihai, S.-I.; Mitulescu, T.-C. Digital Twin—A Review of the Evolution from Concept to Technology and Its Analytical Perspectives on Applications in Various Fields. *Appl. Sci.* **2024**, *14*, 5454. [\[CrossRef\]](#)

18. Attaran, S.; Attaran, M.; Celik, B.G. Digital Twins and Industrial Internet of Things: Uncovering Operational Intelligence in Industry 4.0. *Decis. Anal. J.* **2024**, *10*, 100398. [\[CrossRef\]](#)
19. Katsoulakis, E.; Wang, Q.; Wu, H.; Shahriyari, L.; Fletcher, R.; Liu, J.; Achenie, L.; Liu, H.; Jackson, P.; Xiao, Y.; et al. Digital Twins for Health: A Scoping Review. *NPJ Digit. Med.* **2024**, *7*, 77. [\[CrossRef\]](#)
20. Peldon, D.; Banihashemi, S.; LeNguyen, K.; Derrible, S. Navigating Urban Complexity: The Transformative Role of Digital Twins in Smart City Development. *Sustain. Cities Soc.* **2024**, *111*, 105583. [\[CrossRef\]](#)
21. Yang, B.; Zhang, B.; Han, Y.; Liu, B.; Hu, J.; Jin, Y. Vision transformer-based visual language understanding of the construction process. *Alex. Eng. J.* **2024**, *99*, 242–256. [\[CrossRef\]](#)
22. Sun, Q.; Zhang, J.; Fang, Z.; Gao, Y. Self-Enhanced Attention for Image Captioning. *Neural Process. Lett.* **2024**, *56*, 131. [\[CrossRef\]](#)
23. Boulila, W.; Ghadorh, H.; Masood, S.; Alzahem, A.; Koubaa, A.; Ahmed, F.; Khan, Z.; Ahmad, J. A transformer-based approach empowered by a self-attention technique for semantic segmentation in remote sensing. *Heliyon* **2024**, *10*, e29396. [\[CrossRef\]](#)
24. Mauricio, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [\[CrossRef\]](#)
25. Jamil, S.; Piran, M.J.; Kwon, O.-J. A Comprehensive Survey of Transformers for Computer Vision. *Drones* **2023**, *7*, 287. [\[CrossRef\]](#)
26. Kameswari, C.S.; Kavitha, J.; Reddy, T.S.; Chinthaguntla, B.; Jagatheesaperumal, S.K.; Gaftandzhieva, S.; Doneva, R. An Overview of Vision Transformers for Image Processing: A Survey. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 30. [\[CrossRef\]](#)
27. Rangel, G.; Cuevas-Tello, J.C.; Nunez-Varela, J.; Puente, C.; Silva-Trujillo, A.G. A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks. *J. Sens.* **2024**, *2024*, 2797320. [\[CrossRef\]](#)
28. Zhou, Q.; Li, X.; He, L.; Yang, Y.; Cheng, G.; Tong, Y.; Ma, L.; Tao, D. TransVOD: End-to-End Video Object Detection with Spatial-Temporal Transformers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7853–7869. [\[CrossRef\]](#)
29. Zhang, B.; Liu, L.; Phan, M.H.; Tian, Z.; Shen, C.; Liu, Y. Segvit v2: Exploring efficient and continual semantic segmentation with plain vision transformers. *Int. J. Comput. Vis.* **2024**, *132*, 1126–1147. [\[CrossRef\]](#)
30. Hussain, A.; Hussain, T.; Ullah, W.; Baik, S.W. Vision transformer and deep sequence learning for human activity recognition in surveillance videos. *Comput. Intell. Neurosci.* **2022**, *2022*, 3454167. [\[CrossRef\]](#)
31. Amr, A. Future of Industry 5.0 in Society: Human-Centric Solutions, Challenges and Prospective Research Areas. *J. Cloud Comput.* **2022**, *11*, 40.
32. Ghobakhloo, M.; Iranmanesh, M.; Foroughi, B.; Tirkolaee, E.B.; Asadi, S.; Amran, A. Industry 5.0 Implications for Inclusive Sustainable Manufacturing: An Evidence-Knowledge-Based Strategic Roadmap. *J. Clean. Prod.* **2023**, *417*, 138023. [\[CrossRef\]](#)
33. Kalinaki, K.; Yahya, U.; Malik, O.A.; Lai, D.T.C. A Review of Big Data Analytics and Artificial Intelligence in Industry 5.0 for Smart Decision-Making. In *Human-Centered Approaches in Industry 5.0: Human-Machine Interaction, Virtual Reality Training, and Customer Sentiment Analysis*; IGI Global: Hershey, PA, USA, 2024; pp. 24–47.
34. Zafar, M.H.; Langås, E.F.; Sanfilippo, F. Exploring the synergies between collaborative robotics, digital twins, augmentation, and industry 5.0 for smart manufacturing: A state-of-the-art review. *Robot. Comput.-Integr. Manuf.* **2024**, *89*, 102769. [\[CrossRef\]](#)
35. Yitmen, I.; Almusaed, A. Synopsis of Industry 5.0 Paradigm for Human-Robot Collaboration. In *Industry 4.0 Transformation Towards Industry 5.0 Paradigm—Challenges, Opportunities and Practices*; Yitmen, I., Almusaed, A., Eds.; IntechOpen: London, UK, 2024; pp. 24–47.
36. Masoomi, B.; Sahebi, I.G.; Ghobakhloo, M.; Mosayebi, A. Do Industry 5.0 Advantages Address the Sustainable Development Challenges of the Renewable Energy Supply Chain? *Sustain. Prod. Consum.* **2023**, *43*, 94–112. [\[CrossRef\]](#)
37. Rame, R.; Purwanto, P.; Sudarno, S. Industry 5.0 and Sustainability: An Overview of Emerging Trends and Challenges for a Green Future. *Innov. Green Dev.* **2024**, *3*, 100173. [\[CrossRef\]](#)
38. Khan, M.; Haleem, A.; Javaid, M. Changes and Improvements in Industry 5.0: A Strategic Approach to Overcome the Challenges of Industry 4.0. *Green Technol. Sustain.* **2023**, *1*, 100020. [\[CrossRef\]](#)
39. Amr, A.; Noor, H.S.A. Human-Centric Collaboration and Industry 5.0 Framework in Smart Cities and Communities: Fostering Sustainable Development Goals 3, 4, 9, and 11 in Society 5.0. *Smart Cities* **2024**, *7*, 1723–1775. [\[CrossRef\]](#)
40. Ordieres-Meré, J.; Gutierrez, M.; Villalba-Díez, J. Toward the Industry 5.0 Paradigm: Increasing Value Creation through the Robust Integration of Humans and Machines. *Comput. Ind.* **2023**, *150*, 103947. [\[CrossRef\]](#)
41. Murtaza, A.A.; Saher, A.; Zafar, M.H.; Moosavi, S.K.R.; Aftab, M.F.; Sanfilippo, F. Paradigm Shift for Predictive Maintenance and Condition Monitoring from Industry 4.0 to Industry 5.0: A Systematic Review, Challenges and Case Study. *Results Eng.* **2024**, *24*, 102935. [\[CrossRef\]](#)
42. Ghobakhloo, M.; Iranmanesh, M.; Mubarak, M.F.; Mubarik, M.; Rejeb, A.; Nilashi, M. Identifying Industry 5.0 Contributions to Sustainable Development: A Strategy Roadmap for Delivering Sustainability Values. *Sust. Prod. Consum.* **2022**, *33*, 716–737. [\[CrossRef\]](#)
43. Longo, F.; Padovano, A.; Umbrello, S. Value-Oriented and Ethical Technology Engineering in Industry 5.0: A Human-Centric Perspective for the Design of the Factory of the Future. *Appl. Sci.* **2020**, *10*, 4182. [\[CrossRef\]](#)

44. Khan, L.U.; Yaqoob, I.; Imran, M.; Han, Z.; Hong, C.S. 6G wireless systems: A vision, architectural elements, and future directions. *IEEE Access* **2020**, *8*, 147029–147044. [[CrossRef](#)]
45. Francisti, J.; Balogh, Z.; Reichel, J.; Magdin, M.; Koprda, Š.; Molnár, G. Application Experiences Using IoT Devices in Education. *Appl. Sci.* **2020**, *10*, 7286. [[CrossRef](#)]
46. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge computing: Vision and challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. [[CrossRef](#)]
47. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
48. Nascimento, F.H.; Cardoso, S.A.; Lima, A.M.; Santos, D.F. Synchronizing a collaborative arm’s digital twin in real-time. In Proceedings of the 2023 Latin American Robotics Symposium (LARS), 2023 Brazilian Symposium on Robotics (SBR), and 2023 Workshop on Robotics in Education (WRE), Salvador, Brazil, 9–11 October 2023; pp. 230–235.
49. Tao, F.; Zhang, H.; Liu, A.; Nee, A.Y. Digital twin in industry: State-of-the-art. *IEEE Trans. Ind. Inform.* **2018**, *15*, 2405–2415. [[CrossRef](#)]
50. Molnár, G.; Sik, D. Smart devices, smart environments, smart students—A review on educational opportunities in virtual and augmented reality learning environment. In Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications, Naples, Italy, 23–25 October 2019; pp. 495–498.
51. Nagy, E.; Karl, E.; Molnár, G. Exploring the Role of Human-Robot Interactions, within the Context of the Effectiveness of a NAO Robot. *Acta Polytech. Hung.* **2024**, *21*, 177–190. [[CrossRef](#)]
52. Nagy, E. Robots in educational processes. *J. Appl. Multimed.* **2022**, *17*, 1–7.
53. Szabo, A.B.; Katona, J. A Machine Learning Approach for Skin Lesion Classification on iOS: Implementing and Optimizing a Convolutional Transfer Learning Model with Create ML. *Int. J. Comput. Appl.* **2024**, *46*, 666–685.
54. Gyonyoru, K.I.K.; Katona, J. Student Perceptions of AI-Enhanced Adaptive Learning Systems: A Pilot Survey. In Proceedings of the IEEE 7th International Conference and Workshop in Óbuda on Electrical and Power Engineering, Budapest, Hungary, 17–18 October 2024; pp. 93–98.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.