

Vision-Based Human Action Recognition for Automated Maintenance Operation Detection in Industrial Equipment Using Machine Learning Algorithms

Natalia Koteleva

Department of Automation of Technological Processes and Production, Empress Catherine II Saint Petersburg Mining University, 2, 21 Line of Vasilyevsky Island, 199106 St. Petersburg, Russia; koteleva_ni@pers.spmi.ru

Abstract

The maintenance and repair of industrial equipment is a process whose automation can improve the efficiency and safety of industrial production. This paper presents a software service that recognizes service engineer actions from video captured during equipment maintenance. The development stages of this service are described, along with the key features selected for training the machine learning model and their use in implementation. Experiments were conducted to test the software service under laboratory conditions, using the disassembly of a centrifugal oil pump and the replacement of a device in a smart panel for electricity distribution and metering as case studies. The trained model recognizes the required actions with at least 87% accuracy. Finally, the paper describes the testing and hardware requirements, the main functionality, and the advantages of using the service. The results show that identifying the actions of a service engineer requires only the position of their hands in two coordinates, x and y . Dimensionality reduction methods require the extraction of three types of features—geometric, angular, and distance—as well as the use of statistical parameters of time series data in conjunction with the energy and entropy of the wavelet transform coefficients of the time series data. These transformations enable the use of simplified machine learning methods and do not require high-powered devices on-site. It is shown that the presented method allows for a significant reduction in the incoming information for the service to operate (by 31.72 times), which is expressed by storing the coordinates of the hands instead of the video (5.9 MB versus 186 KB).

Keywords: data analysis; machine learning; computer vision; maintenance and repair; automation control system; Human Action Recognition



Received: 28 July 2026

Revised: 22 August 2026

Accepted: 3 September 2026

Published: 7 September 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The relevance of various areas of industrial enterprise operations changes over time. While the relevance of equipment maintenance and repair (M&R) remains static, approaches to its implementation change as new technologies develop. During maintenance, companies may pursue various strategies. While the scientific literature does not clearly categorize these strategies, four main ones are most commonly identified: reactive, preventive, predictive, and proactive. Modern companies are striving to reduce the share of these two maintenance strategies and are moving toward predictive maintenance, and in some cases, condition-based maintenance, i.e., proactive maintenance. However, without organizational and technical modernization of both the maintenance process itself and the systems supporting it, including automation systems, this transition is impossible. However, it is

necessary. Various statistical data support this. For example, the paper [1] shows that several investigations reveal that 30–40% of all accidents and precursor events in the chemical process industry are maintenance related. A primary challenge in deploying advanced predictive maintenance and condition-based monitoring is the influence of environmental and operational variations (EOV), such as temperature shifts or load changes. These fluctuations alter sensor signals, often masking structural defects or triggering costly false alarms. To ensure certain defect diagnosis, recent research emphasizes baseline compensation and adaptive data normalization to isolate EOV effects from true physical degradation [2]. Modern statistics also demonstrate a link between maintenance and incidents. In this case, the aviation industry can be considered, where incidents are recorded more openly, including the causes of these incidents. Current statistics [3] show that 75 percent of aviation accidents are related to human error. Of these, 30 percent are directly related to maintenance. Furthermore, a number of studies examine the factors that cause errors during equipment maintenance [4]. Of these, 30 percent are various violations of routine maintenance procedures. Therefore, it is necessary to shift the search for new technologies suitable for equipment maintenance systems. The modern development of industrial enterprises occurs in parallel with the advancement of digital technologies, many of which have already reached a level of maturity where their seamless implementation in production will bring benefits [5,6]. This static understanding of the need for repair and maintenance, coupled with the dynamics of new technologies, creates a special zone around the task of equipment M&R. If this zone provides a balance between long-established principles of organizational and technical information support for M&R and new paradigms, then the enterprise's operational efficiency can increase manifold without significant technological modifications [7,8]. Traditional maintenance monitoring methods often rely on manual data entry, which leads to errors and delays. Vision-Based Human Action Recognition (VHAR) addresses this issue. However, the key challenge is choosing the right digital technologies for its implementation. Augmented reality technologies, machine vision, artificial intelligence, and big data are a minimal set of digital technologies whose implementation is already yielding tangible benefits for oil and gas companies [9,10], mining [11,12], metallurgy [13,14], and the electric power industry [15,16].

The implementation of software services in existing production facilities is not new [17,18]. The organization of communication channels can pose a challenge during implementation [19,20]. Approaches providing information support to service engineers during maintenance and repair are being widely developed by the scientific community [21,22]. Most of these are based on augmented, virtual, and mixed reality technologies [23,24]. For example, work [25] provides a practical example of routine maintenance tasks performed by unskilled service workers. The effectiveness of augmented reality technology is demonstrated, highlighting the successful completion of all assigned tasks. The paper [26] presents the structure of the maintenance support system based on the application of augmented reality technology. A comprehensive equipment maintenance support system is developed to perform maintenance tasks. The paper [27] shows how virtual reality is used to simplify maintenance methods and create technical documentation. While general VHAR framework provides a foundation, the recent technical literature emphasizes specialized vision-based hand-pose estimation and skeleton-action recognition to track precise manual operations [28]. Furthermore, integrating temporal-segmentation and industrial assembly-recognition models allows systems to decompose continuous maintenance video into discrete steps, automatically verifying procedural compliance [29]. A number of works are devoted to the topics of transition from corrective, scheduled and condition-based maintenance to new strategies [30,31]. For example, the paper [32] shows the transition of the offshore wind turbine blade maintenance strategy to a maintenance strategy based on

reinforcement learning. The paper [33] develops a general policy of globally dynamic group maintenance for multi-component degrading systems, which automatically updates all information about the system state instantly after the completion of each group maintenance task. The paper [34] presents the PPOE approach to making maintenance decisions, allowing to reduce the total costs of maintenance. The paper [35] shows the optimization of the maintenance process and changes in maintenance policies through the implementation and use of a digital twin. A number of studies have explored the combined idea of transitioning to new maintenance strategies and using digital twins [36,37]. Examples of new research and developments linking digital technologies and maintenance include [38], which proposes using language models and natural language processing to decompose maintenance actions. Ref. [39] also emphasizes the need to develop a field related to large language models and equipment maintenance. State-of-the-art machine-learning research focuses on reliability-oriented modeling and generative data-physics fusion to predict remaining useful life under data scarcity [40]. However, these methods primarily analyze isolated machinery physics rather than human-induced runtime errors. The method proposed in this study fills this gap, leveraging feature learning to govern operational workflows and minimize procedural human errors in real time.

Recent advancements in industrial human activity recognition heavily rely on deep video-based architectures, such as Two-Stream Convolutional Networks, Temporal Segment Networks, and video Transformers [41,42]. While these frameworks achieve high accuracy in detecting actions within continuous, unbroken data streams and identifying temporal action boundaries, their practical deployment in industrial maintenance environments faces severe constraints. Processing dense RGB video blocks and calculating temporal dynamics requires dedicated high-end GPU clusters. This makes them entirely incompatible with the low-power edge hardware and augmented reality wearables typically used by service engineers in the field. Moreover, ingesting raw video frames inherently captures sensitive corporate infrastructure, proprietary machinery designs, and ambient personnel data, violating strict enterprise data security policies. In addition, many continuous temporal segmentation networks struggle to scale down to the subtle, fine-grained hand movements that characterize manual maintenance micro-steps, often generating high rates of false positives.

This study proposes a method for improving the organizational, technical, and information support for equipment maintenance. This method is based on a software service that enables automation of maintenance process management, specifically identifying service personnel actions during maintenance in real time, tracking the sequence of actions, assessing the execution time of both all maintenance operations and each individual action, accounting for downtime in the repair area, and developing recommendations for reorganizing operations performed during the maintenance process.

2. Materials and Methods

The core concept behind the software service is to use the position of the service engineer's hands as the primary feature for identifying actions. Figure 1 shows the logical diagram of the software service operation.

The workflow of the developed pipeline begins with spatial state feature extraction from the raw video stream using the Google MediaPipe Hands framework [43]. For each video frame, MediaPipe tracks 21 landmarks per hand, capturing articulated finger joints and wrist positions. By extracting the (x, y) planar coordinates for both the left and right hands, the raw video stream is compressed into a multi-variate continuous signal consisting of exactly 84 time series (21 landmarks $\times$ 2 coordinates $\times$ 2 hands) [44].

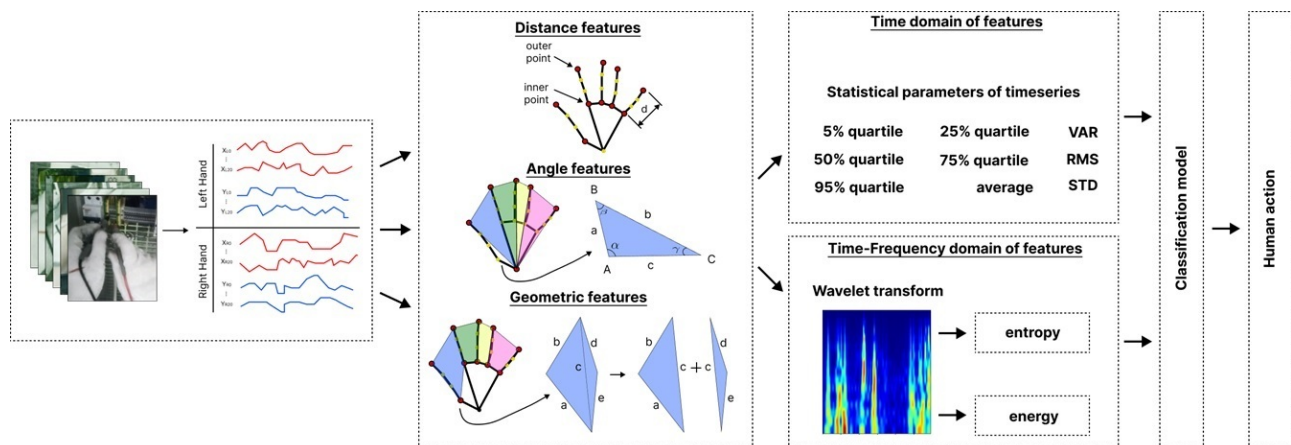


Figure 1. Scheme of operation of the software service.

While modern skeleton-based frameworks often feed these raw coordinates directly into high-overhead Graph Convolutional Networks or Recurrent Neural Networks [45,46], the method introduces a deterministic, geometry-aware feature engineering layer. This layer bridges the gap between raw coordinates and traditional machine learning classifiers by injecting domain-specific kinematic invariants and robust geometric features (Euclidean distances, joint angles, and triangular areas calculated via Heron's formula). To ensure strict reproducibility and maximize the system's resilience against spatial variations, the raw Cartesian coordinates are transformed into a normalized geometric feature space. At first, translation and scale invariance is achieved through distance normalization. The Euclidean distances between hand landmarks are inherently immune to spatial translation. To eliminate variations caused by the hand's distance to the camera or physical scale differences, every computed distance is divided by a baseline anchor distance, specifically defined as the span between the wrist and the metacarpophalangeal joint of the index finger. Because this baseline anchor changes proportionally with the hand's distance to the camera, the resulting ratio mathematically cancels out all camera-distance and scaling effects.

Second, high rotation and viewpoint robustness—approaching quasi-invariance under small out-of-plane angles—is established by calculating joint angles for triplets of landmarks using the law of cosines. Since these calculated angles are scalar angular values rather than directional vectors, they remain completely unchanged regardless of global 2-D coordinate system rotations. While 3-D perspective transformations and sharp out-of-plane rotations inherently distort 2-D projected angles, localizing these calculations to intra-hand joint triplets significantly minimizes perspective distortions compared to global frame tracking. At third, handedness invariance is secured by evaluating composite area features for landmark triangles using Heron's formula. Because Euclidean distances, calculated angles, and triangular areas are fundamentally symmetric under spatial reflection, this geometric representation yields identical feature structures for both left and right hands. Consequently, the pipeline naturally erases the distinction between left-hand and right-hand tracking data, guaranteeing complete handedness invariance. This geometric reduction maps the noisy 84-dimensional coordinate space into a highly dense, geometry-robust 36-dimensional feature space, enabling real-time deployment on computationally constrained industrial edge hardware [47].

The software service's algorithm development addresses four major tasks: hand landmark detection, temporal segmentation and sequence alignment, calculation of geometric and spatiotemporal features, and the final generation of a complex feature vector. To detect hand landmarks, a dataset representing a maintenance video is processed frame-by-frame using a custom execution wrapper around the Google MediaPipe Hands API v0.10.11.

The underlying detector operates under its standard robust industrial default settings: tracking a maximum of two hands ($\text{max_num_hands} = 2$) in continuous video mode ($\text{static_image_mode} = \text{False}$), with a minimum detection and tracking confidence threshold set to 0.5. An important issue is how to deal with missing hands in the frame. To maintain low computational latency on industrial edge servers, a deterministic zero-padding strategy is implemented. The pipeline does not spend additional processing time distinguishing between physical hand absence (e.g., a worker stepping away or an object obscuring the limb) and tracking failures. If the landmark detector fails to return data for a frame, all 42 corresponding coordinate streams ($21 \text{ landmarks} \times 2 \text{ dimensions}$) are instantly populated with a value of 0.0. This guarantees a rigid and highly predictable data layout across all iterations. To extract features from continuous actions, the 84-dimensional raw signal is segmented using a sliding window approach. The selection of the sliding window width represents a trade-off between temporal resolution and context capture. While a short 1-s window (sliding window width = 30 frames) increases the number of training samples and minimizes system latency, it fails to capture the macro-context of long-term industrial operations (such as coupling or pump shielding) and limits the resolution of the Continuous Wavelet Transform at lower frequencies. Empirical testing demonstrated that a window size of 100 frames ($\sim 3.3 \text{ s}$) is optimal. It captures complete, cyclic hand-kinematic patterns required for accurate feature extraction via Heron's formula, while maintaining a sufficient sampling volume for downstream machine learning classifiers. Within each 100-frame sliding window, the 84 raw coordinate time series are mathematically transformed into 36 geometric invariant series (10 palm distances, 24 triangular finger angles, and 2 composite area features evaluated analytically using Heron's formula).

To extract highly localized non-stationary dynamics from these geometric signatures, a Continuous Wavelet Transform layer is applied to each of the 36 timeseries. The transform is parameterized to strictly match the physical characteristics of the capture environment. The complex Shannon wavelet is chosen as Mother Wavelet for its superior ability to isolate frequency transitions in structural human-engineered movements.

For each of the 36 geometric series within a segmented window, two deterministic time-frequency descriptors are computed over the resulting Continuous Wavelet Transform coefficients: the global Shannon entropy and the average coefficient energy.

At the final stage, the input vector for the downstream classification engine is constructed by concatenating two structural feature blocks across the 36 geometric series: 1. Statistical Descriptors with 7 metrics (variance, standard deviation, and five percentiles: 5%, 25%, 50%, 75%, 95%), yielding 252 features (36×7); 2. Time-Frequency Descriptors with 2 metrics (the entropy and energy of the continuous wavelet transform coefficients of the time series), yielding 72 features (36×2).

This aggregates into a static, uniform feature vector of exactly 324 dimensions per action window. This dense but lightweight representation allows traditional machine learning models to classify the macro-actions of the maintenance dataset with negligible computational latency.

These combined features were subsequently used to train the final classification model. Various machine learning architectures can be deployed to classify and identify the service engineer's actions based on this vector.

3. Results and Discussion

First, we provide an example of the service engineer's hand position when installing devices. There are a lot of maintenance videos. We have chosen one of them and have analyzed the hand position. Let us look at freely available YouTube videos on installing a Siemens 1215 PLC on a DIN [48].

Figure 2 shows part of the control cabinet maintenance process, namely PLC maintenance.

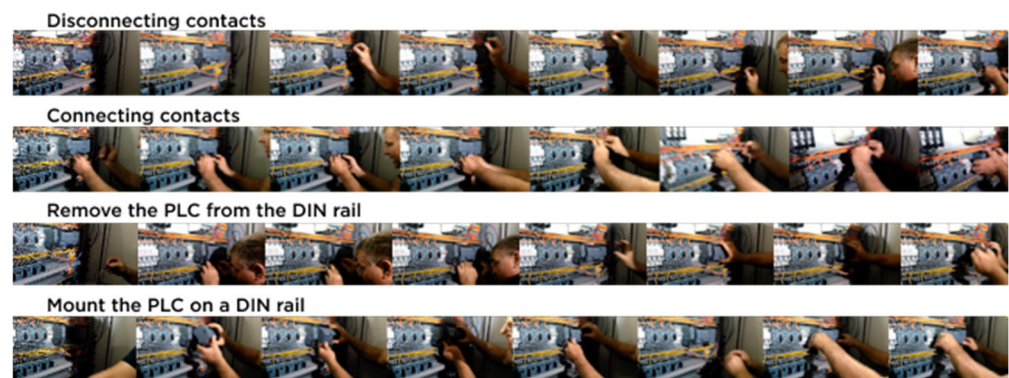


Figure 2. Video sequence for four steps involved in installing a PLC.

The presented video sequence clearly shows the hand positions of a service engineer during PLC installation. The video sequences, from top to bottom, horizontally show the following actions frame by frame: disconnecting contacts, connecting contacts, removing the PLC from a DIN rail, and installing the PLC on a DIN rail. The video sequence shown in the figure clearly demonstrates the differences in the service engineer's hand positions during these actions. These visible differences demonstrate the usefulness of analyzing the service engineer's hand position as a key characteristic.

Of course, such a conclusion cannot be drawn from a single image. It is provided here for illustrative purposes. When forming the hypothesis, a larger number of video files were analyzed, taking into account the repetition of similar operations, overlapping operations, noise and interference in the frame, etc.

The developed software service was tested in laboratory conditions. Two tasks similar to industrial ones were used for testing: servicing a Grundfos centrifugal pump (Bjerringbro, Denmark) (CR1S-4 A-FGJ-AE-HQQE) and servicing a Schneider Electric electrical panel (Rueil-Malmaison, France).

The procedure for creating the dataset warrants special emphasis. The structured data collection protocol was deployed in a laboratory environment as a multi-day study. A group of diverse human participants, wearing and without gloves, acting as service engineers, performed the maintenance sequences across multiple independent calendar days. Prior to any video recording, each service engineer underwent an instructional training phase to learn the precise technical specifications of the maintenance routine. A trial maintenance procedure was performed, and official recording commenced only after the engineer's operational competence was fully verified. Video capture was strictly segmented: a video file was initiated precisely at the start of a maintenance action and immediately terminated and closed upon the action's completion. Thus, each generated video file contains exactly one isolated action instance, preventing temporal boundary blurring. For the centrifugal pump use case, exactly 30 independent experiment files were captured from 30 diverse participants for each monitored action class. These were split into a balanced allocation: 15 complete video experiments (for each class) were designated as the training matrix, and 15 (for each class) entirely separate video experiments (recorded during independent sessions and on different days) were reserved as the holdout validation set. For the Schneider Electric electrical panel maintenance use case, 35 videos (for each class) were recorded in total and partitioned into 27 training videos and 8 holdout validation videos (for each class). Each video sequence was processed using a signature coordinate transformation procedure. Only the extracted x , y , and z spatial coordinates of the hand landmarks were saved for subsequent analysis. The raw videos themselves were immediately discarded

and never stored or used in the model training pipeline. Because the algorithm operates exclusively on these extracted numerical coordinates, it is mathematically impossible to reconstruct the original video stream. All further feature engineering and classification transformations were performed solely on the coordinate datasets, the detailed structure of which is described in Section 2 of this article.

Figure 3 shows a visualization of the sequence of actions for replacing the sealing gasket of a centrifugal pump.

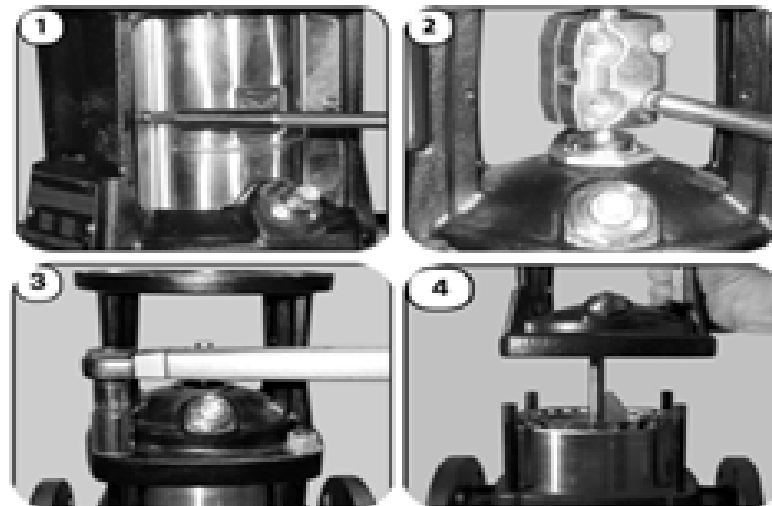


Figure 3. Steps to replace the pump seal. 1—removing the protective cover 2—removing the coupling halves, 3—loosening the nuts, and 4—disconnecting the engine.

The process under study consists of four actions: removing the protective cover, removing the coupling halves, loosening the nuts, and disconnecting the engine. To train the model, the maintenance process was repeated several times by different people, wearing and without gloves, whose work was recorded on video. Different video cameras were used for the recording. The resulting videos (15 sets for each class) were used to train the model. Table 1 and Figure 4 show the model training results.

Table 1. Model training results.

Classifier	Train Score	Test Score	Train Time
Logistic regression	1.000000	0.933333	0.016969
Neural net	1.000000	0.933333	2.908715
Nearest neighbors	0.929825	0.866667	0.000284
Linear SVM	1.000000	0.866667	0.012345
Random forest	1.000000	0.800000	0.064759
Decision tree	1.000000	0.600000	0.024723
Gradient boosting	1.000000	0.533333	0.923079
Naive Bayes	0.807018	0.533333	0.000860
Gaussian process	1.000000	0.466667	0.008130
AdaBoost	0.508772	0.400000	0.079609

Figure 4 shows the integer confusion matrix for Pump.

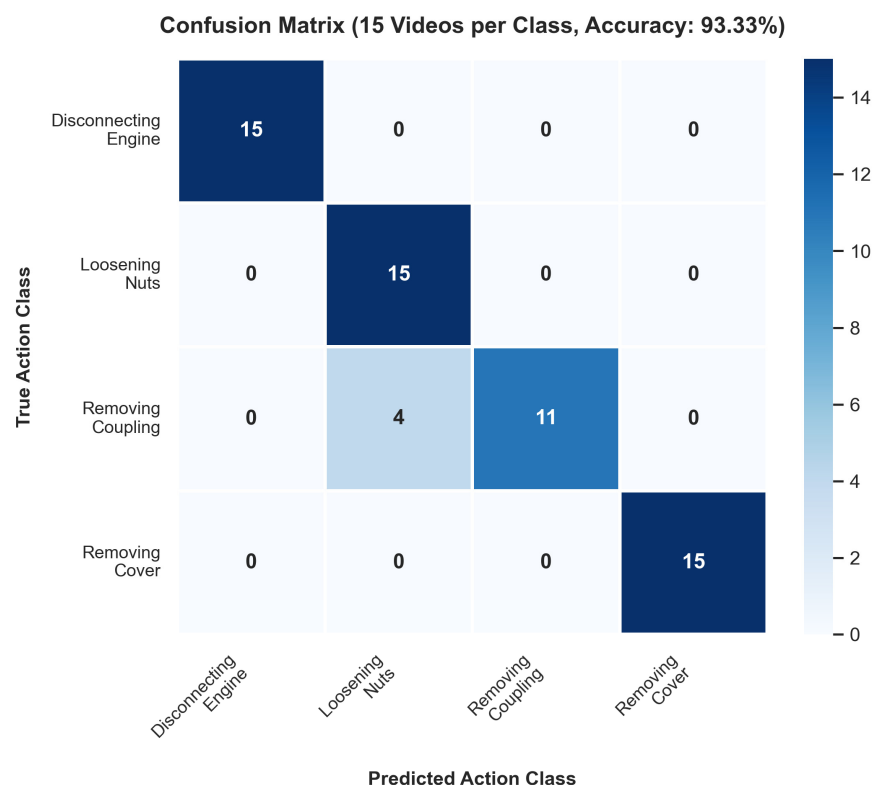


Figure 4. Integer confusion matrix.

The model accuracy on the test data was 0.93, a good result and confirms the applicability of this model in the software service. To comprehensively evaluate the robustness and discriminative power of the engineered hand-position feature family, a baseline screening was conducted across ten fundamentally diverse classifier architectures. It is critical to emphasize that all models- ranging from traditional linear estimators to ensemble trees and shallow neural networks-were evaluated utilizing their standard, default hyperparameter configurations. No iterative hyperparameter optimization or fine-tuning was performed on the holdout validation set. This approach intentionally avoids selection bias and optimistic performance shifts. The high validation accuracies achieved by top-tier models (0.933) do not stem from customized model tuning, but rather demonstrate that the proposed geometric features provide naturally clean, structured, and highly separable data boundaries across independent operator sessions.

Table 2 presents a comparison of classification accuracy using various features. The best results were achieved using the full set of features described in the article.

Table 2. Features comparison.

Class Name	Accuracy	F1	ROC AUC
Distance Features	0.56	0.51	0.72
Angle Features	0.59	0.56	0.77
Geometric Features	0.51	0.41	0.63
Time-Frequency Features	0.73	0.73	0.95
Time and Time-Frequency Features	0.93	0.93	0.99

Figure 5 shows the web interface of the developed software service. On the left, a video is shown in simulated real-time mode; on the right, the probability of each action is determined.

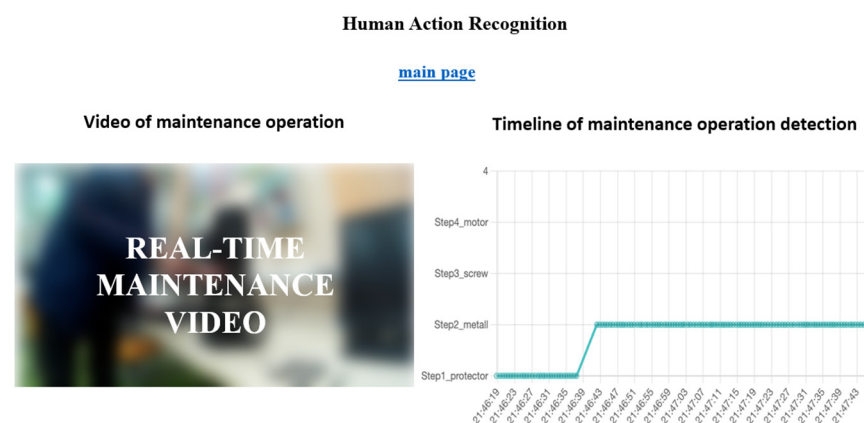


Figure 5. Web interface of the software service.

Figure 6 shows the maintenance process for a Schneider Electric smart panel, specifically the installation of its key component—the Smartlink bus. This process involves six steps. Thirty-five videos, filmed from different angles, using different devices, with and without gloves, were used to train the model.

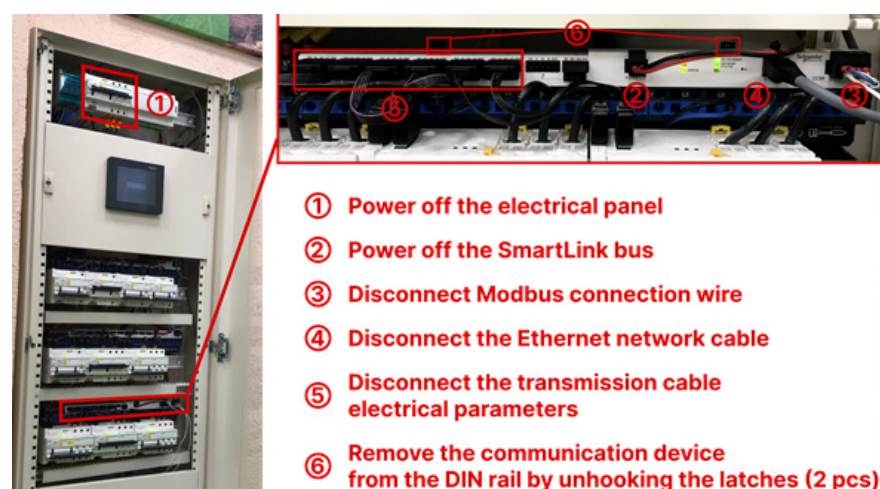


Figure 6. Schneider Electric's Smart Shield Maintenance Process.

Table 3 and Figure 7 show the model training results. The random forest classification algorithm demonstrated the best results, with an accuracy of 0.87 on the test sample. This result is acceptable for using the trained model in the software service.

Table 3. Model training results.

Classifier	Train Score	Test Score	Train Time
Random forest	1.000000	0.87500	0.187500
Gradient boosting	1.000000	0.71875	29.359375
Decision tree	1.000000	0.65625	0.046875
Neural net	0.945946	0.65625	15.921875
Gaussian process	0.878378	0.65625	0.093750
Logistic regression	0.864865	0.62500	0.828125
Linear SVM	0.702703	0.59375	0.015625
Naive Bayes	0.702703	0.56250	0.125000
Nearest neighbors	0.689189	0.53125	0.078125
AdaBoost	0.621622	0.43750	2.078125

Figure 7 shows the integer confusion matrix for Schneider Electric’s Smart Shield Maintenance Process.

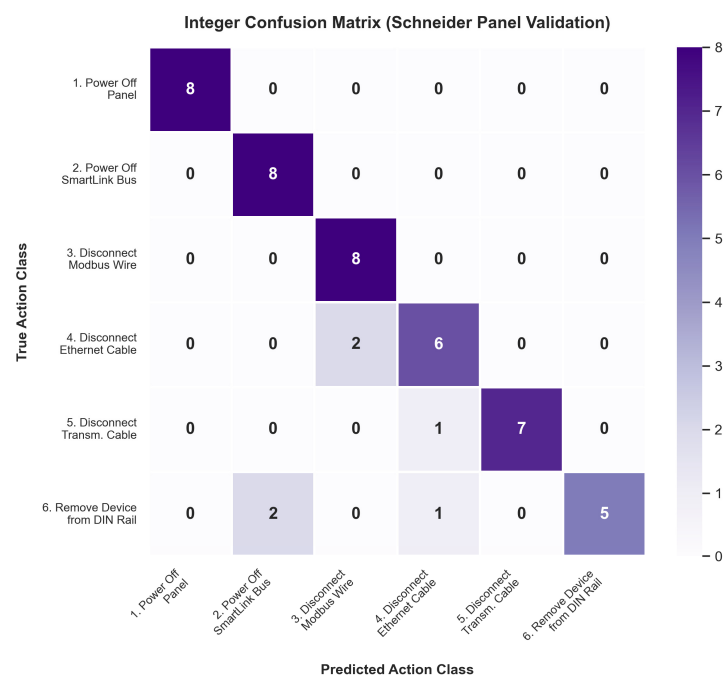


Figure 7. Confusion matrix.

Table 4 shows the results.

Table 4. Results.

Class Name	Precision	Recall	F1
Power Off Panel	1.000	1.000	1.000
Power Off SmartLink Bus	0.800	1.000	0.889
Disconnect Modbus Wire	0.800	1.000	0.889
Disconnect Ethernet Cable	0.750	0.750	0.750
Disconnect Transm. Cable	1.000	0.875	0.933
Remove Device from DIN Rail	1.000	0.625	0.769

The distribution of misclassifications in the resulting confusion matrix reflects the specific nature of the Schneider Electric’s Smart Shield Maintenance Process. Analyzing these patterns highlights both the strengths and current limitations of the action recognition system. Classes 1, 2, and 3 are recognized by the system with absolute precision. This demonstrates that the initial stages of the procedure—powering off the panel, disconnecting the SmartLink bus, and removing the Modbus wire—possess unique visual signatures. The operator’s hand movements and spatial coordinates within these zones are highly distinct, allowing the neural network to segment them perfectly. Class 4 predicted as Class 3 (2 errors): Disconnecting the Ethernet cable is confused with disconnecting the Modbus wire. This misclassification stems from the high visual similarity of the manual manipulation required to unplug interface cables, combined with nearly identical hand motility. Class 5 predicted as Class 4 (1 error): Disconnecting the transmission cable is mistakenly identified as disconnecting the Ethernet cable. This single error is caused by the closely matched physical geometry of the connectors and their immediate proximity on the device interface. The final step (removing the device from the DIN rail) shows the highest variance. During the physical detachment of the hardware, the operator’s hands inevitably obstruct a significant portion of the module. This partial occlusion causes the model to lose

track of key structural points of the DIN rail, leading it to falsely detect visual artifacts from earlier stages, such as manipulating the power bus (Step 2) or the network cable (Step 4). An overall model accuracy of 87.5% represents a strong performance for multi-class video action recognition. Every recorded error has a clear physical and visual justification.

It is worth noting that several classifiers achieve perfect accuracy (1.00) on the training data. Given the relatively straightforward and structured nature of the classification tasks, combined with the compact dimensionality of the engineered geometric feature set, the models easily find an ideal separating hyperplane for the training datasets. However, the consistently high holdout validation scores (0.933 and 0.875) on completely unseen participants demonstrate that the system avoids harmful overfitting and maintains strong generalization capabilities. As a preliminary experimental study focusing on architectural feasibility, the current results successfully confirm the validity of the hand-position tracking approach before larger-scale deployment.

Thus, the software service enables the identification of service engineers' actions during equipment maintenance and repair.

It should be emphasized that the developed software service is independent and autonomous. It can be implemented on various hardware and integrated into various existing information system architectures. The software service does not require significant computing resources.

However, additional research was conducted to confirm this. For this purpose, video processing was performed on computers with calculations performed on the processor and graphics card.

Table 5 shows the results of video processing on computers with CPU.

Table 5. Results of the software service's work on the CPU.

Characteristics of a Computing Machine	Processing Time, s	FPS, Frames per Second
M1 Pro (10-core) ARM	65	9
Intel Core i9-10940X 3.30GHz	116	5
AMD Ryzen 7 5700G 3.8GHz	79	8
Intel Core i7-11800H 2.30GHz	107	6
Quad-core Cortex-A76	196	3

Table 6 shows the results of video processing on computers with GPU.

Table 6. Results of the software service's work on the GPU.

Characteristics of a Computing Machine	Processing Time, s	FPS, Frames per Second
M1 Pro (10-core) ARM	25	24
Intel Core i9-10940X 3.30GHz	39	15
AMD Ryzen 7 5700G 3.8GHz	40	15
Intel Core i7-11800H 2.30GHz	33	18
Quad-core Cortex-A76	70	9

Detailed results for all five measurements are presented graphically in Figures 8 and 9.

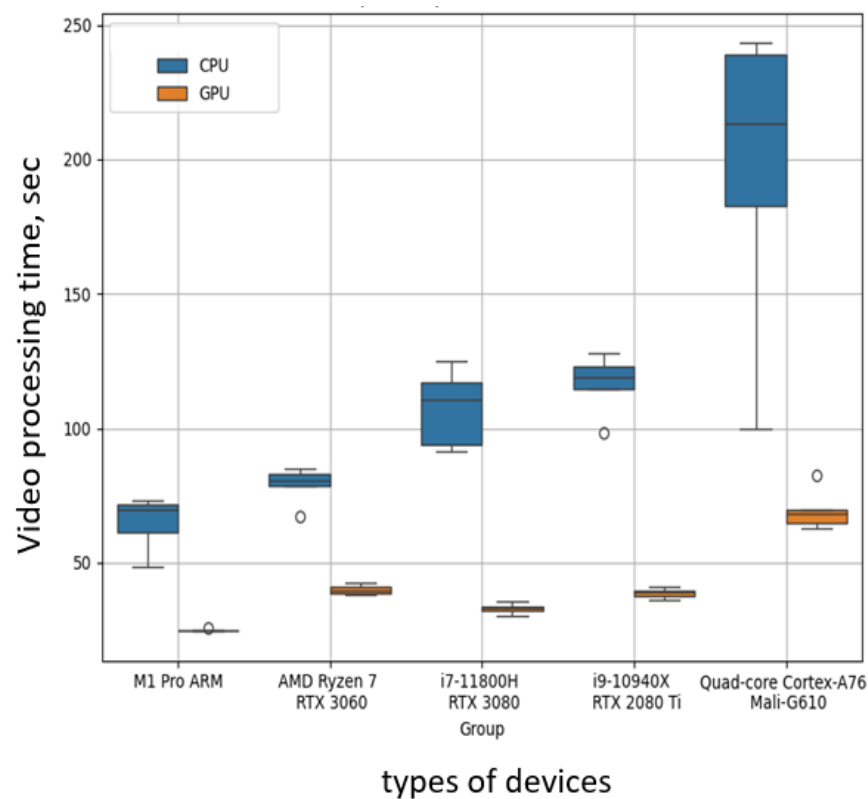


Figure 8. Testing results of the software service.

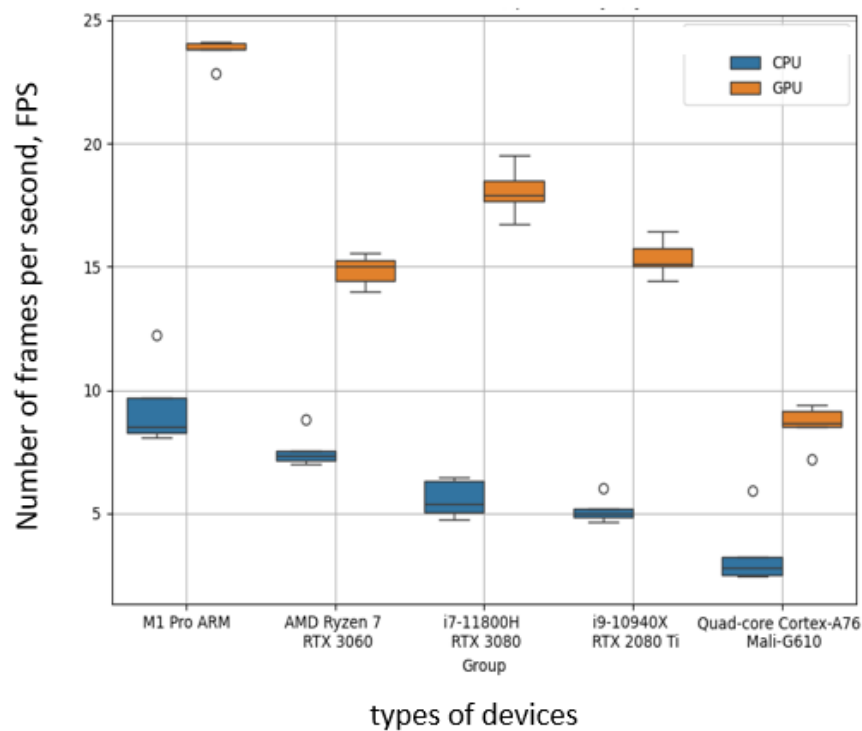


Figure 9. Testing results of the software service.

Based on the computational performance benchmarks, the software service operates with an asynchronous data-handling architecture. When deployed with a standard industrial camera capturing at a native rate of 30 FPS, the 15 FPS classification throughput does not imply frame reduction or data loss. Instead, the video ingestion layer continuously queues incoming data, executing sequential frame processing with a stable, predictable

time lag. This intentional delay is utilized to accumulate statistical performance metrics across consecutive frames. Rather than outputting volatile, independent frame-level predictions, the system aggregates the localized temporal probabilities of hand states. This statistical accumulation acts as a mathematical filter, mitigating momentary classification noise, motion blur, and spatial occlusions, thereby ensuring highly stable action boundary recognition. Furthermore, the software service is designed to function as a subsystem under a higher-level hierarchical control framework (such as a Manufacturing Execution System or an automated logging platform). This higher-level supervisor does not demand hard real-time execution limits at the millisecond scale. Instead, it requires reliable transaction state updates within a flexible, near-real-time window. A deterministic processing latency-bounded by the duration of the localized maintenance steps-is completely acceptable for operational quality auditing. This architectural configuration provides an optimal balance between low computational edge resource utilization and robust semantic accuracy.

Table 7 presents the functionality that can be implemented in maintenance and repair management systems using the software service and potential improvements. The functions listed in the table are a minimum; their number can certainly be expanded and adapted to the needs of each specific enterprise.

Table 7. Functional capabilities of the software service.

Software Service Functionality	Ways to Improve Maintenance and Repair Management Systems
Recording the execution time of each operation	planning maintenance and repairs, reducing downtime, and reducing maintenance time
	improving maintenance quality
Tracking the order of operations	timely identification of technical problems, availability, and resource availability for maintenance and repair work
Automatic creation of service engineer work time photos	planning maintenance and repair work
	optimization of maintenance and repair work

A significant advantage of the developed software service, according to the authors, is the significant reduction in the data used to implement the above-mentioned functionality, as well as the method for obtaining this information. Let us consider this using the example of a single 20-s video file used to evaluate the required hardware in the experiments described above. The video file size is 5.9 MB. The size of the input data for the software service, which enables the implementation of the functions described above, is 186 KB. This reduction in the size of the input data is 31.72 times, which undoubtedly represents an additional advantage for enterprises.

Thus, the implementation of even a basic set of digital tools in maintenance and repair management systems forms the foundation for a comprehensive enterprise transformation. The impact of implementing a software service within an industrial enterprise can consist of several components, for example: 1. Reduction in direct repair and replacement costs. Early detection of service engineer errors during maintenance prevents emergency breakdowns and allows equipment to operate at 100% of its intended service life and even exceed it. 2. Reduction and minimization of equipment downtime. 3. Reduction in maintenance costs through planning and performing maintenance based on actual condition. 4. Increased energy efficiency and resource conservation by minimizing the time equipment operates in emergency and pre-emergency conditions.

4. Conclusions

This paper introduced a lightweight software service designed to analyze service engineers' hand positions during industrial maintenance procedures using low-dimensional spatial coordinate sequences. The proposed architecture offers distinct advantages, including the extraction of behavioral transaction states in near real time, minimal data ingestion requirements, and highly reduced local data storage and processing footprints compared to heavy video-based deep learning pipelines. Furthermore, by discarding raw video data and operating exclusively on extracted coordinate vectors, the service inherently satisfies strict enterprise data privacy and security criteria. In terms of architectural scalability, the underlying design principles allow the service to be seamlessly integrated into virtually any existing enterprise information system architecture. Rather than requiring dedicated computational clusters, the software can function as a modular add-on to complement existing automated systems, expert frameworks, and augmented reality (AR)-based assistant environments deployed on wearable edge hardware. Additionally, the accumulated operational sequence database can form the foundation of a standalone analytics platform, enabling comprehensive post hoc reviews of past repairs and serving as an empirical source for broader maintenance planning and schedule optimization networks. Finally, the authors explicitly emphasize that further extensive testing and long-term pilot studies are mandatory before these laboratory findings can be fully extrapolated to enterprise scales. While the system demonstrates clear architectural viability, preliminary visual evaluations have already indicated its feasibility in continuous inference modes. However, downstream performance metrics—such as tangible maintenance cost reductions, the minimization of equipment downtime, and improvements in overall execution quality—must be treated strictly as potential systemic benefits until validated through continuous industrial deployment. Future work will focus on establishing a comprehensive testing procedure and conducting extended experiments specifically on unbroken, continuous maintenance sequences to provide rigorous empirical validation.

Funding: This research received no external funding.

Institutional Review Board Statement: Federal Law No. 323-FZ of the Russian Federation, “Basic Law on the Protection of the Health of Citizens of the Russian Federation,” clearly stipulates that the necessity of ethical review is limited to medical interventions. This study is purely technical, non-medical, and non-psychological; no sensitive personal data was collected, and therefore, according to the Labor Law of the Russian Federation, ethical review is not required.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available upon reasonable request.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Okoh, P.; Haugen, S. A study of maintenance-related major accident cases in the 21st century. *Process Saf. Environ. Prot.* **2014**, *92*, 346–356. [\[CrossRef\]](#)
2. Rezazadeh, N.; De Luca, A.; Perfetto, D.; Salami, M.R.; Lamanna, G. Systematic critical review of structural health monitoring under environmental and operational variability: Approaches for baseline compensation, adaptation, and reference-free techniques. *Smart Mater. Struct.* **2025**, *34*, 073001. [\[CrossRef\]](#)
3. Nawaz, M.O.; Zia, M.Q.; Shams, T.A. Understanding human error in military aviation maintenance: The role of Performance shaping factors, cognitive workload and error orientation. *Reliab. Eng. Syst. Saf.* **2026**, *273*, 112337. [\[CrossRef\]](#)
4. Morag, I.; Chemweno, P.; Pintelon, L.; Sheikhalishahi, M. Identifying the causes of human error in maintenance work in developing countries. *Int. J. Ind. Ergon.* **2018**, *68*, 222–230. [\[CrossRef\]](#)

5. Nevskaya, E.E.; Medvedev, D.A.; Makhovikov, A.B.; Filyasova, Y.A. Information technologies for systematization of methods improving safety in coal mines. *Sustain. Dev. Mt. Territ.* **2025**, *17*, 627–635. (In Russian) [[CrossRef](#)]
6. Zhang, B.; Zhang, J.; Chen, C. Digital technology innovation and corporate resilience. *Glob. Financ. J.* **2024**, *63*, 101042. [[CrossRef](#)]
7. Nevskaya, M.; Shabalova, A.; Kosovtseva, T.; Nikolaychuk, L. Applications of simulation modeling in mining project risk management: Criteria, algorithm, evaluation. *J. Infrastruct. Policy Dev.* **2024**, *8*, 5375. [[CrossRef](#)]
8. Stoianova, A.D.; Trofimets, V.Y.; Matrokhina, K.V. Methodological approach to rating of arctic zone companies based on ESG indicators. *Min. Inf. Anal. Bull. Sci. Tech. J.* **2024**, *6*, 149–162. [[CrossRef](#)]
9. Soromotin, A.V.; Martyushev, D.A.; Melekhin, A.A. Prediction of Hydrodynamic Parameters of the State of the Bottomhole Zone of Wells Using Machine Learning Methods. *Georesursy* **2024**, *26*, 109–117. (In Russian) [[CrossRef](#)]
10. Zakharov, L.A.; Martyushev, D.A.; Ponomareva, I.N. Predicting dynamic formation pressure using artificial intelligence methods. *J. Min. Inst.* **2022**, *253*, 23–32. [[CrossRef](#)]
11. Romashev, A.O.; Nikolaeva, N.V.; Gatiatullin, B.L. Adaptive approach formation using machine vision technology to determine the parameters of enrichment products deposition. *J. Min. Inst.* **2022**, *256*, 677–685. [[CrossRef](#)]
12. Khokhlov, S.V. Application of digital simulation methods for predicting parameters of blasted rock muckpile. *J. Min. Inst.* **2025**, *275*, 196–204.
13. Sultanbekov, R.; Beloglazov, I.; Islamov, S.; Ong, M.C. Exploring of the incompatibility of marine residual fuel: A case study using machine learning methods. *Energies* **2021**, *14*, 8422. [[CrossRef](#)]
14. Boykov, A.V.; Payor, V.A. Machine vision system for monitoring the process of levitation melting of non-ferrous metals. *Tsvetnye Met.* **2023**, *4*, 85–89. [[CrossRef](#)]
15. Zhukovskiy, Y.; Buldysko, A.; Revin, I. Induction Motor Bearing Fault Diagnosis Based on Singular Value Decomposition of the Stator Current. *Energies* **2023**, *16*, 3303. [[CrossRef](#)]
16. Korolev, N.A.; Vasilev, B.Y. The Optimization Potential in Terms of Energy Efficiency for an Asynchronous Electric Drive Through Voltage and Frequency Control with Its Technical Condition. *Energies* **2026**, *19*, 2854. [[CrossRef](#)]
17. Zeb, S.; Mahmood, A.; Ali Khowaja, S.; Dev, K.; Ali Hassan, S.; Gidlund, M.; Bellavista, P. Towards defining industry 5.0 vision with intelligent and softwarized wireless network architectures and services: A survey. *J. Netw. Comput. Appl.* **2024**, *223*, 103796. [[CrossRef](#)]
18. Ahmad, A.; Altamimi, A.B.; Aqib, J. A reference architecture for quantum computing as a service. *J. King Saud Univ.-Comput. Inf. Sci.* **2024**, *36*, 102094. [[CrossRef](#)]
19. Dvoynikov, M.V.; Kutuzov, P.A. Analysis of Efficiency of Communication Channels for Monitoring and Operational Control of Oil and Gas Wells Drilling Process. *Int. J. Eng. Trans. A Basics* **2025**, *38*, 120–131. [[CrossRef](#)]
20. Priyanka, E.B.; Thangavel, S.; Kalavathidevi, T. Miniaturized antenna design for communication establishment of peer-to-peer communication in the oil pipelines. *Pet. Res.* **2021**, *6*, 291–302. [[CrossRef](#)]
21. Obura, D.; Nabifo, V.; Apiny, I.; Akamushaba, O.; Dadebo, D. Application of a decision support system for monitoring and maintenance management of water valley tank facilities: A technological transformation toward sector sustainability. *Groundw. Sustain. Dev.* **2024**, *25*, 101109. [[CrossRef](#)]
22. Braghirolli, L.F.; Mendes, L.G.; Engbers, H.; Leohold, S.; Triska, Y.; Flores, M.R.; de Souza, R.O.; Freitag, M.; Frazzon, E.M. Improving production and maintenance planning with meta-learning-based failure prediction. *J. Manuf. Syst.* **2024**, *75*, 42–55. [[CrossRef](#)]
23. Tadeja, S.K.; Bozzi, L.O.S.; Samson, K.D.G.; Pattinson, S.W.; Bohné, T. Exploring the repair process of a 3D printer using augmented reality-based guidance. *Comput. Graph.* **2023**, *117*, 134–144. [[CrossRef](#)]
24. Wang, Z.; Bai, X.; Zhang, S.; Billingham, M.; He, W.; Wang, P.; Lan, W.; Min, H.; Chen, Y. A comprehensive review of augmented reality-based instruction in manual assembly, training and repair. *Robot. Comput.-Integr. Manuf.* **2022**, *78*, 102407. [[CrossRef](#)]
25. Wagner, M.; Leubner, C.; Strunk, J. Mixed Reality or Simply Mobile? A Case Study on Enabling Less Skilled Workers to Perform Routine Maintenance Tasks. *Procedia Comput. Sci.* **2023**, *217*, 728–736. [[CrossRef](#)]
26. Yang, X.; Huang, F.; Jiang, J.; Chen, Z. A virtual-reality spatial matching algorithm and its application on equipment maintenance support: System design and user study. *Signal Process. Image Commun.* **2024**, *129*, 117188. [[CrossRef](#)]
27. Burova, A.; Mäkelä, J.; Heinonen, H.; Becerril Palma, P.; Hakulinen, J.; Opas, V.; Siltanen, S.; Raisamo, R.; Turunen, M. Asynchronous industrial collaboration: How virtual reality and virtual tools aid the process of maintenance method development and documentation creation. *Comput. Ind.* **2022**, *140*, 103663. [[CrossRef](#)]
28. Liu, Y.; Liu, R.; Hu, Y.; Wu, M.; Xin, W.; Miao, Q.; Wu, S.; Li, L. A Systematic Review of Skeleton-Based Action Recognition: Methods, Challenges, and Future Directions. *IEEE Trans. Neural Netw. Learn. Syst.* **2026**, *37*, 2046–2065. [[CrossRef](#)] [[PubMed](#)]
29. Lu, P.; Jia, G. Semantically Guided Temporal Action Segmentation for Industrial Assembly Operations. In Proceedings of the 2026 7th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Dalian, China, 8–10 May 2026; pp. 799–803. [[CrossRef](#)]

30. Rodríguez, M.; González-Prida, V.; Sánchez, A.; Crespo, A. Application of Degradation and Optimization Models for digitization of Maintenance Management in Railway Infrastructures. *IFAC-PapersOnLine* **2024**, *58*, 115–120. [\[CrossRef\]](#)
31. Kang, C.A.O.; Zhang, Y.; Feng, J. Failure rate analysis and maintenance plan optimization method for civil aircraft parts based on data fusion. *Chin. J. Aeronaut.* **2024**, *38*, 103219. [\[CrossRef\]](#)
32. Pasca Hendradewa, A.; Yin, S. Comparative analysis of offshore wind turbine blade maintenance: RL-based and classical strategies for sustainable approach. *Reliab. Eng. Syst. Saf.* **2025**, *253*, 110477. [\[CrossRef\]](#)
33. Yang, L.; Zhou, S.; Ma, X.; Chen, Y.; Jia, H.; Dai, W. Group machinery intelligent maintenance: Adaptive health prediction and global dynamic maintenance decision-making. *Reliab. Eng. Syst. Saf.* **2024**, *252*, 110426. [\[CrossRef\]](#)
34. Xu, D.; Wang, Y.; Liu, X.; Ji, Z. A dynamic maintenance policy for degradation system by state monitoring and deep reinforcement learning. *Comput. Electr. Eng.* **2024**, *119*, 109603. [\[CrossRef\]](#)
35. Palmitessa, E.; Premoli, A.; Roda, I.; Macchi, M. Integrating maintenance and energy problems through a Digital Twin-based decision support framework under the guidance of Asset Management. *IFAC-PapersOnLine* **2024**, *58*, 7–12. [\[CrossRef\]](#)
36. Ma, S.; Flanigan, K.A.; Bergés, M. State-of-the-art review and synthesis: A requirement-based roadmap for standardized predictive maintenance automation using digital twin technologies. *Adv. Eng. Inform.* **2024**, *62*, 102800. [\[CrossRef\]](#)
37. Jeon, C.-H.; Shim, C.-S.; Lee, Y.-H.; Schooling, J. Prescriptive maintenance of prestressed concrete bridges considering digital twin and key performance indicator. *Eng. Struct.* **2024**, *302*, 117383. [\[CrossRef\]](#)
38. Giordano, V.; Fantoni, G. Decomposing maintenance actions into sub-tasks using natural language processing: A case study in an Italian automotive company. *Comput. Ind.* **2025**, *164*, 104186. [\[CrossRef\]](#)
39. Zhang, C.; Xu, Q.; Yu, Y.; Zhou, G.; Zeng, K.; Chang, F.; Ding, K. A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing. *Robot. Comput.-Integr. Manuf.* **2025**, *92*, 102883. [\[CrossRef\]](#)
40. Song, L.-K.; Faes, M.G.R.; Tao, F. Generative data-physics fusion for hierarchical reliability assessment. *Reliab. Eng. Syst. Saf.* **2026**, *272*, 112642. [\[CrossRef\]](#)
41. Rajalingam, A.; Pandi, V.S.; Awwad, H.; Ravi, N.; Karthikeyan, G.; Pavithra, M. Improve the Real-Time Industrial Inspection and Maintenance: Intelligent Recognition with AI-Powered Object Detection in Augmented Reality (AR) Environments. In *Proceedings of the 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, 28–30 November 2025; pp. 1–7. [\[CrossRef\]](#)
42. Helian, B.; Huang, G.; Geimer, M. Temporal sequence-based object detection and action recognition for mobile machinery on construction sites. *Adv. Eng. Inform.* **2025**, *68*, 103691. [\[CrossRef\]](#)
43. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.-L.; Grundmann, M. MediaPipe Hands: On-device real-time hand tracking. *arXiv* **2020**, arXiv:2006.10214. [\[CrossRef\]](#)
44. Samaan, G.H.; Wadie, A.R.; Attia, A.K.; Asaad, A.M.; Kamel, A.E.; Slim, S.O.; Abdallah, M.S.; Cho, Y.-I. MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition. *Electronics* **2022**, *11*, 3228. [\[CrossRef\]](#)
45. Liu, M.; Liu, H.; Hu, Q.; Ren, B.; Yuan, J.; Lin, J.; Wen, J. 3D skeleton-based action recognition: A review. *arXiv* **2025**, arXiv:2506.00915.
46. Wang, W.; Zhang, Y.-D. A short survey on deep learning for skeleton-based action recognition. In *Proceedings of the 14th IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC '21 Companion)*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–6. [\[CrossRef\]](#)
47. Varshini, T.S.; Rukmani, P. MP-GestLSTM: Real time gesture detection using MediaPipe and LSTM. *Syst. Sci. Control Eng.* **2025**, *13*, 2587853. [\[CrossRef\]](#)
48. Kovrigin, A. Installing a PLC Siemens 1215 [Video]. YouTube. 10 July 2019. Available online: <https://youtu.be/V62V4OsFKik> (accessed on 17 January 2024).

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.